Cal Colistra Github link:

Introduction to Data Science

Project 3

2/28/2022

Purpose of Project 3:

Previously there have been various studies that aimed to identify certain factors that affect life expectancy. These studies considered demographic variables, income composition and mortality rates. Also, some of these studies were only based on the data from one year. The purpose of this project is to take into account more factors such as the effect of immunization and human development index. The data set provided for this project considers data from a period of 2010 to 2015 for all countries. Unlike the past studies, this project focuses on mortality factors, economic factors, social factors and other health related factors. Since this dataset considers different countries over a span of 5 years, it will be easier to determine which factors are contributing to life expectancy.

Methodology:

The data set was provided by the World Health Organization and was found under The Global health Observatory data repository. The data consists of 11 columns and 1108 rows for 193 countries. To cleanse the data I deleted rows where population = 0 because it wouldn't make sense to consider data of a country with a population of 0. Next, I decided to fill in where countries had 0 total expenditure and 0 percentage expenditure. I use the average of the columns that had a total expenditure and percentage expenditure greater than 0. Next, I deleted 7 rows where GDP was 0. After this I noticed there were a couple rows that had multiple values equaling 0. Palau had life expectancy, adult mortality and alcohol equaling 0 and Tuvalu had life expectancy and adult mortality equaling 0. After data cleansing the total number of countries in the data set became 847.

Summary:

List of Countries with **lowest** average mortality rates:

Country	format(avg(Adult_Mortality), 2)
Tunisia	10.67
Italy	31.50
Iceland	35.33
Israel	42.17
Montenegro	48.50
Kiribati	49.83
Spain	51.33
Netherlands	51.67
Australia	52.33
Switzerland	53.00
Myanmar	53.83
Cyprus	55.17
Croatia	55.50
Sweden	56.17
Canada	56.50
Ireland	56.83
Malta	57.50
Luxembourg	58.00
Japan	58.67

List of Countries with **highest** average mortality rates:

Country	format(avg(Adult_Mortality),
Lesotho	436.00
Central African Republic	435.50
Zimbabwe	421.00
Swaziland	409.50
Nigeria	366.00
Chad	363.33
Sierra Leone	362.67
Angola	353.67
Malawi	346.83
South Sudan	346.83
South Africa	322.50
Cameroon	310.67
Zambia	292.17
Eritrea	292.00
Haiti	287.00
Uganda	283.67
Guinea-Bissau	282.83
Papua New Guinea	282.80
Liberia	274.17

List of Countries with **lowest** average population:

Country	format(avg(Population)
Sri Lanka	2,522.83
Maldives	8,454.33
Georgia	9,383.33
Kiribati	14,193.50
Tonga	15,189.83
Israel	43,490.33
Seychelles	63,329.67
Samoa	104,704.83
Japan	106,435.17
Sao Tome and Principe	127,502.83
Iceland	130,239.33
China	135,528.33
Vanuatu	171,856.83
Albania	199,472.67
Uzbekistan	209,960.17
Suriname	215,914.83
Luxembourg	216,144.17
Slovenia	218,044.50
Montenegro	229,739.17

List of Countries with **highest** average population:

Country	format(avg(Population)
India	281,099,848.50
Indonesia	175,751,726.33
Pakistan	123,341,557.50
Nigeria	90,382,108.33
Brazil	70,799,653.83
Bangladesh	59,971,506.33
Russian Federation	55,182,376.17
Turkey	51,513,111.50
Ethiopia	49,418,339.50
Mexico	42,514,971.50
Algeria	37,971,074.17
Philippines	36,966,300.00
Argentina	36,019,452.00
Colombia	33,247,267.83
Italy	30,963,837.83
Kenya	30,465,692.17
South Africa	29,322,660.17
France	26,338,424.83
Sudan	25,267,513.83

List of Countries with **lowest** average GDP:

Country	format(avg(GDP),
Senegal	94.40
Burundi	192.02
Sierra Leone	298.39
Niger	322.34
Guinea	353.51
Madagascar	377.95
Malawi	399.30
Ethiopia	405.60
Haiti	413.83
Liberia	414.58
Central Afri...	415.03
Guinea-Bissau	434.25
Rwanda	449.40
Mozambique	454.98
Togo	467.78
Zimbabwe	477.77
Afghanistan	519.25
Eritrea	532.46
Cambodia	542.29

List of Countries with **highest** average GDP:

