

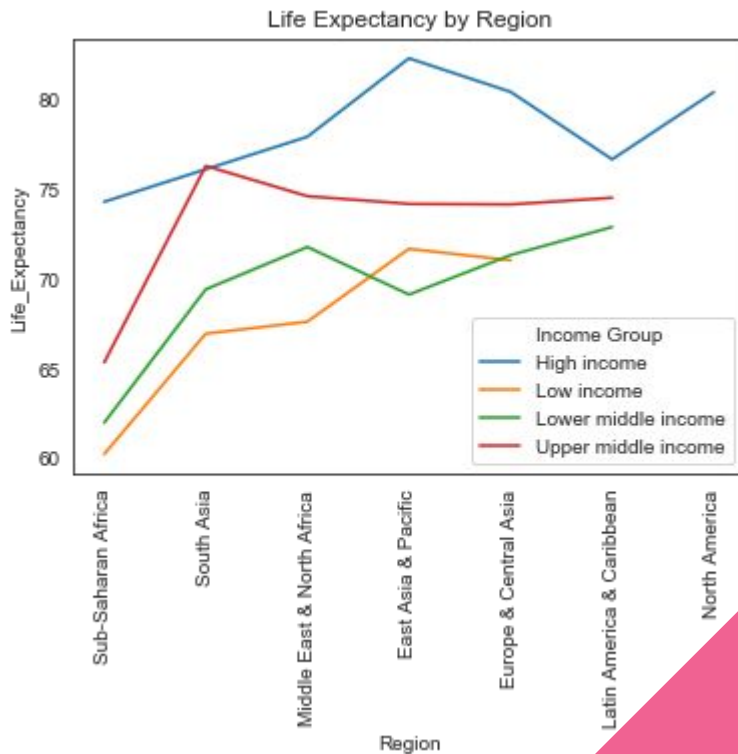


What factors have the greatest impact on life expectancy?

Life Expectancy by Region

This graph displays life expectancy by Region for year 2016.

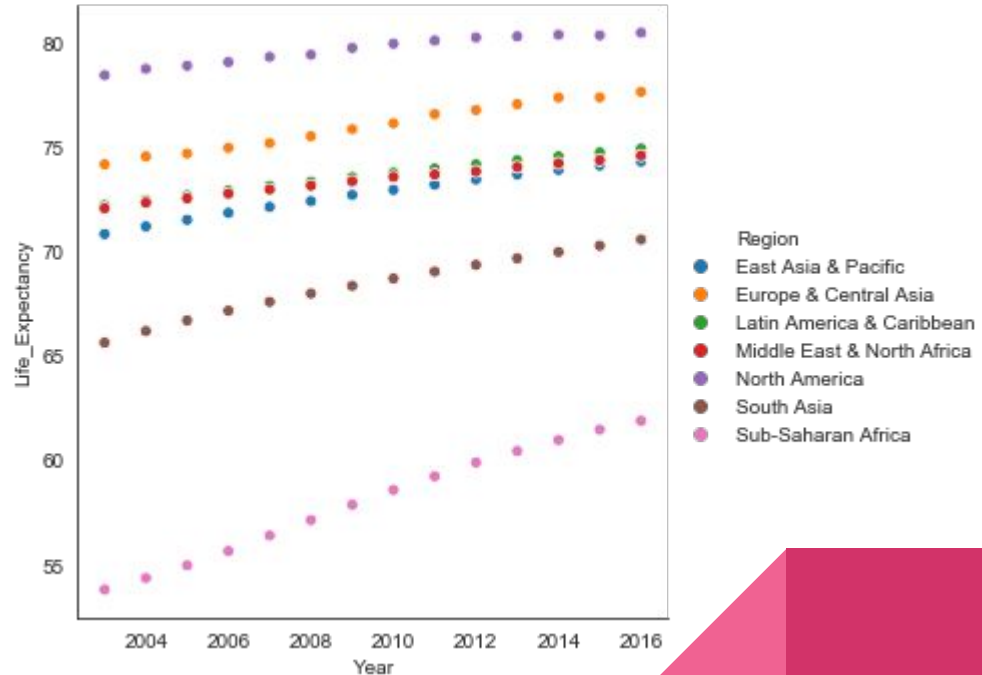
Higher incomes are definitively associated with higher life expectancies



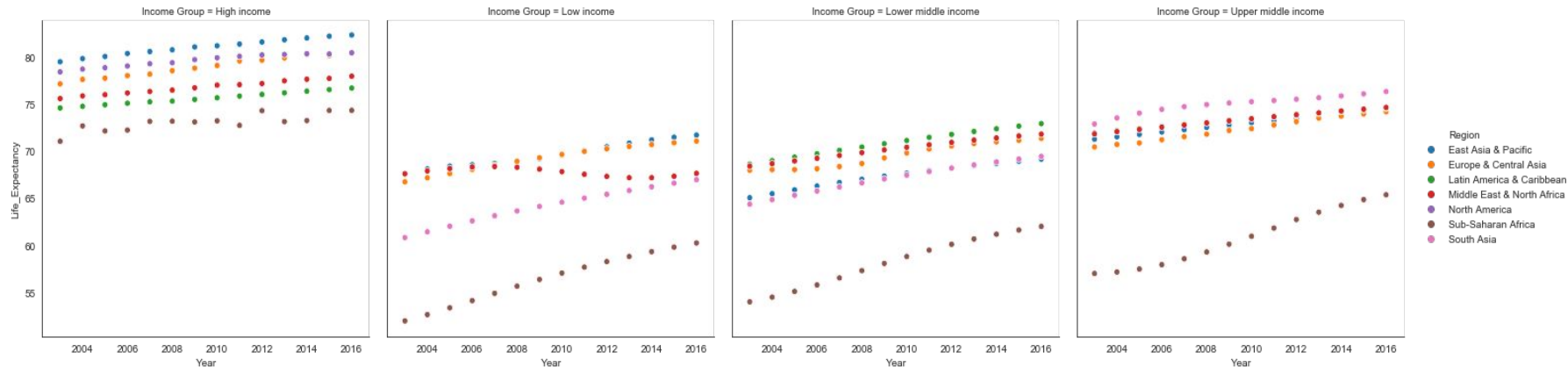
Increase in Life Expectancy over last 14 years

This depicts the change in life expectancy over the years by major region

One could conclude that living in North America or Europe & Central Asia gives the best chance for living longest



Money is a Matter



When you view the change in life expectancy for each region by income grouping, region has less significance. (Except North America and Sub-saharan Africa at the two extremes.) High Income is a clear contributor to long life.

Problem Statement


Is there any conclusive evidence of conditions or situations which will allow a person to increase their life expectancy?



Problem Solved

Be filthy rich and
live in East Asia & Pacific region.

Ok. We can't all do that, so is/are there some other things that might be attainable
to increase my life expectancy?



Background

The Health Nutrition and Population Statistics database

- Provides key health, nutrition and population statistics gathered from a variety of international and national sources
- Is aggregated by the World Bank

The data is collected and maintained by multiple sources



About the dataset

This dataset combines key health statistics from a variety of sources to provide a look at global health and population trends. It includes information on nutrition, reproductive health, education, immunization, and diseases from over 200 countries.

Update Frequency: Biannual

Last Updated: September 5, 2018



Data Overview

The data is organized by country code and indicator code

The country code identifying a specific country or geographic region

The indicator code identifying the statistic being measured

Data is measured from 1960 to 2017

There are about 25,000 records where all years contain nulls for a country and indicator - These records are removed from the data set



Data Overview

There are 405 indicator codes under 21 major topical areas in the raw data set

There are 404 indicator codes after eliminating all records where there are no measures recorded

There are 259 country/region codes in the raw data set

There are 258 country/region codes after eliminating all records where there are no measures recorded



Data Overview

The raw data has columns for country/region code & name, indicator code & name, and measures

The measures are recorded in columns for years from 1960 to 2017

Note: There is a column for 2018 but it is almost empty in this dataset so it is excluded from the analysis

However, since our question needs to be answered by evaluating the measures by type of indicator, the data needs to be partially transposed



Initial Data Analysis

While the years will be of interest because they can inform of overall shifts in life expectancy due to nonspecific factors, for the most part, only the most recent 15 years will be evaluated

There might be something of interest in trends over the years so a quick analysis will be performed to identify anything that might prove of importance to the bigger question

The indicators, though, are the primary drivers or factors in predicting life expectancy



Initial Data Analysis

Limit analysis to years 2003 to 2017

2017 is most current full year of data available

At some point, the trends change, and 15 years is sufficient to identify trends for projection

Indicators change over the years; too many years with inconsistent data will bias the results



Data Preparation

Transposed indicators and years so that the indicators become the columns

Indicators will be the features evaluated

Assigned column labels to indicators using a lookup value for topical areas and indicator itself (Note the titles in the data are unwieldy)

Added identifier for type of measure, e.g. ratio/percent, integer, scale

Added identifier for gender-specific indicators



Target Measure

The target measure is life expectancy at birth

There are 3 measures of life expectancy

- All genders

- Female

- Male



Features Overview

The major topical areas for indicators are:

Economic

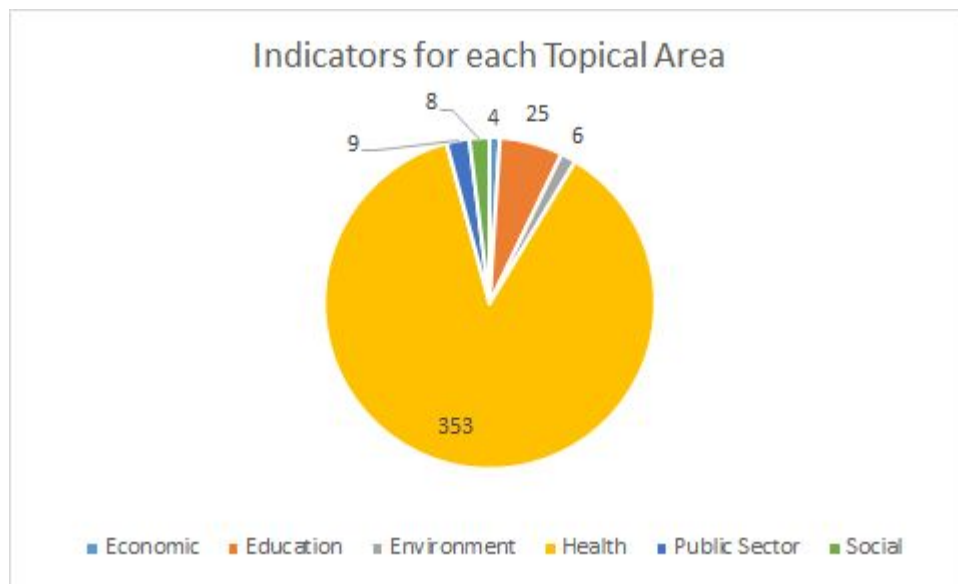
Education

Environment

Health

Public Sector

Social



Economic Indicators

There are 4 economic indicators in the dataset

There are 3 poverty indicators that are the headcount ratio of people living below the national poverty line

