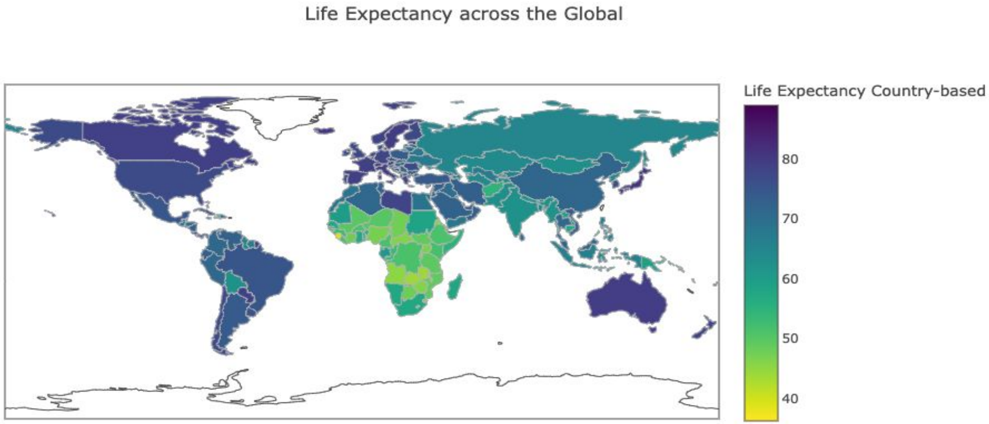
Global Life Expectancy

By Polina Ovchinnikova

Background/Motivation

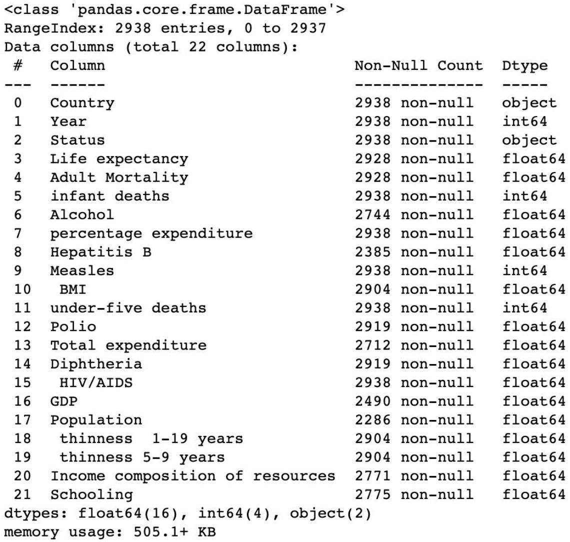
- The online data set is the Life Expectancy (WHO), which keeps track of all countries’ health status and related factors
- Look into how status of a country “Developed” vs “Developing” plays a role in Life Expectancy


Life Expectancy across the Global
Life Expectancy Country-based
80
70
60
50
40

*Notes: the data set is available to the public for health data analysis and downloaded from kaggle

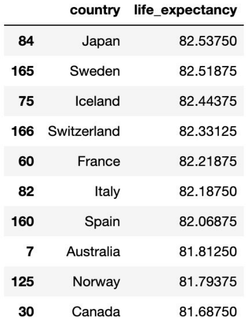
Data Description

- Data is collected from 2000-2015 and from 193 countries
- 22 columns and 2938 rows
- “Developing” countries take up to 82.57% of the total data



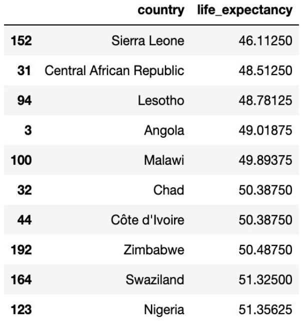
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          2938 non-null   object 
 1   Year                             2938 non-null   int64  
 2   Status                           2938 non-null   object 
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64  
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64  
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64  
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years              2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

Top 10 Best

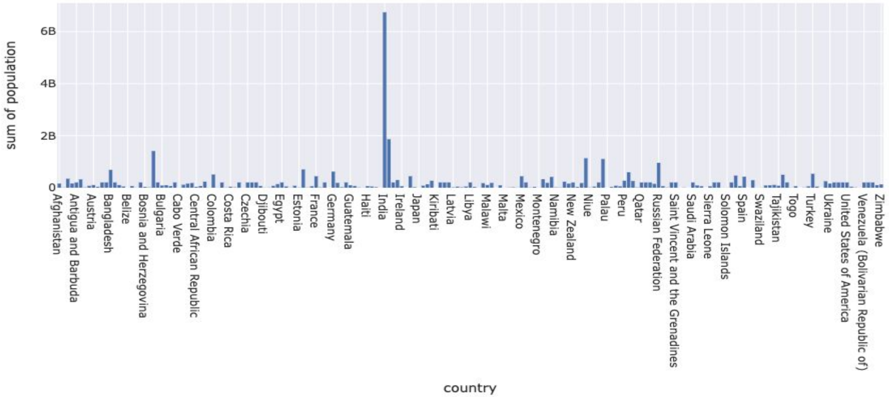


	country	life_expectancy
84	Japan	82.53750
165	Sweden	82.51875
75	Iceland	82.44375
166	Switzerland	82.33125
60	France	82.21875
82	Italy	82.18750
160	Spain	82.06875
7	Australia	81.81250
125	Norway	81.79375
30	Canada	81.68750