Country	format(avg(GDP),
Switzerland	70,817.14
Luxembourg	63,600.08
Australia	61,393.17
Austria	40,276.87
Netherlands	34,698.75
Denmark	33,730.66
Canada	33,583.38
Israel	29,932.44
Ireland	27,964.95
Finland	26,069.41
Iceland	24,840.91
Italy	24,713.54
Sweden	23,298.50
France	21,954.84
Greece	19,802.19
Norway	18,417.94
Germany	16,925.93
Japan	16,718.99
Spain	15,505.83

List of Countries with **lowest** average Schooling:

Country	format(avg(Schooling),
South Sudan	4.08
Eritrea	5.05
Niger	5.07
Djibouti	6.10
Central African Republic	6.97
Sudan	7.03
Chad	7.10
Burkina Faso	7.18
Pakistan	7.70
Mali	7.85
Mauritania	8.05
Ethiopia	8.32
Guinea	8.52
Senegal	8.53
Haiti	8.92
Myanmar	9.05
Guinea-Bissau	9.08
Sierra Leone	9.17
Equatorial Guinea	9.20

List of Countries with **highest** average Schooling:

Country	format(avg(Schooling), 2)
Australia	20.08
Iceland	18.78
Ireland	18.50
Denmark	18.20
Netherlands	17.77
Norway	17.58
Spain	17.22
Argentina	17.17
Slovenia	17.07
Finland	16.95
Germany	16.90
Greece	16.90
Lithuania	16.57
Estonia	16.43
Italy	16.42
Portugal	16.37
Belgium	16.23
France	16.15
Canada	15.97

List of Countries with **lowest** average alcohol consumption:

Country	format(avg(Alcohol), 2)
South Sudan	0.00
Bangladesh	0.01
Mauritania	0.01
Afghanistan	0.01
Pakistan	0.03
Comoros	0.05
Niger	0.05
Iraq	0.06
Indonesia	0.07
Guinea	0.08
Bhutan	0.09
Timor-Leste	0.10
Tajikistan	0.11
Kiribati	0.17
Nepal	0.18
Mali	0.21
Senegal	0.23
Vanuatu	0.30
Myanmar	0.32

List of Countries with **highest** average alcohol consumption:

Country	format(avg(Alcohol), 2)
Belarus	12.78
Lithuania	11.82
Austria	10.09
Croatia	10.05
France	9.60
Bulgaria	9.43
Luxembourg	9.39
Ireland	9.39
Portugal	9.34
Germany	9.26
Belgium	9.13
Poland	9.11
Russian Fe...	8.96
Slovenia	8.64
Australia	8.40
Switzerland	8.20
Denmark	8.19
Serbia	7.81
Finland	7.76

<u>Do densely populated countries tend to have lower life expectancy?</u>

Countries with **highest population**:

- India, Indonesia, Brazil, Pakistan, Nigeria, Bangladesh, Russia

Countries with **lowest Life Expectancy**:

- Haiti, Sierra Leone, Central African Republic, Angola, Lesotho, Chad, Nigeria, Zimbabwe

<u>Answer:</u> No, other than Nigeria, the densely populated countries don't match the countries with low life expectancy

Python:

How does adult mortality affect life expectancy?

Adult mortality and life expectancy have a negative correlation (-0.75212). This means that when a country has a high life expectancy, their adult mortality rates will be lower. It also means the opposite: when a country has lower life expectancy, their adult mortality rates will be higher.

Does life expectancy have positive or negative correlation with eating habits, drinking alcohol, social factors, and economic factors?

Correlation table:

	Year	Life_Expectancy	Adult_Mortality	Alcohol	Percentage_Expenditure	BMI	Total_Expenditure	GDP	Population	Schooling
Year	1.000000	0.065634	-0.037776	-0.358063	0.010327	0.048177	0.013623	-0.012964	0.023228	0.071349
Life_Expectancy	0.065634	1.000000	-0.752118	0.399364	0.393558	0.548622	0.237063	0.466719	-0.034120	0.804695
Adult_Mortality	-0.037776	-0.752118	1.000000	-0.217801	-0.250781	-0.412567	-0.138235	-0.299109	0.024284	-0.555038
Alcohol	-0.358063	0.399364	-0.217801	1.000000	0.356947	0.265208	0.237866	0.404765	-0.021096	0.492724
Percentage_Expenditure	0.010327	0.393558	-0.250781	0.356947	1.000000	0.219910	0.277197	0.903099	-0.033152	0.383069
BMI	0.048177	0.548622	-0.412567	0.265208	0.219910	1.000000	0.161114	0.279447	-0.071932	0.552770
Total_Expenditure	0.013623	0.237063	-0.138235	0.237866	0.277197	0.161114	1.000000	0.261385	-0.075150	0.252240
GDP	-0.012964	0.466719	-0.299109	0.404765	0.903099	0.279447	0.261385	1.000000	-0.029179	0.477967
Population	0.023228	-0.034120	0.024284	-0.021096	-0.033152	-0.071932	-0.075150	-0.029179	1.000000	-0.046832
Schooling	0.071349	0.804695	-0.555038	0.492724	0.383069	0.552770	0.252240	0.477967	-0.046832	1.000000