Measured for urban, rural, and all

Gross National Income (GNI) per capita in USD is calculated using the Atlas method (to convert other currencies to current USD)



Education indicators

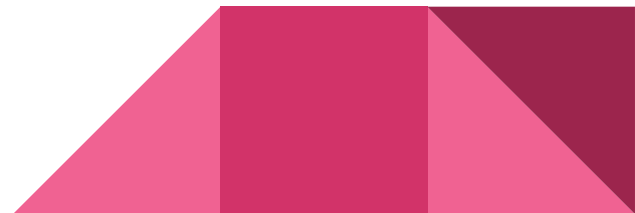
There are 25 education indicators of which all but one come from UNESCO

Education input is reflected in the indicator

Government expenditure on education, total (% of GDP)

Education outcomes are captured in 9 indicators

Education participation is captured in 15 indicators



Education outcomes

Primary completion rates are captured for male, female and all (3)

Primary completion rate is calculated by dividing the number of new entrants (enrollment minus repeaters) in the last grade of primary education, regardless of age, by the population at the entrance age for the last grade of primary education and multiplying by 100.

Literacy rates are captured in categories for youths and adults for male, female and all (6)



Education Outcomes

Literacy rate is the percentage of people ages 15-24 (youth) or ages 15 and above (adult) who can both read and write with understanding a short simple statement about their everyday life

The statistic is often self-reported and has multiple measurement methods depending on the country reporting.

Literacy rate is therefore highly suspect in terms of accuracy and should probably not be used



Education Participation

Enrollment rates are captured for primary and secondary education for both males and females in addition to total (12)

Gross enrollment includes all enrolled students regardless of age

Net enrollment includes only students of the age for the education level

Ratios are calculated by dividing the number of students enrolled by the population of the age group which officially corresponds to the education level, and multiplying by 100



Education Participation

Enrollment rates are also captured for tertiary education or post-secondary education (2)

Only gross enrollment and female gross enrollment are available in the data

Male gross enrollment is implied to be the difference between total gross enrollment and female gross enrollment and can be an added feature if enrollment rates appear to have impact on life expectancy

The final indicator in this category is related to orphans and appears quite limited and therefore will be ignored



Environment indicators

There are 6 indicators that identify population size, the year over year growth in population and the ratio to the total population for urban and rural areas.

Urban population is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects

Rural population is calculated as the difference between the total population and the urban population

While not separately included in the dataset, total population can easily be added by adding urban and rural population for each country/region

Note: it can also be checked against the population indicator by gender in the Population Structure subcategory of Health indicators

Social Indicators

There are 8 indicators classified as social

Women who were first married by age 15 refers to the percentage of women ages 20-24 who were first married by age 15 (1)

Share of women employed in the nonagricultural sector is the share of female workers in wage employment in the nonagricultural sector (industry and services), expressed as a percentage of total employment in the nonagricultural sector (1)



Social Indicators

Labor force is the supply of labor available for producing goods and services in an economy and includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers (2)

Total labor force

Female labor force as a percentage of the total labor force

Net migration is the net total of migrants during the period (1)

Considered highly unreliable and will not use



Social Indicators

Unemployment refers to the share of the labor force that is without work but available for and seeking employment (3)

Total unemployed

Unemployed males

Unemployed females



Public sector indicators

There are 9 indicators in this category all measuring the Human Capital Index (HCI)

The HCI calculates the contributions of health and education to worker productivity

The final index score ranges from zero to one and measures the productivity as a future worker of child born today relative to the benchmark of full health and complete education



Public sector indicators

The HCI is calculated in total and then using upper bounds of each component and separately using lower bounds of each component

Each is done for total, male and female

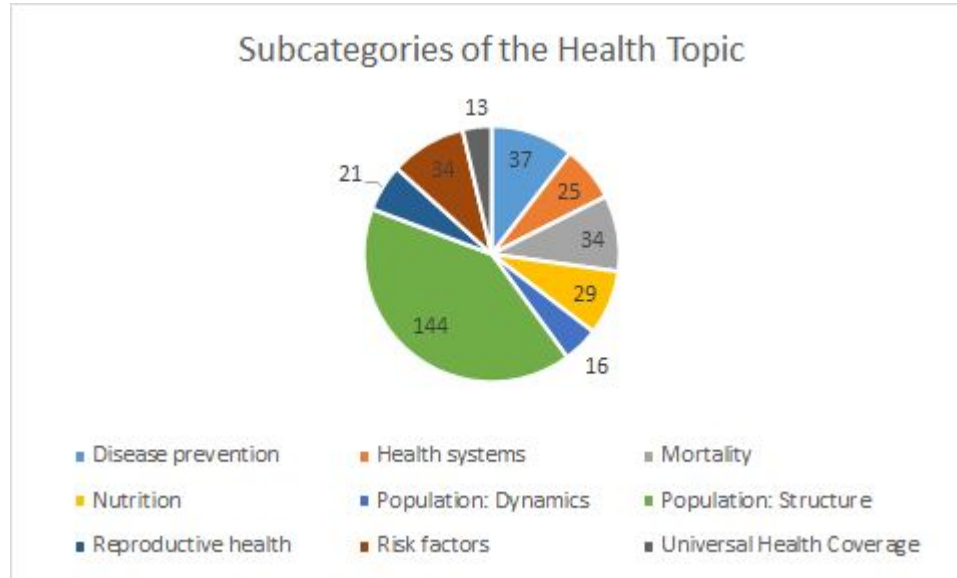
Uncertainty in HCI is high when there is great overlap in the 3 HCI calculations

Need to caveat any conclusions based from these indicators due to uncertainty in data integrity



Health indicators

There are 353 health indicators and within the health category, there are nine subcategories:



Health: Disease Prevention

Child immunization rate is the percentage of children ages 12-23 months who received vaccinations before 12 months or at any time before the survey

There are statistics for 6 different vaccines

Tuberculosis case detection rate is calculated by WHO as number of reported new and relapse cases divided by estimate of the number of incident cases x 100 (1)

Tuberculosis treatment success rate (% of new cases)



Health: Disease Prevention

These 4 indicators are based on data collected in surveys and is specific to a 2-week time frame and will not be used due to inconsistency in timing

ARI treatment (% of children under 5 taken to a health provider)

Children with fever receiving antimalarial drugs (% of children under age 5 with fever)

Diarrhea treatment (% of children under 5 receiving oral rehydration and continued feeding)

Diarrhea treatment (% of children under 5 who received ORS packet)



Health: Disease Prevention

There are five measures related to water and sanitation that are calculated as number of population with access divided by the total population and expressed as a percentage with separate calculations for urban, rural and total (15)

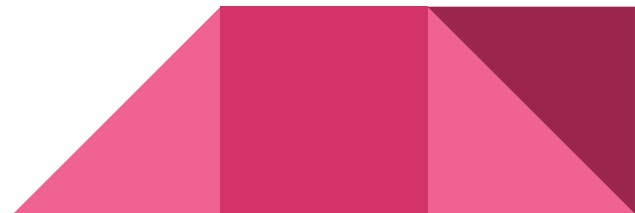
People using at least basic drinking water services

People using at least basic sanitation services

People with basic handwashing facilities including soap and water

People using safely managed drinking water services

People using safely managed sanitation services



Health: Disease Prevention

Comprehensive correct knowledge of HIV/AIDS is scored by asking the respondents to name 2 prevention techniques, scored divided by age range and gender (4)

Condom use is scored by asking respondents 2 scenarios and divided by gender response (4)

Use of insecticide-treated bed nets for children under 5 (1)

Use of Intermittent Preventive Treatment of malaria for pregnant women (1)



Health: Mortality

In the mortality category, 3 indicators (life expectancy) are the target

16 indicators are either a different expression of the target or a direct component of the target and will be excluded

Mortality caused by road traffic injury per 100,000 people (1)

Tuberculosis death rate per 100,000 people (1)

Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (1)



Health: Mortality

These measures are captured per 100,000 population in total and by gender:

Mortality rate attributed to household and ambient air pollution,
age-standardized (3)

Mortality rate attributed to unintentional poisoning (3)

Percentage of persons aged 30 that will die from CVD, cancer, diabetes or
CRD before age 70 (3)

Suicide rate (3)



Health: Health Systems

These indicators are per 1,000 people (4):

Hospital beds

Community health workers

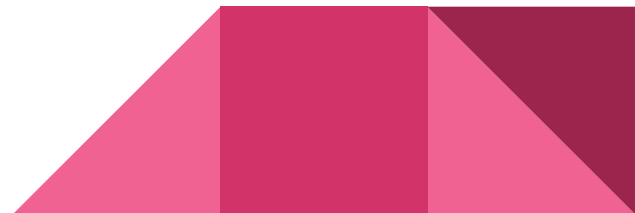
Nurses and midwives

Physicians

These indicators are per 100,000 people (2):

Specialist surgical workforce

Number of surgical procedures



Health: Health Systems

4 of the expenditure indicators:

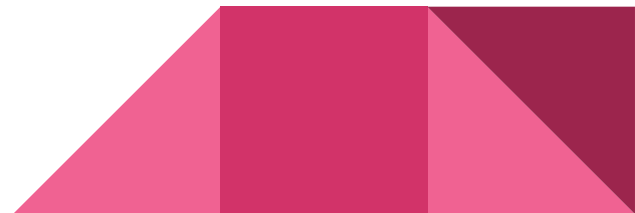
Level of current health expenditure expressed as a percentage of GDP