Top 10 Worst



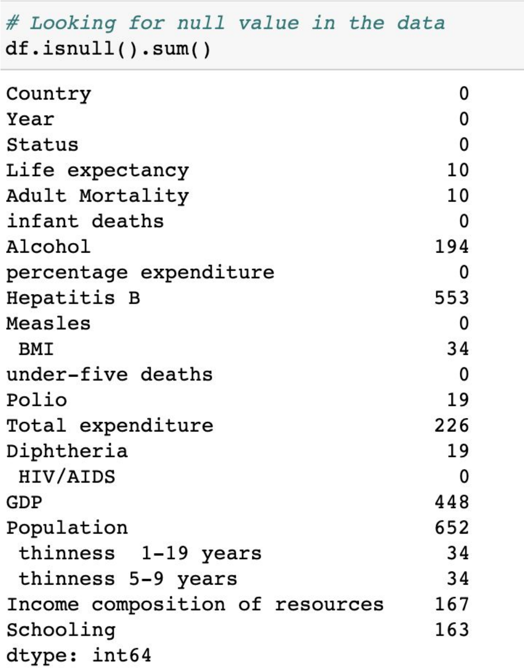
	country	life_expectancy
152	Sierra Leone	46.11250
31	Central African Republic	48.51250
94	Lesotho	48.78125
3	Angola	49.01875
100	Malawi	49.89375
32	Chad	50.38750
44	Côte d'Ivoire	50.38750
192	Zimbabwe	50.48750
164	Swaziland	51.32500
123	Nigeria	51.35625



Data Cleaning

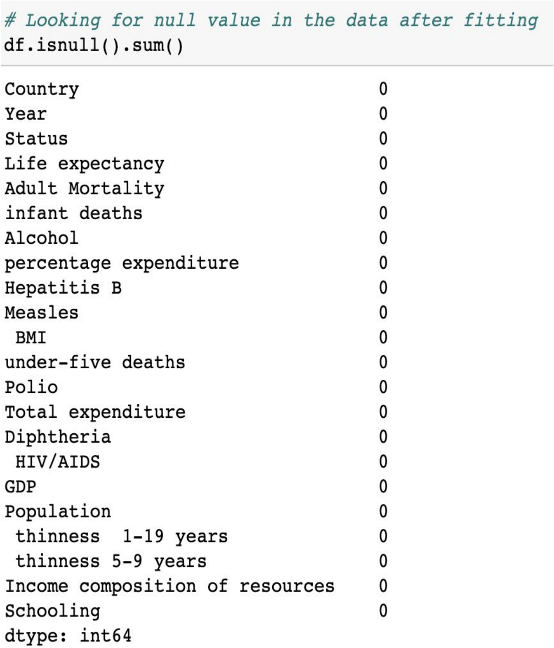
The dealt consisted of some null values, that have been delta with filling them in with the data’s mean values.

- The majority of missing values in the data came from: Population, Hepatitis B, and GDP



```
# Looking for null value in the data
df.isnull().sum()
```

```
Country                              0
Year                                 0
Status                               0
Life expectancy                     10
Adult Mortality                     10
infant deaths                        0
Alcohol                            194
percentage expenditure               0
Hepatitis B                        553
Measles                              0
 BMI                                34
under-five deaths                    0
Polio                               19
Total expenditure                  226
Diphtheria                          19
 HIV/AIDS                            0
GDP                                448
Population                         652
 thinness  1-19 years               34
 thinness 5-9 years                 34
Income composition of resources    167
Schooling                          163
dtype: int64
```

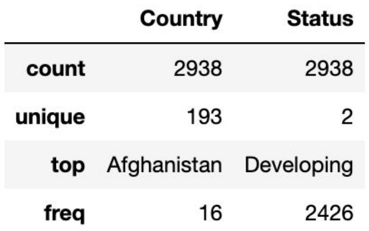


```
# Looking for null value in the data after fitting
df.isnull().sum()
```