Correlations with life expectancy:

Eating habits(BMI) = .548622 (**Positive** correlation, strength = medium)

Drinking alcohol = .399364 (**Positive** correlation, strength = weak)

Social factors (Schooling) = .804695 (**Positive** correlation, strength = strong)

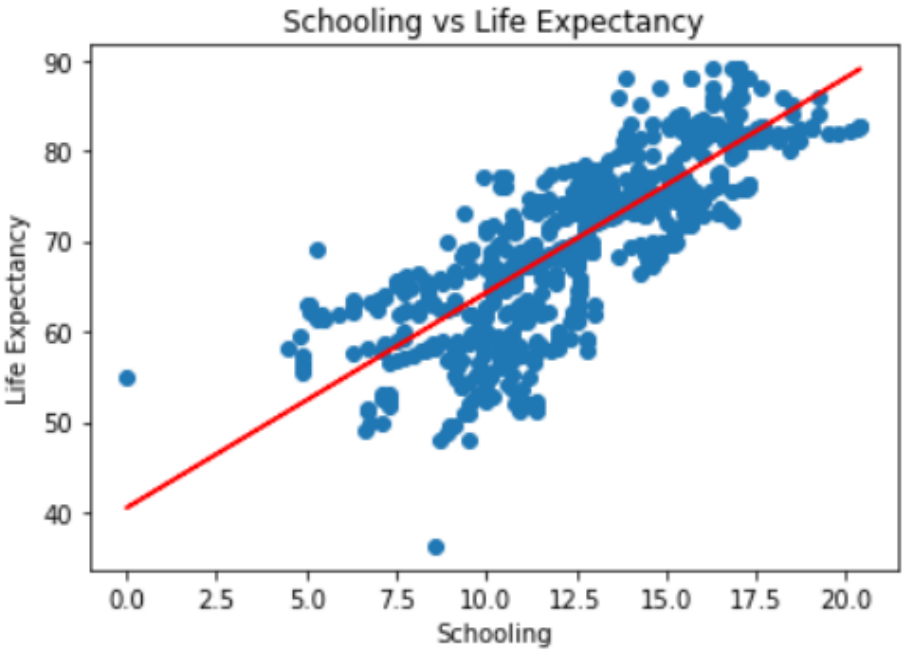
Economic Factors

- Percent Expenditure = 0.393558 (**Positive** correlation, strength = weak)
- Total Expenditure = 0.237063 (**Positive** correlation, strength = weak)

What is the impact of schooling on the lifespan of humans?

- Life expectancy and Schooling have a strong positive correlation (0.8). This means that countries with more schooling have higher life expectancy rates.

Create and **plot** several linear regression models for **life expectancy** (as the **dependent variable)** and the independent variables that had the highest correlation with life expectancy as found in your correlation analysis.