Current expenditures on health per capita in current US dollars

Current expenditures on health per capita expressed in international dollars at purchasing power parity (PPP)

Capital health expenditure as a percentage of GDP

Infrastructure and stocks of vaccines



Health: Health Systems

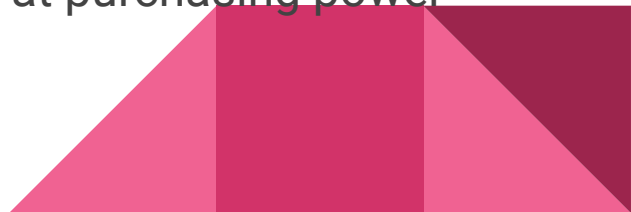
4 indicators for share of current health expenditures funded from external sources:

- Total external funding

- Amount channeled through government

- Amount per capita expressed in current US dollars

- Amount per capita expressed in international dollars at purchasing power parity (PPP)



Health: Health Systems

5 Indicators for general government health expenditure:

As a percent of current health expenditure

As a percentage of GDP

As a percent of general government expenditure

Per capita expressed in international dollars at purchasing power parity (PPP)

Per capita in current US dollars



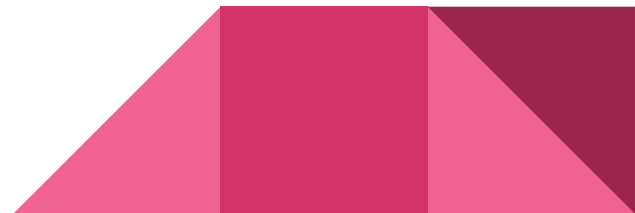
Health: Health Systems

3 indicators for share of out-of-pocket payments by households of total current health expenditures

As a percent of current health expenditure

Per capita expressed in international dollars at purchasing power parity (PPP)

Per capita in current US dollars



Health: Health Systems

3 indicators for share of current health expenditures funded from domestic private sources including prepaid health insurance

As a percent of current health expenditure

Per capita expressed in international dollars at purchasing power parity (PPP)

Per capita in current US dollars



Health: Nutrition

There are 5 miscellaneous indicators:

Consumption of iodized salt (% of households)

Exclusive breastfeeding (% of children under 6 months)

Infant and young child feeding practices, all 3 IYCF (% children ages 6-23 months)

Low-birthweight babies (% of births)

Vitamin A supplementation coverage rate (% of children ages 6-59 months)



Health: Nutrition

There are 4 indicators for anemia

Prevalence of anemia among children (% of children under 5)

Prevalence of anemia among non-pregnant women (% of women ages 15-49)

Prevalence of anemia among pregnant women (%)

Prevalence of anemia among women of reproductive age (% of women ages 15-49)



Health: Nutrition

Percentage of adults ages 18 and over whose Body Mass Index (BMI) is more than 25 kg/m² - total and by gender (3)

Percentage of children under age 5 whose weight for height is more than two standard deviations above the median for the international reference population of the corresponding age as established by the WHO's new child growth standards released in 2006 - total and by gender (3)

Note the 2006 data measure reference - exercise care in relying




Health: Nutrition

Proportion of children under age 5 whose weight for height is more than three standard deviations below the median for the international reference population - total and by gender (3)

Proportion of children under age 5 whose weight for height is more than two standard deviations below the median for the international reference population - total and by gender (3)

Percentage of children under age 5 whose weight for age is more than two standard deviations below the median for the international reference population - total and by gender (3)



Health: Nutrition

Percentage of children under age 5 whose height for age is more than two standard deviations below the median for the international reference population - total and by gender (3)

Number of people who are undernourished (1)

Percentage of the population whose food intake is insufficient to meet dietary energy requirements continuously (undernourished) (1)




Health: Population Dynamics

Some of the indicators are calculated based on bad assumptions and will be ignored for this reason (3)

Example: percentage of young dependents uses a crude calculation based solely on age ranges with no regard for outliers which could be substantive as some teens work and some over-64 are primary supporters

The indicators for completeness of death/birth registrations will not be used as features but may prove useful in vetting out countries/regions that are creating bias (6)



Health: Population Dynamics

These 3 indicators might have an impact:

Population growth (annual %)

Death rate, crude (per 1,000 people)

Birth rate, crude (per 1,000 people)

Rate of natural increase, excludes migration (to be calculated)

Crude birth rate - crude death rate



Health: Population Dynamics

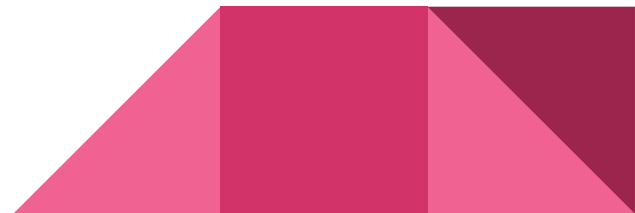
Final 4 indicators to be evaluated in this category:

Mean age at first marriage, female

Mean age at first marriage, male

Female headed households (% of households with a female head)

Women who were first married by age 18 (% of women ages 20-24)



Health: Population Structure

There are 144 indicators in this section that are almost entirely numbers of persons, in total and by gender, in various age ranges

While arguments for using this data in this analysis could be made, it is perceived to be beyond the scope of this project

The features to be considered as projection variables for life expectancy do not include the gender distribution at specific ages

The overall population distribution will be included



Health: Population Structure

The 6 indicators included from this category are:

- Population, total, male and female (3)

- Percentage to total population, female (1)

- Percentage to total population, male (1)

- Sex ratio at birth (male births per female births) (1)



Health: Reproductive Health

There are 2 measures, one collected and one calculated, for maternal mortality ratio which is defined as the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination

National estimate, per 100,000 live births

Modeled estimate, per 100,000 live births




Health: Reproductive Health

There are 3 measures for maternal death which is defined as death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes

Number of maternal deaths

Lifetime risk of maternal death is probability that a 15-year-old female will die eventually from a maternal cause, expressed as percentage and 1 in: rate varies by country



Health: Reproductive Health

Maternity leave is the mandatory minimum number of calendar days that legally must be paid by the government, the employer or both

Maternity leave benefits refers to the total percentage of wages covered by all sources during paid maternity leave

Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)

Adolescent fertility rate (births per 1,000 women ages 15-19)

Fertility rate, total (births per woman)



Health: Reproductive Health


Demand for family planning satisfied by any methods (% of married women with demand for family planning)

Demand for family planning satisfied by modern methods (% of married women with demand for family planning)

Contraceptive prevalence, modern methods (% of women ages 15-49)

Contraceptive prevalence, any methods (% of women ages 15-49)

Unmet need for contraception (% of married women ages 15-49)



Health: Reproductive Health

Wanted fertility rate (births per woman)

Pregnant women receiving prenatal care of at least four visits (% of pregnant women)

Pregnant women receiving prenatal care (%)

Births attended by skilled health staff (% of total)

Postnatal care coverage (% mothers)

Newborns protected against tetanus (%)



Health: Risk Factors

The indicators in this health subcategory will receive first priority in the evaluation process as it is presumed that high risk factors to good health have direct impact on life expectancy, including:

Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)

Cause of death, by injury (% of total)

Cause of death, by non-communicable diseases (% of total)



Health: Risk Factors

More indicators:

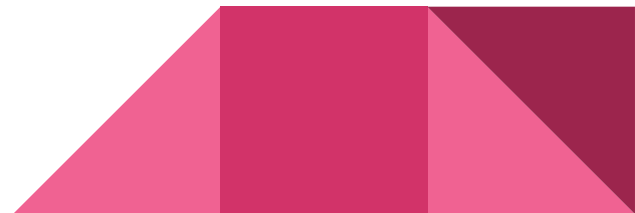
Incidence of malaria (per 1,000 population at risk)

Malaria cases reported/confirmed

Prevalence of syphilis (% of women attending antenatal care)

Diabetes prevalence (% of population ages 20 to 79)

Incidence of tuberculosis (per 100,000 people)



Health: Risk Factors

More indicators:

Alcohol consumption per capita, 15+ years of age, total and by gender (3)

Smoking prevalence, ages 15+, total and by gender (3)

Risk of catastrophic expenditure for surgical care (% of people at risk)

Risk of impoverishing expenditure for surgical care (% of people at risk)

People practicing open defecation (% of population), total, urban and rural (3)



Health: Risk Factors

Recent breakthroughs in medicine have virtually eliminated death from HIV

While, it still impacts quality of life, the disease may no longer be a risk factor to life expectancy

There are 15 indicators related to HIV and they will be included but perhaps not weighted in the evaluation

Adults (ages 15+) living with HIV

AIDS estimated deaths (UNAIDS estimates)



Health: Risk Factors

HIV cont.,

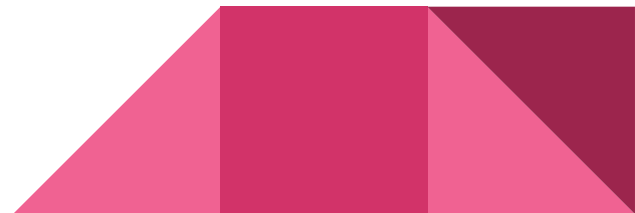
Women's share of population ages 15+ living with HIV (%)

Prevalence of HIV, total (% of population ages 15-49)

Children (0-14) living with HIV

Prevalence of HIV, female (% ages 15-24)

Prevalence of HIV, male (% ages 15-24)



Health: Risk Factors

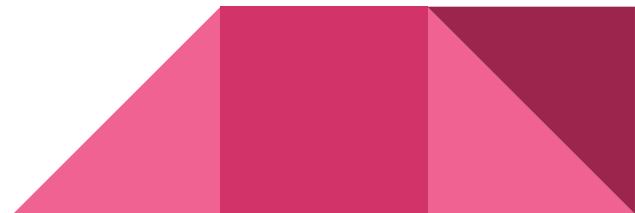
HIV cont.,

Adults (ages 15+) newly infected with HIV

Children (ages 0-14) newly infected with HIV

Adults (ages 15+) and children (ages 0-14) newly infected with HIV

Incidence of HIV (% of uninfected population ages 15-49)



Health: Risk Factors

HIV cont.,

Children orphaned by HIV/AIDS

Antiretroviral therapy coverage for PMTCT (% of pregnant women living with HIV)

Antiretroviral therapy coverage (% of people living with HIV)

Adults (ages 15+) and children (0-14 years) living with HIV



Health: Universal Health Coverage

These 13 indicators were collected during 3 different studies

They are populated only for 1 to 4 years, depending on the country and indicator

The most recent year for information is 2015

These indicators will not be used in the evaluation



Data Cleansing: Country/Region

The data is collected by several different organizations and is not always collected from all of the same countries

The data is sometimes collected by country, inclusive of its territories and sometimes collected with the territories identified separately

There are 41 country codes that are used to aggregate data in different ways and the data associated with those codes represent duplication of data that is included at a more specific level

The data related to these codes will be removed from the evaluation set



Data Cleansing: Target

There are 18 country codes with varied data collected in most years and without data for the target (life expectancy)

The data for these 18 country codes is removed from evaluation

The remaining data includes as many as 214 data features completed for a given country code and as few as 5 out of the total 219 potential features

Depending on results from analyses, additional cleansing may be needed



Data Analysis

At this point the dataset still contains a great deal of missing values

Additional analysis and review shows that the year 2017 is actually missing a great deal of data, especially health data, and this indicates it is not fully current

Reduce the years in the analysis by removing 2017

Note: the predictions will need to be caveated as to the lack of more current information



Data Analysis

Using the feature GNI (gdp replacement), 12 additional countries are removed for scarce and sparse data

9 countries warrant additional analysis

- GNI might be found using alternate sources for missing years

- GNI may be interpolated from other years' data

- Other missing data may be derivable from other available data



Data Analysis

Identify nulls found within features and eliminate features that lack sufficient data for evaluation


This evaluation considers how much of the feature is missing overall

- All with over 60% missing flagged for removal

 - 68 features removed

- All with less than 20% missing flagged for derivation procedures,

- Remaining to be considered against expected usefulness of feature and ease of deriving missing values



Data Analysis and Cleansing

127 features considered and reviewed for derivation of missing data

Population data in Environment category divides the population between Urban and Rural and includes the proportion between the two as well as year over year change

56 of the 61 missing values in Rural year over year change are attributed to 100% of the population reported as Urban and would result in a divide by zero error -- these are manually set to zero



Data Analysis and Cleansing

The remaining 5 missing values are for one country and are the most recent 5 years

Using the basic least squares method of projection, the urban and rural populations are projected

The values for percent of rural to urban and the year over year growth rates are calculated from the population projections. Projections are consistent with trends of decreasing rural population while modest overall growth




Data Analysis and Cleansing

Labor force data in the social category are missing for all years for 6 country codes

The data for total labor force and % of females in labor force are missing for years 2012 to 2016 for Eritrea

The labor force grew at about 1% in each of the previous 10 years and the female percentage increased at just under 1%

2012 and 2016 are filled continuing that pattern with a 1% and 0.99% year over year increase, respectively



Data Analysis and Cleansing

The 6 countries missing the labor force data entirely will be kept in the main data set initially

Analysis with and without the features will determine whether the countries will be removed or the features ignored or some combination

Aruba has been identified as missing data in some health areas and has the most inconsistent data measures of the 6 and may be removed later in the analysis even if the other 5 remain



Data Cleansing

In the education category:

Primary and secondary gross enrollments had the greatest amount of available data

Missing data in these features was derived using a variety of methods applied based on the available data and is separately documented

Tertiary (post-secondary) enrollment features with missing values were also derived



Data Cleansing

Countries removed due to lack of available data in the education category are:

Haiti

Micronesia

Bosnia



Data Cleansing

This leaves the health category which has several sub-categories

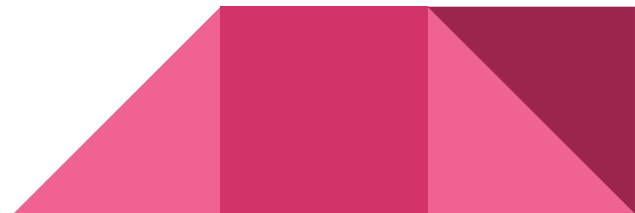
In the health, mortality subcategory there are 3 measures that are recorded every 5 years with annual measures recorded starting in 2016

- Death from unintentional poisoning

- Death from suicide

- Death from CVD, cancer, diabetes or CRD before 70

Each is measured in total and by gender



Data Cleansing

To provide consistent number of years to evaluate for the models

Features will be filled from most recent measure

2003 - 2004 will be the 2005 measure

2006 - 2009 will be the 2010 measure

2011 - 2014 will be the 2015 measure



Data Cleansing

There are 5 countries with no metrics for these categories but for which other data makes them desirable to keep in the overall analysis

Fill the data for these 4 countries with the overall average in these measures

Hong Kong China

Macao China

Puerto Rico

West Bank/Gaza

Aruba is removed at this point as missing data in education and social too



Identifying best features

We now have 35 features, excluding year and the 3 targets, for 182 country codes

- 6 sets of features have both a total and one or more gender-specific measures
- 1 set of features measures rural and urban population 3 different ways



KBest Top Ranked Features

The first method applied is Select KBest from ScikitLearn module where the features that have the highest f-value are identified

The top 7 features for predicting life expectancy without regard to gender are:

- School enrollment, secondary (% gross)
- School enrollment, secondary, female (% gross)
- School enrollment, secondary, male (% gross)
- School enrollment, tertiary (% gross)
- Rural population (% of total population)
- Urban population (% of total population)
- Birth rate, crude (per 1,000 people)



KBest Top Ranked Features

The next top 8 features are:

Adolescent fertility rate (births per 1,000 women ages 15-19)

Fertility rate, total (births per woman)

Mortality rate attributed to unintentional poisoning (per 100,000 population)

Mortality rate attributed to unintentional poisoning, female (per 100,000 female population)

Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)

Tuberculosis death rate (per 100,000 people)

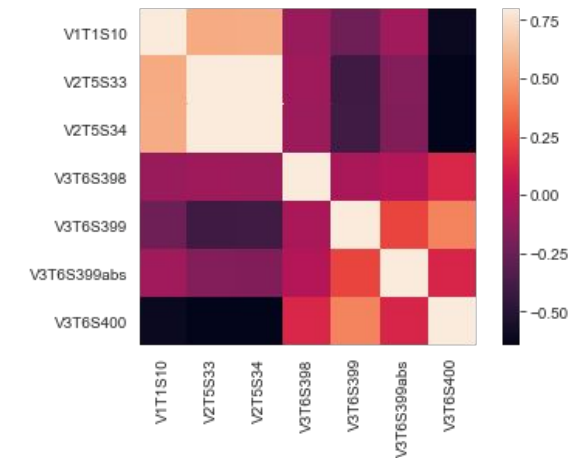
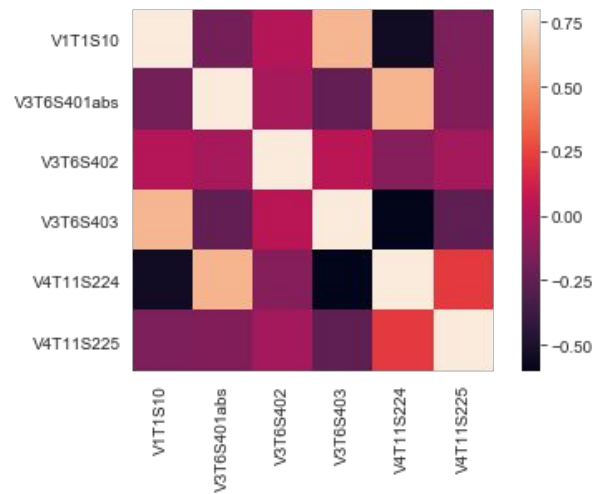
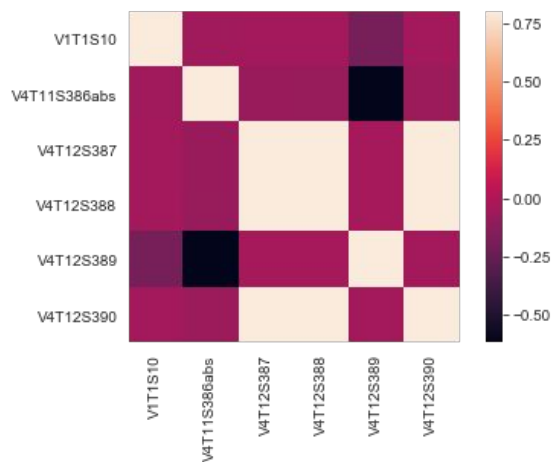
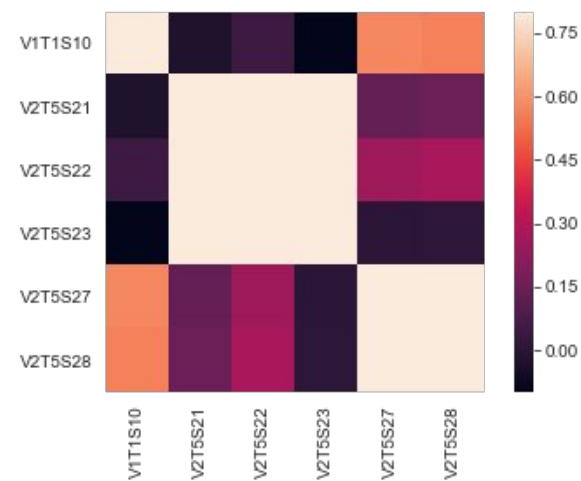
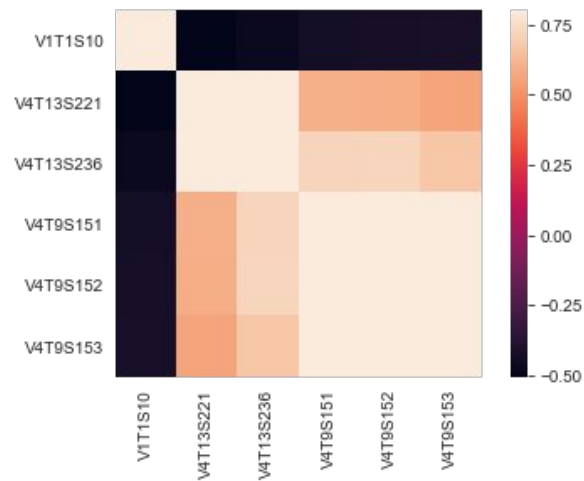
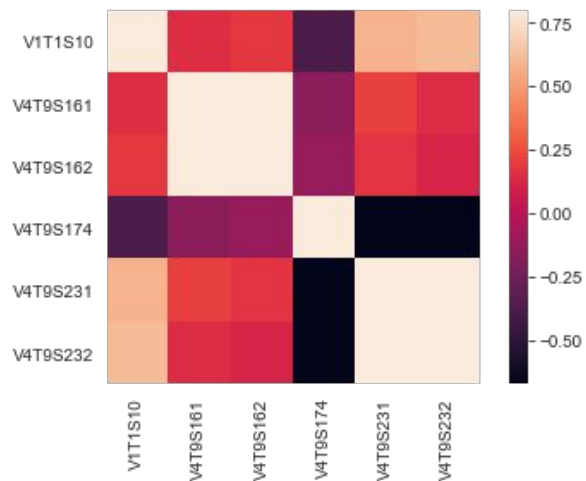
Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%)