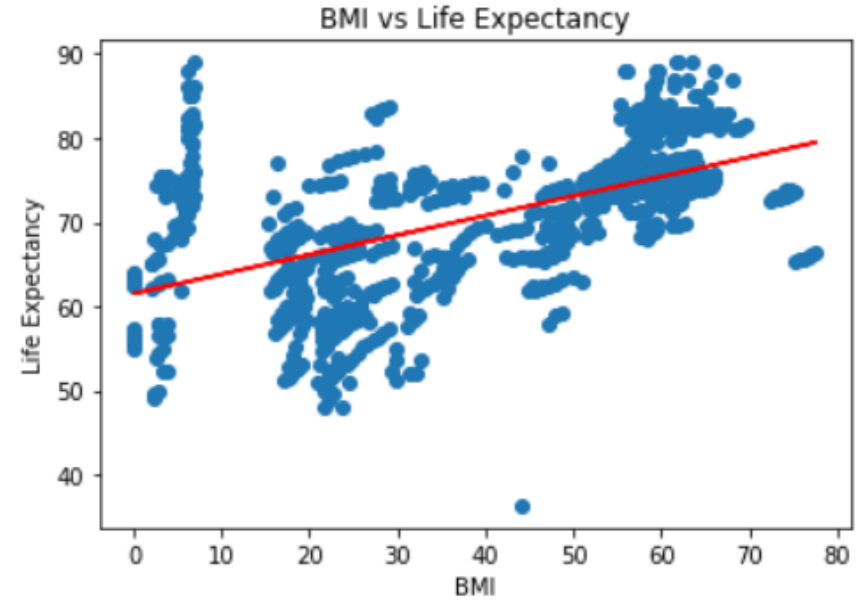


MSE: 27.99

R2: 0.6475

Equation:

y = (2.38)x + 40.47

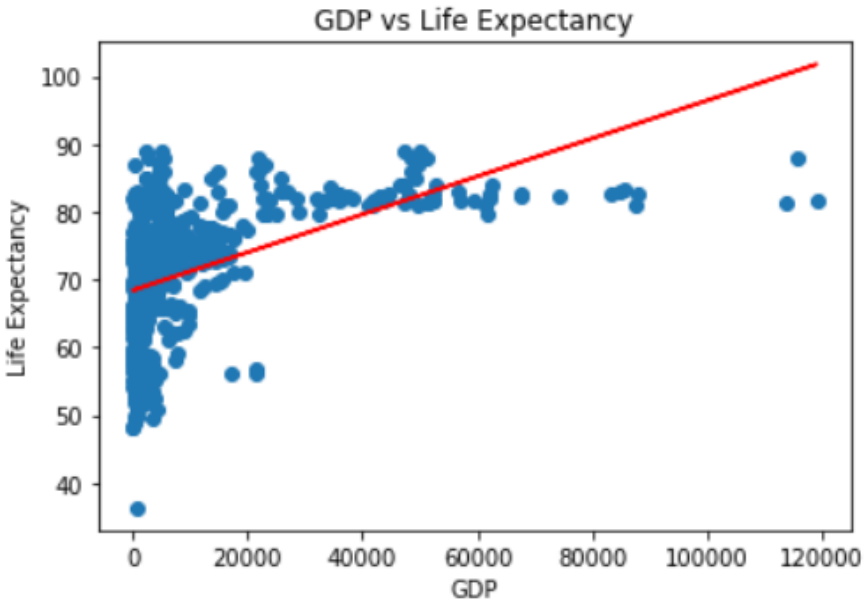


MSE: 55.51

R2: 0.3

Equation:

y = (0.232)x + 61.5



MSE: 62.12

R2: 0.2178

Equation:

y = (0.00027)x + 68.4

Which model performs the best?

- When trying to figure out which factors had the most significant effect on life expectancy I chose to compare schooling, BMI, and GDP because these had the strongest correlation with life expectancy. The model that performed best is the "Schooling and Life Expectancy" model because, when compared to BMI and GDP, it has the lowest mean square error and the highest R2 (coefficient determination).

Conclusion:

The predicting factors that were previously chosen such as demographic variables and income composition don't seem to really affect life expectancy. But, mortality rates (also a previously chosen factor) does seem to at least have a strong correlation with life expectancy. The correlation between life expectancy and mortality rates is -0.75. This means if a country has high mortality rates, they are more likely to have low life expectancy and vice versa. This makes sense because a high mortality rate means that the country has a high death rate and this could be considered as the opposite of a high life expectancy.

Countries with low life expectancy (<65) would most likely improve their average lifespan if they increase healthcare expenditure. Although the correlation between

percent of healthcare expenditure and life expectancy is only .39, it is still a positive correlation and therefore an increase in healthcare expenditure would most likely increase life expectancy over time.

An odd correlation made from the data set is a positive correlation between life expectancy and alcohol consumption. I would expect that this correlation would be negative but according to the data, it is positive. I would expect this because alcohol is known to be bad for one's health and therefore I would think that it would decrease life expectancy rates. It is possible that this positive correlation between life expectancy and alcohol was caused by error in the data, but this is very unlikely because the data came from a well known national organization. Another possibility is that countries with higher alcohol consumption may be more wealthy or developed, and for that same reason their life expectancy is also higher.