Top Correlated Features

After applying the correlation matrix method and viewing results in a heatmap, the top directly correlated features are:

- School enrollment, secondary (% gross)
- School enrollment, secondary, male (% gross)
- School enrollment, tertiary (% gross)
- Urban population (% of total population)
- GNI per capita, Atlas method (current US\$)





Top Correlated Features

The top inverse correlated features are:

Birth rate, crude (per 1,000 people)

Fertility rate, total (births per woman)

Mortality rate attributed to unintentional poisoning, female (per 100,000 female population)

Mortality rate attributed to unintentional poisoning (per 100,000 population)

Adolescent fertility rate (births per 1,000 women ages 15-19)

Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%)

Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)

Tuberculosis death rate (per 100,000 people)

Rural population (% of total population)



Picking the features

The two different filter methods produced a good deal of overlap in features

Still, further refinement is appropriate

- Features specific to gender when the total for the same feature set is identified will be dropped

- Features for rural will be dropped when it is duplicative of an urban measure in the same feature set

This leaves 9 total features with 8 of them being top results in both methods



About the features

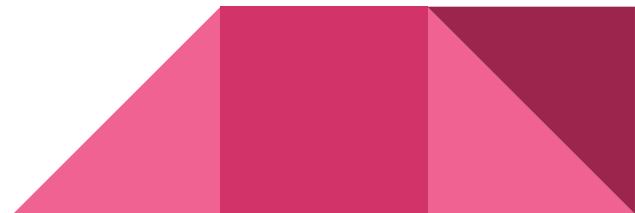
These 4 features have a direct correlation with the target

- School enrollment, secondary (% gross)

- School enrollment, tertiary (% gross)

- Urban population (% of total population)

- GNI per capita, Atlas method (current US\$)



About the features

These 5 features have an inverse correlation to Life Expectancy

- Birth rate, crude (per 1,000 people)

- Mortality rate attributed to unintentional poisoning (per 100,000 population)

- Adolescent fertility rate (births per 1,000 women ages 15-19)

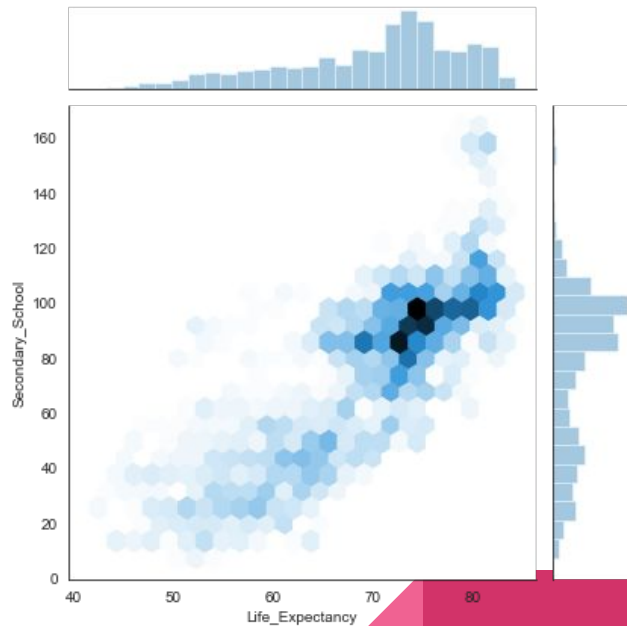
- Tuberculosis death rate (per 100,000 people)

- Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%)



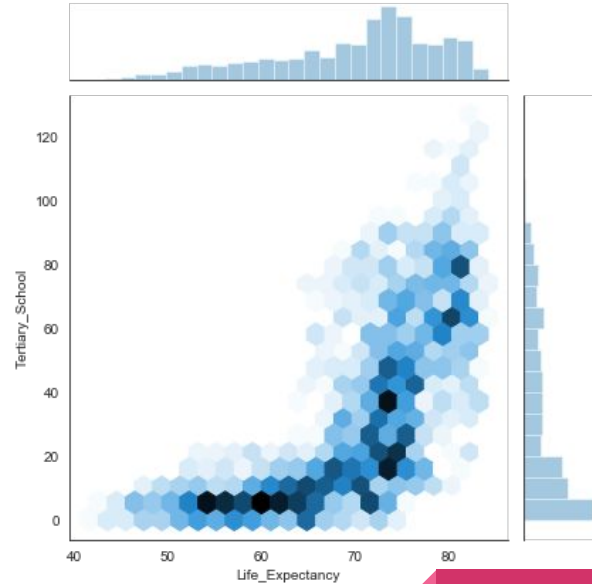
School enrollment, secondary (% gross)

Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education. Secondary education completes the provision of basic education that began at the primary level, and aims at laying the foundations for lifelong learning and human development, by offering more subject- or skill-oriented instruction using more specialized teachers.

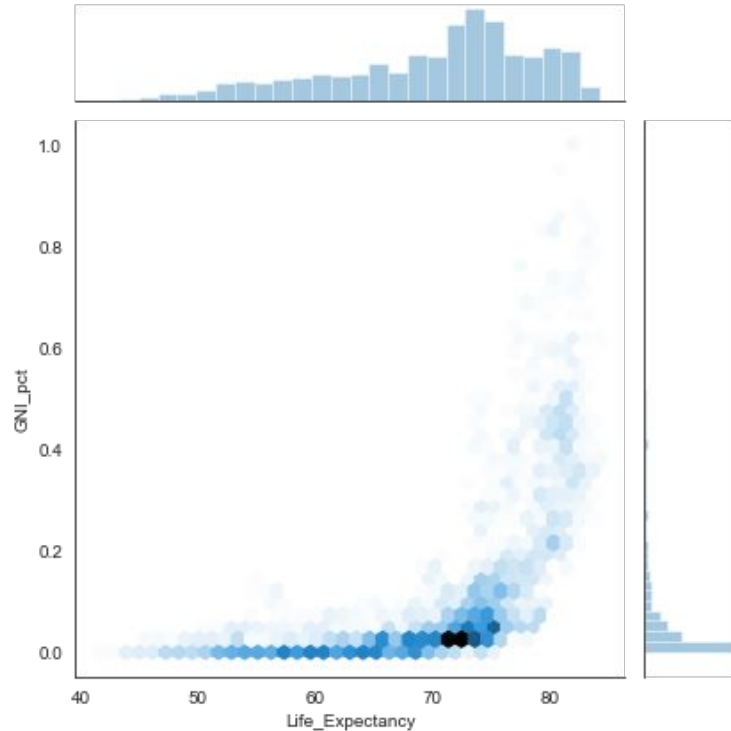


School enrollment, tertiary (% gross)

Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education. Tertiary education, whether or not to an advanced research qualification, normally requires, as a minimum condition of admission, the successful completion of education at the secondary level.

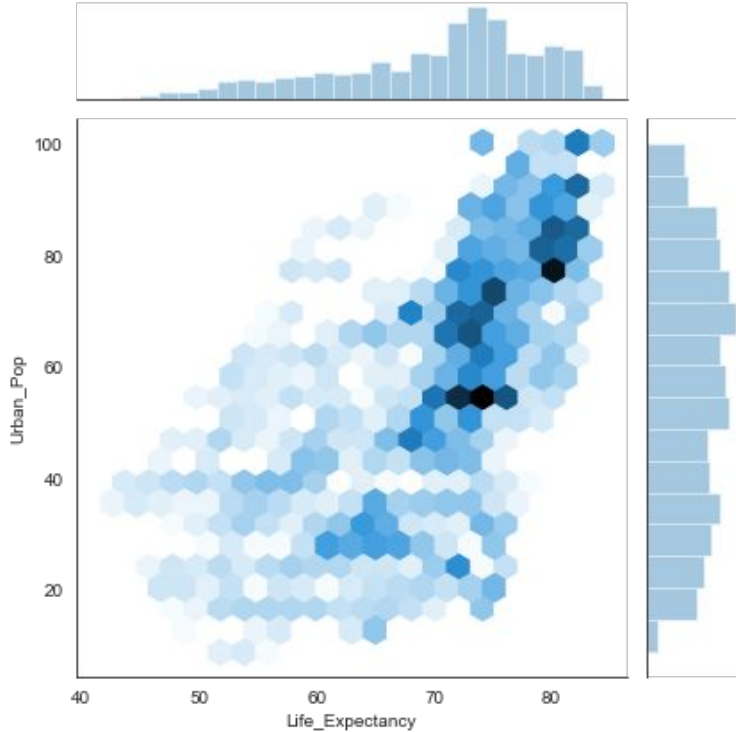


GNI per capita, Atlas method (current US\$)



GNI per capita (formerly GNP per capita) is the gross national income. GNI, calculated in national currency, is usually converted to U.S. dollars using the World Bank Atlas method, divided by the midyear population. GNI was normalized by ranking each country's GNI to total GNI.

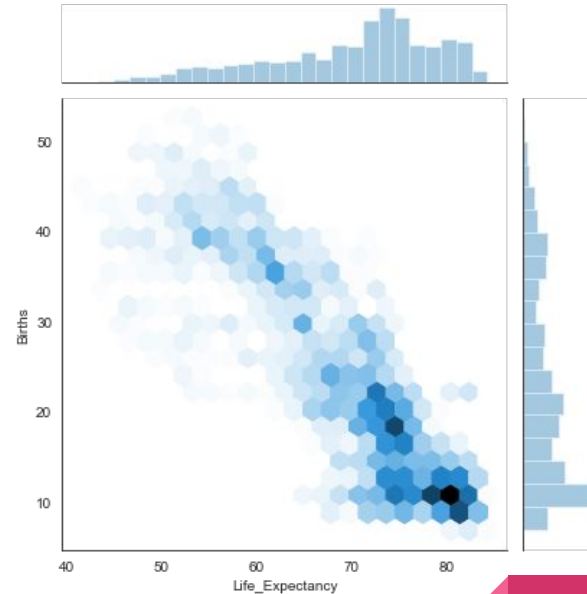
Urban population (% of total population)



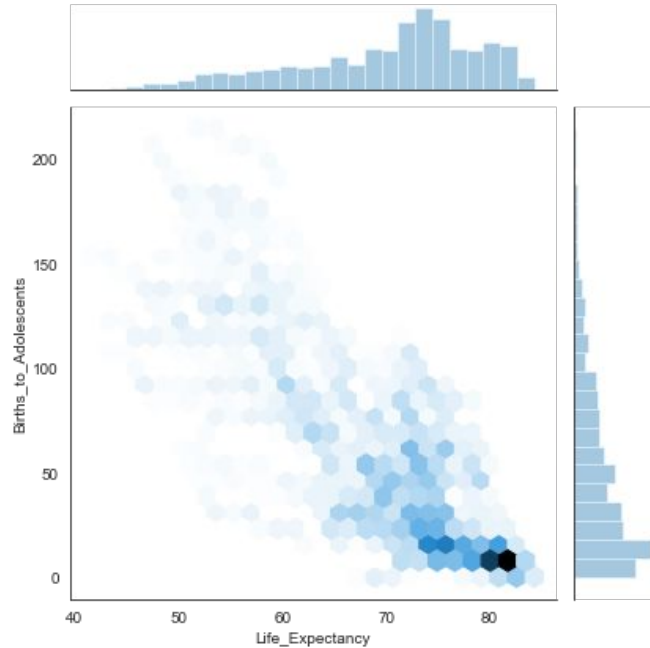
Urban population refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division.

Birth rate, crude (per 1,000 people)

Crude birth rate indicates the number of live births occurring during the year, per 1,000 population estimated at midyear. Subtracting the crude death rate from the crude birth rate provides the rate of natural increase, which is equal to the rate of population change in the absence of migration.

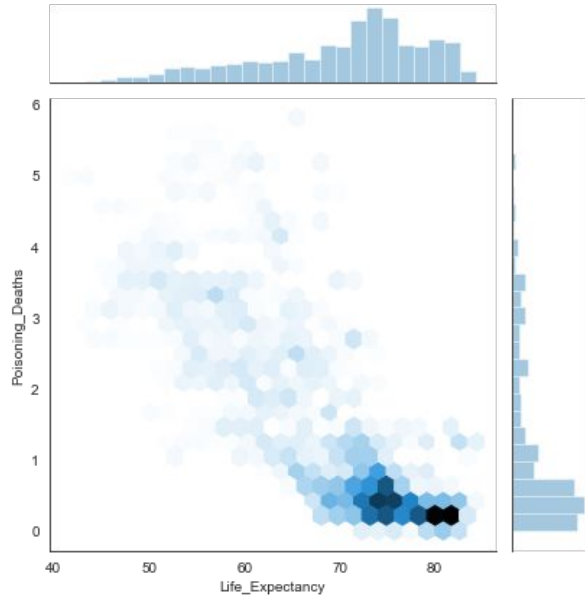


Adolescent fertility rate (births per 1,000 women ages 15-19)



Adolescent fertility rate is the number of births per 1,000 women ages 15-19.

Mortality rate attributed to unintentional poisoning (per 100,000 population)

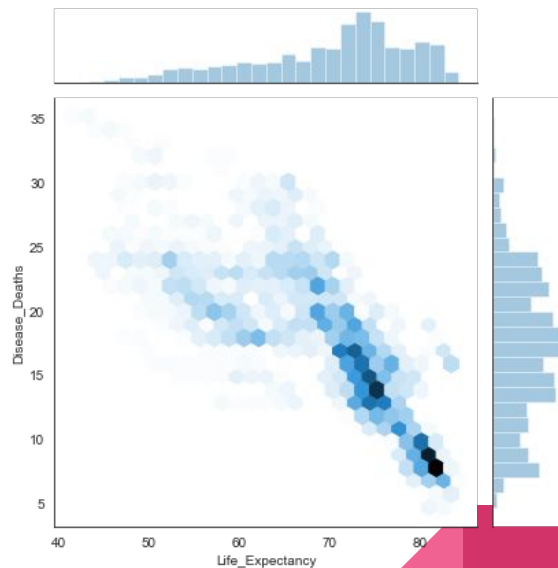


Mortality rate attributed to unintentional poisonings is the number of deaths from unintentional poisonings in a year per 100,000 population. Unintentional poisoning can be caused by household chemicals, pesticides, kerosene, carbon monoxide and medicines, or can be the result of environmental contamination or occupational chemical exposure.

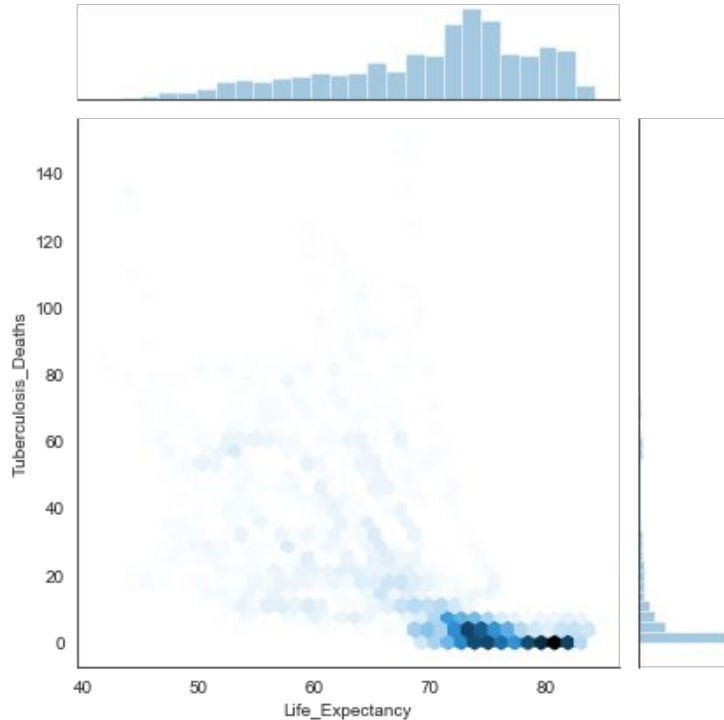
Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%)

Mortality is the percent of 30-year-old-people who would die before their 70th birthday from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease

-- excludes mortality from other cause of death (e.g., injuries or HIV/AIDS)



Tuberculosis death rate (per 100,000 people)



Tuberculosis death rate is the estimated number of deaths from tuberculosis among HIV-negative people, expressed as the rate per 100,000 population. Estimates for all years are recalculated as new information becomes available and techniques are refined, so they may differ from those published previously.

Feature Adjustments

To make GNI more comparable to the other features, it will be converted to a percentage or ratio to the whole

Most values then will be between 0 and 100 although some school enrollments are greater than 100 when people in the population outside of the typical school-level age enroll and unintentional poisoning rates range into the 200s

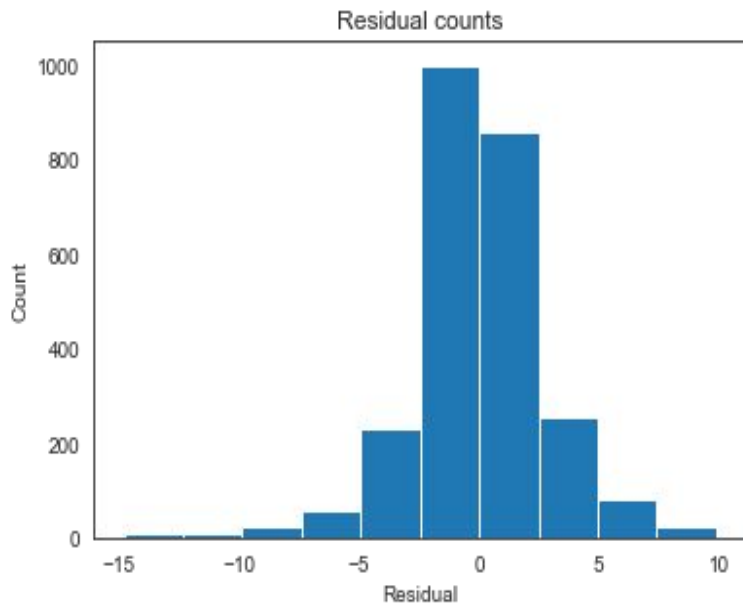
GNI will be expressed as percentage GNI to the highest GNI where the highest GNI will represent 100 percent



Will these features predict life expectancy?