My SQL Queries:

```
use ids6db;
desc Life_Expectancy;
select * from Life_Expectancy;

-- Data Cleansing: --
select count(*) from Life_Expectancy; -- 1108 rows prior to deletion
select count(*) from Life_Expectancy where (Population = 0); -- 252 rows with (population = 0)

-- delete rows where (population = 0):
set SQL_SAFE_UPDATES = 0;
delete from Life_Expectancy where (Population = 0);
-- success: 252 row(s) affected

-- fill in where (total_expenditure = 0) with avg(total_expenditure) where (total_expenditure != 0)
select * from Life_Expectancy where (Total_Expenditure < .5);
select avg(Total_Expenditure) from Life_Expectancy where (Total_Expenditure != 0);
-- avg = 6.2
update Life_Expectancy set Total_Expenditure = 6.2
where (Total_Expenditure = 0);
```

```
-- fill in where (Percentage_Expenditure = 0) with avg(Percentage_Expenditure)
where(percentage_Expenditure != 0)
select * from Life_Expectancy where (Percentage_Expenditure = 0);
select avg(Percentage_Expenditure) from Life_Expectancy where (Percentage_Expenditure != 0);
-- avg = 1041.79
update Life_Expectancy set Percentage_Expenditure = 1041.79
where (Percentage_Expenditure = 0);

-- delete countries with 0 GDP
delete from Life_Expectancy where (GDP = 0);
-- 7 rows affected

-- delete Palau because Life expectancy, adult mortality & alcohol = 0
delete from Life_Expectancy where (Country = "Palau");

-- delete Tuvalu because  life expectancy & adult mortality = 0
delete from Life_Expectancy where (Country = "Tuvalu");

-- display total count of countries after data cleansing:
select count(Country) from Life_Expectancy;  -- 847

-- List of countries with the highest and lowest average mortality rates (years 2010-2015):
select Country, format(avg(Adult_Mortality), 2) from Life_Expectancy
group by Country order by avg(Adult_Mortality) asc;
select Country, format(avg(Adult_Mortality), 2) from Life_Expectancy
group by Country order by avg(Adult_Mortality) desc;
select * from Life_Expectancy where (Country = "Palau");

-- List of countries with the highest and lowest average population (years 2010-2015):
select Country, format(avg(Population), 2) from Life_Expectancy
group by Country order by avg(Population) asc;
select Country, format(avg(Population), 2) from Life_Expectancy
group by Country order by avg(Population) desc;

-- List of countries with the highest and lowest average GDP (years 2010-2015):
select Country, format(avg(GDP), 2) from Life_Expectancy
group by Country order by avg(GDP) asc;
```

```
select Country, format(avg(GDP), 2) from Life_Expectancy
group by Country order by avg(GDP) desc;

-- List of countries with the highest and lowest average Schooling  (years 2010-2015):
select Country, format(avg(Schooling), 2) from Life_Expectancy
group by Country order by avg(Schooling) asc;
select Country, format(avg(Schooling), 2) from Life_Expectancy
group by Country order by avg(Schooling) desc;

-- Which countries have the highest and lowest average alcohol consumption (years 2010-2015)?
select Country, format(avg(Alcohol), 2) from Life_Expectancy
group by Country order by avg(Alcohol) asc;
-- Countries with lowest avg alc consumption:
-- Sudan, Bangladesh, Mauritania, Afghanistan, Pakistan
select Country, format(avg(Alcohol), 2) from Life_Expectancy
group by Country order by avg(Alcohol) desc;
-- Countries with highest avg alc consumption:
-- Belarus, Lithuania, Austria, Croatia, France, Bulgaria

-- Do densely populated countries tend to have lower life expectancy?
Select Country, Life_Expectancy, Population
from Life_Expectancy order by (Population) desc;
-- countries with highest population:
-- India, Indonesia, Brazil, Pakistan, Nigeria, Bangladesh, Russia
Select Country, Life_Expectancy, Population
from Life_Expectancy order by Life_Expectancy asc;
-- countries with lowest Life_Expectancy:
-- Haiti, Sierra Leone, Central African Republic, Angola, Lesotho, Chad, Nigeria, Zimbabwe
-- ans: No, except for Nigeria, the densely populated countries don't match the
-- countries with low life_expectancy
```