Basic Linear Regression

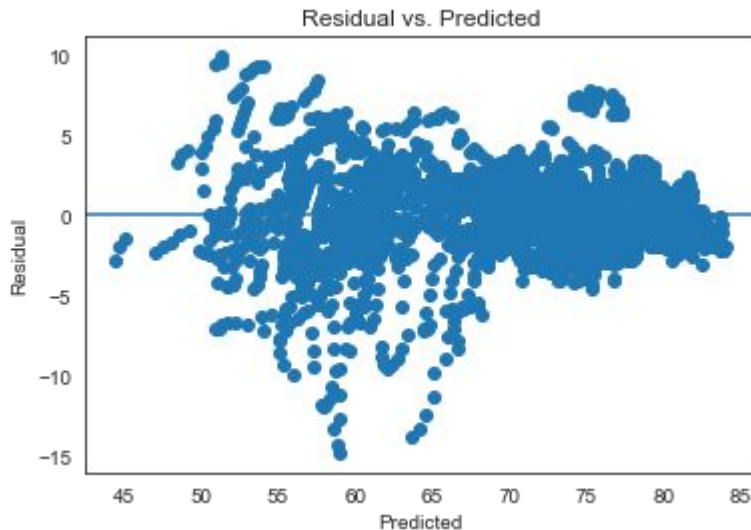


A basic linear regression model with the 9 features produces a good R-square value around .90. It's high enough to show good explanation of variances and not so high as to imply overfitting.

This error graph is reasonably close to a normal distribution for the error. Outliers are more on the negative side.

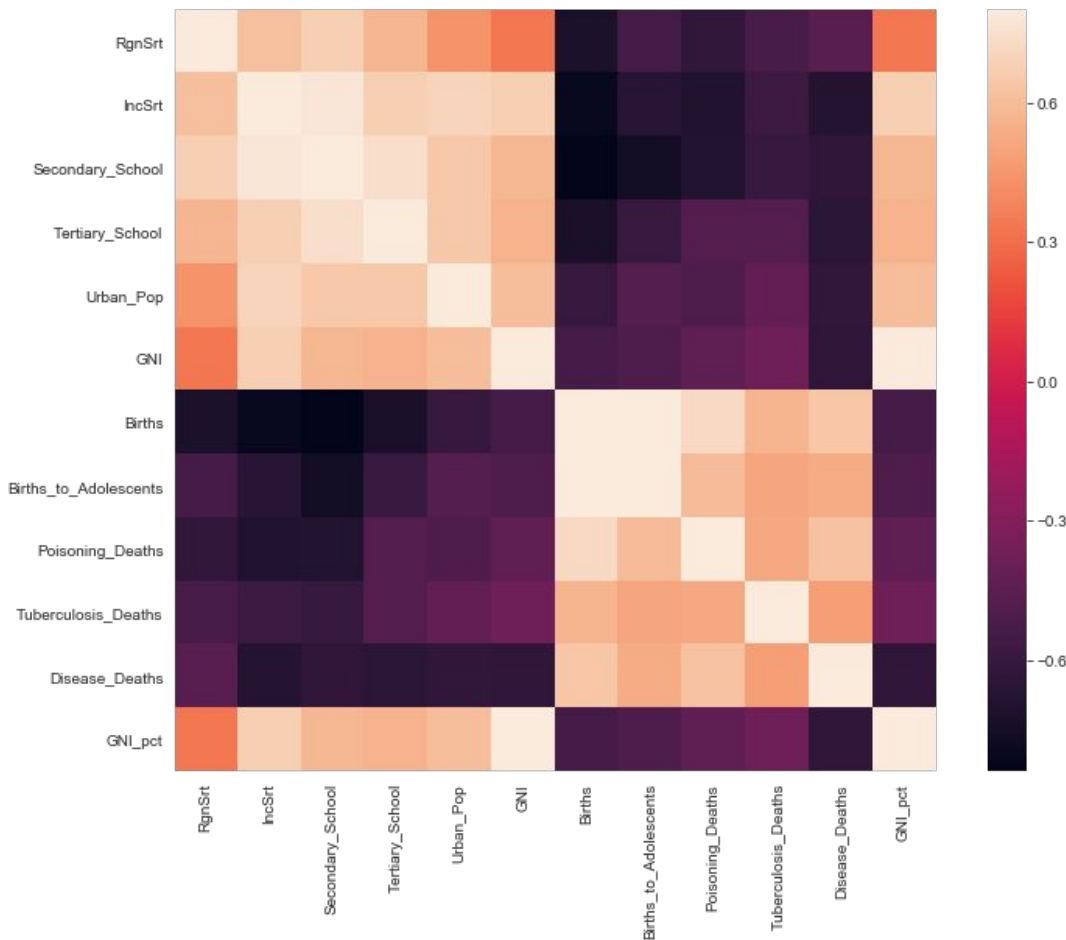
Basic Linear Regression

A scatterplot of what our model predicts versus the actual age recorded is informative. Our model predicts older ages with the most consistency. The top 3rd of ages is in the ± 5 interval. The middle section shows the highest error rate with the outliers leaning to a lower life expectancy.



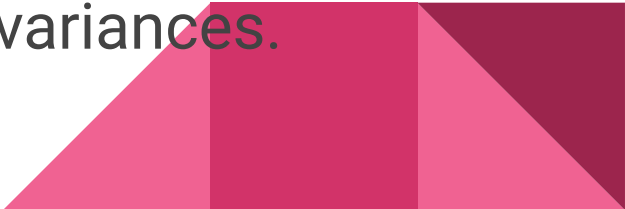
Correlation

There is fairly good correlation, direct or inverse, in these features which unfortunately implies they are all explaining the same variance and that any given one of them would produce substantially the same result as all of them.

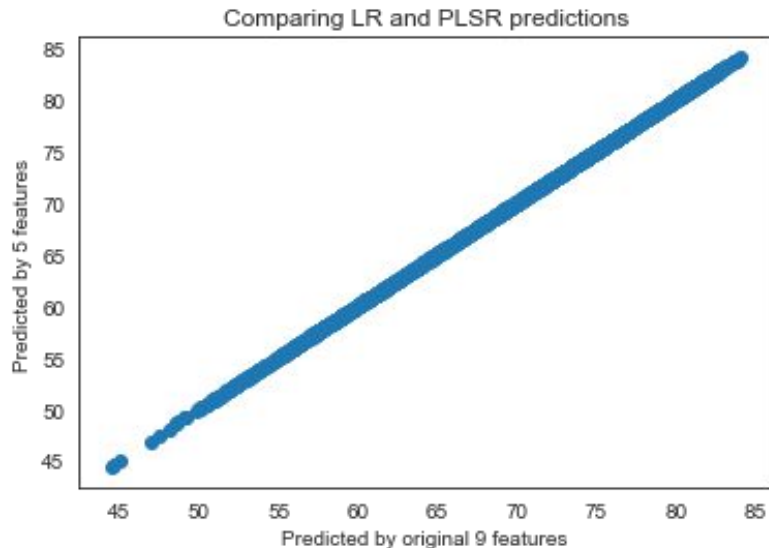


Correlation

Note that earlier correlation was used to identify top predictors of the target from the features. The goal is to find the best predictors with minimal overlap in prediction. In other words, the best case is to have features that are highly correlated to the target while being uncorrelated to one another. This means the features contribute different aspects to the target and support/explain different variances.



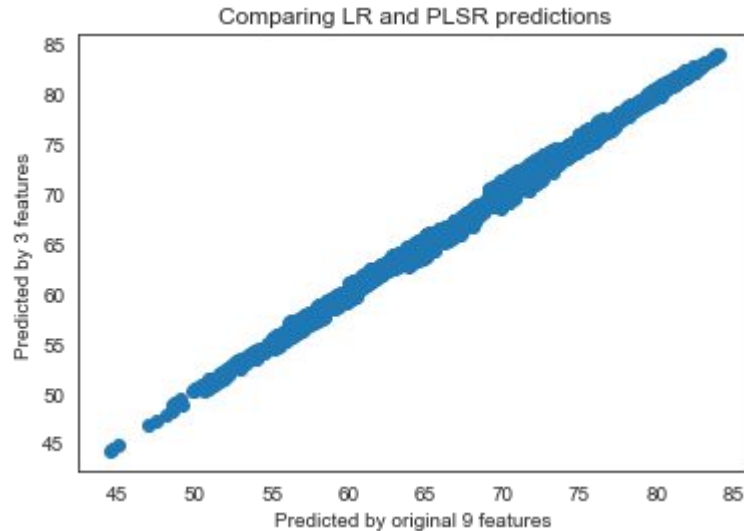
PLSR



After applying partial least squares regression using 5 features (in an attempt to remove duplicative predictors), the result is so similar to basic regression with all 9 features as to be practically the same result. The R-squared values are within 1/100,000th of one another.

PLSR

The next step was to reduce to 3 features and this did give a small shift.



Reduce to 5 features to reduce overfitting

Using the correlation matrix, selected the 5 features with the least correlation to one another.

Urban Population

GNI

Births to Adolescents

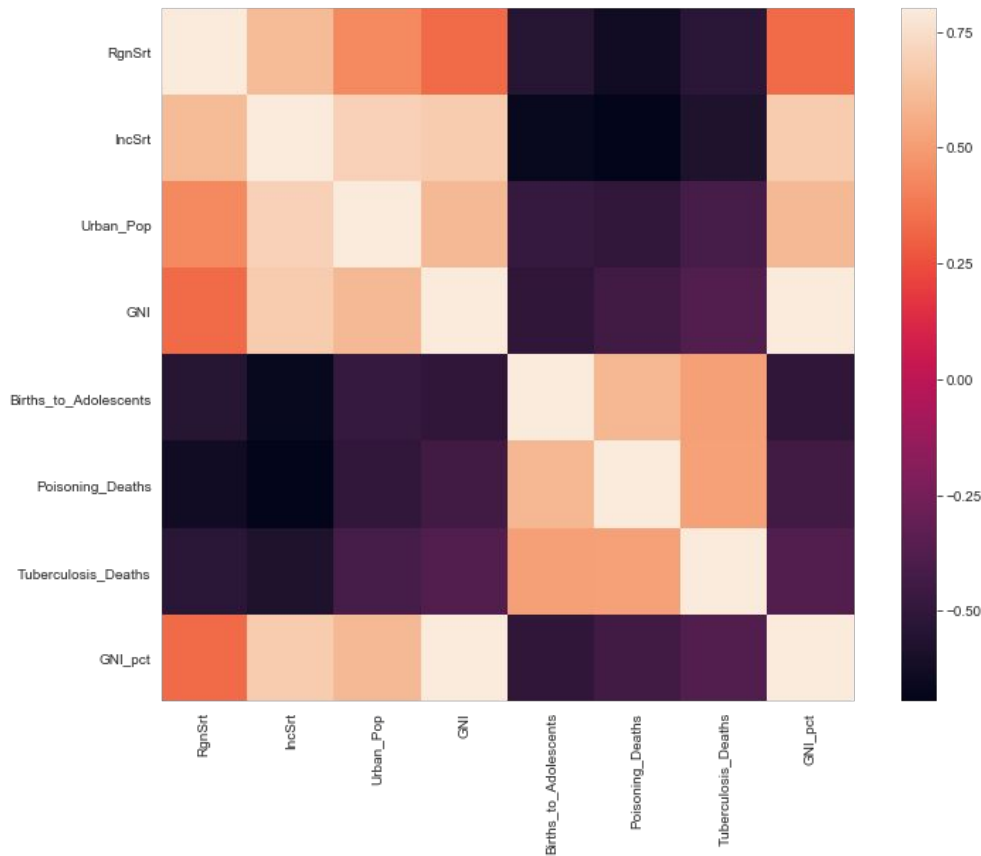
TB Deaths

Accidental Poisoning Deaths



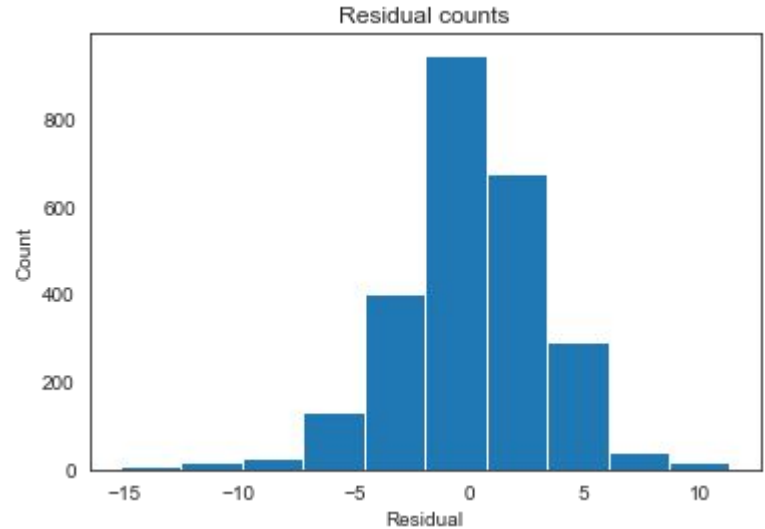
This looks a lot better for explaining different variances.

Even though GNI and Urban population have strong inverse correlation to the other features they appear to have low correlation to one another implying the explanation of different variances.



Linear Regression, Again

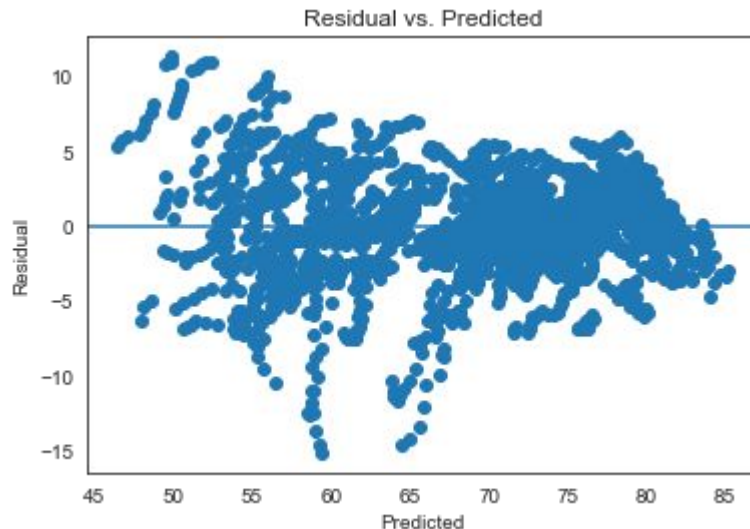
The error distribution is very similar to the original. The R-squared has reduced to about .86 which is still a tolerable result.



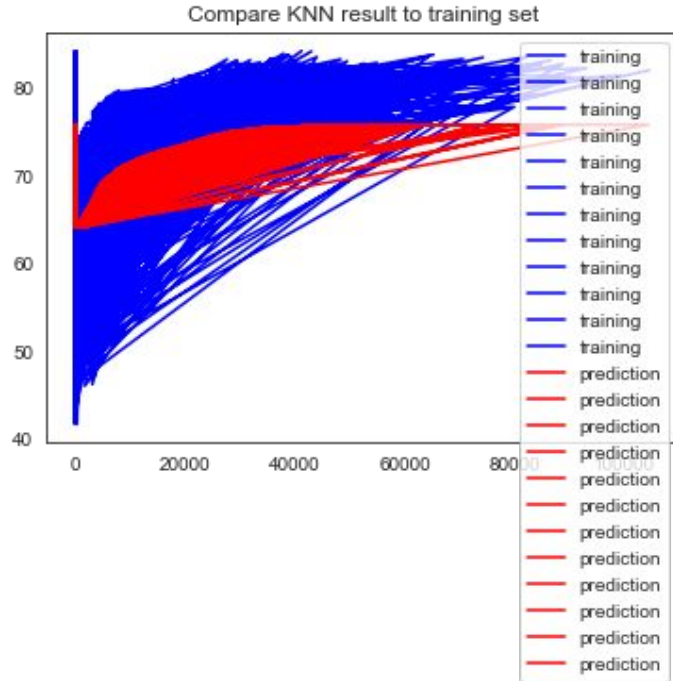
Linear Regression Prediction vs Actual

The predictive power appears to be reduced as well. The results are not as tightly clustered as when using all 9 features. In particular older ages are not as well predicted as the original. And younger ages are being predicted more frequently as an older age.

Going to return to data set with 9 features.



KNN Regression



It's not pretty but it is clear.


KNN is not a good solution for this.

Lasso and Ridge results

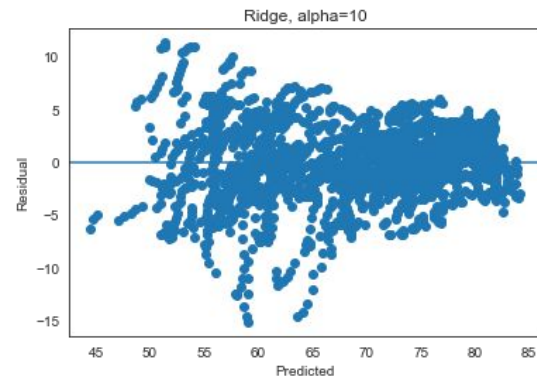
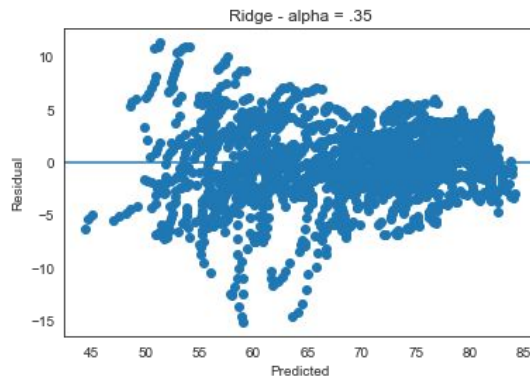
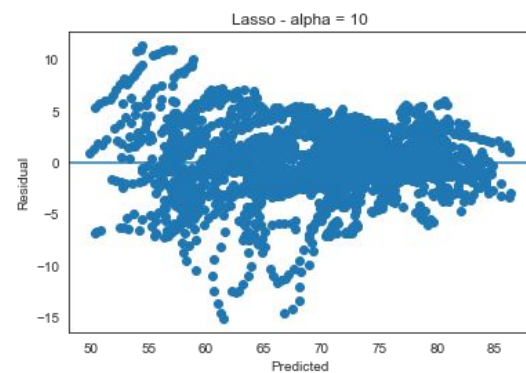
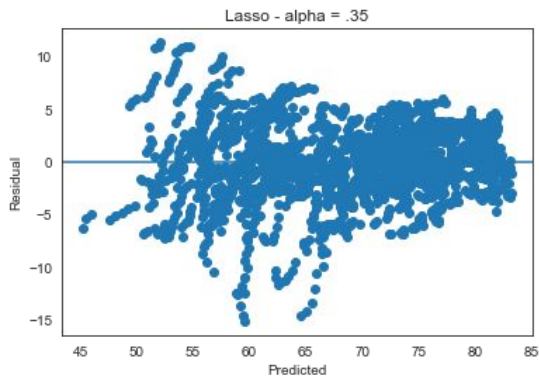
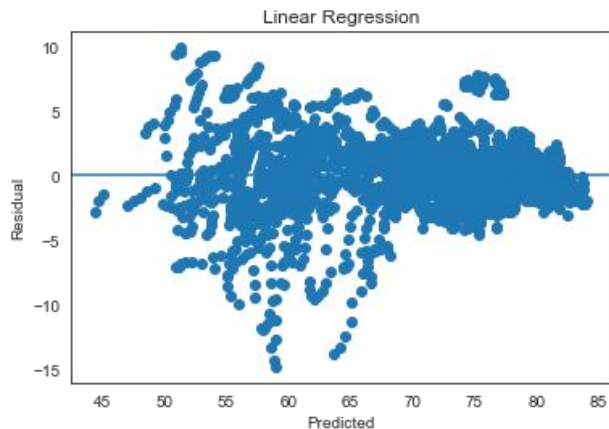
The results when using a lasso regression model are most similar to basic linear regression when a larger lasso is used. The smaller the number of features to lasso, the worse the R-squared becomes.

Ridge regression also produces substantially the same error rates and fit. The increase/decrease of the alpha does not produce different results.

Because all three main regression models produce substantially the same result (depicted on next slide), basic linear regression will be used due to its more common acceptance and understanding.



Comparison of Regression Model Results



Support Vector Machine

Last, but not least, Support Vector Machine was set to the task of predicting life expectancy from the nine features.

The test predictions compared to the actual ages for the data are presented here.

Linear regression is the clear winner for this task.