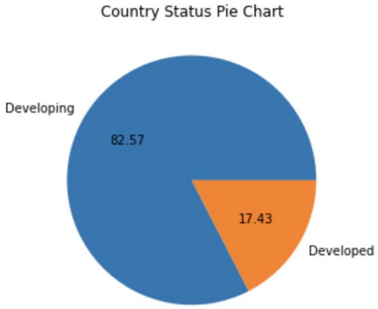
```
Country                            0
Year                               0
Status                             0
Life expectancy                    0
Adult Mortality                    0
infant deaths                      0
Alcohol                            0
percentage expenditure             0
Hepatitis B                        0
Measles                            0
 BMI                               0
under-five deaths                  0
Polio                              0
Total expenditure                  0
Diphtheria                         0
 HIV/AIDS                          0
GDP                                0
Population                         0
 thinness  1-19 years              0
 thinness 5-9 years                0
Income composition of resources    0
Schooling                          0
dtype: int64
```

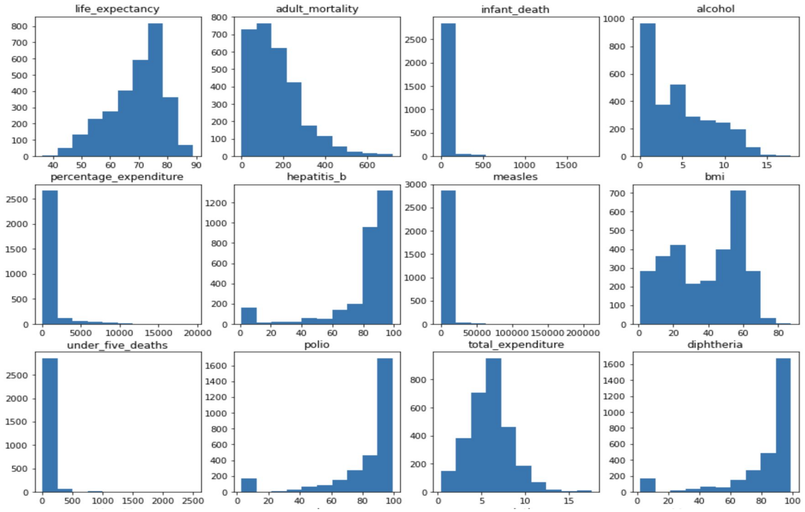
Distribution & Outliers

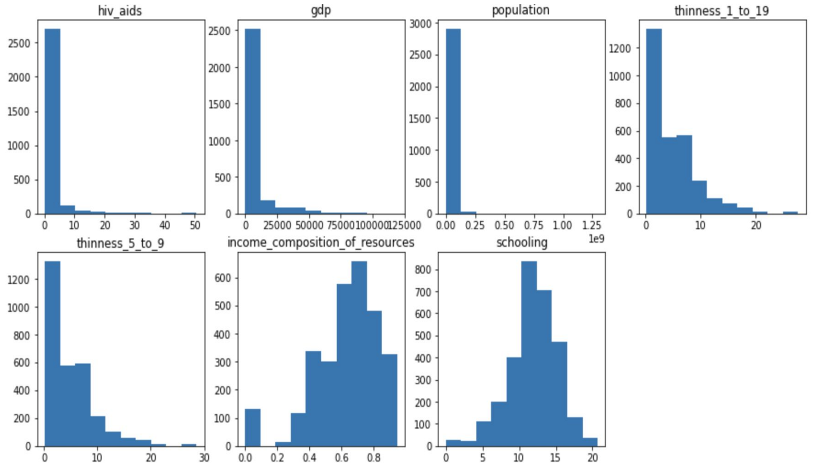
- Then “Measles” &”HIV/AIDS” had the largest % of outliers, with 542, making it 18.45% of the data
- “Afghanistan” has the top frequency



	Country	Status
count	2938	2938
unique	193	2
top	Afghanistan	Developing
freq	16	2426



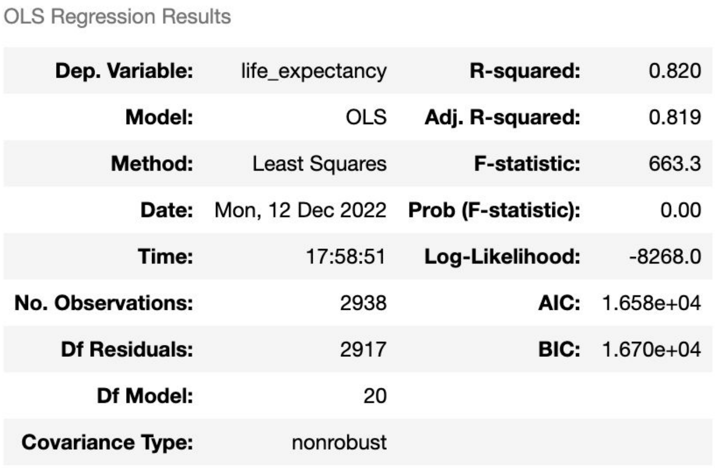




Analyses (or prediction)

Which variable play a major role in Life Expectancy?

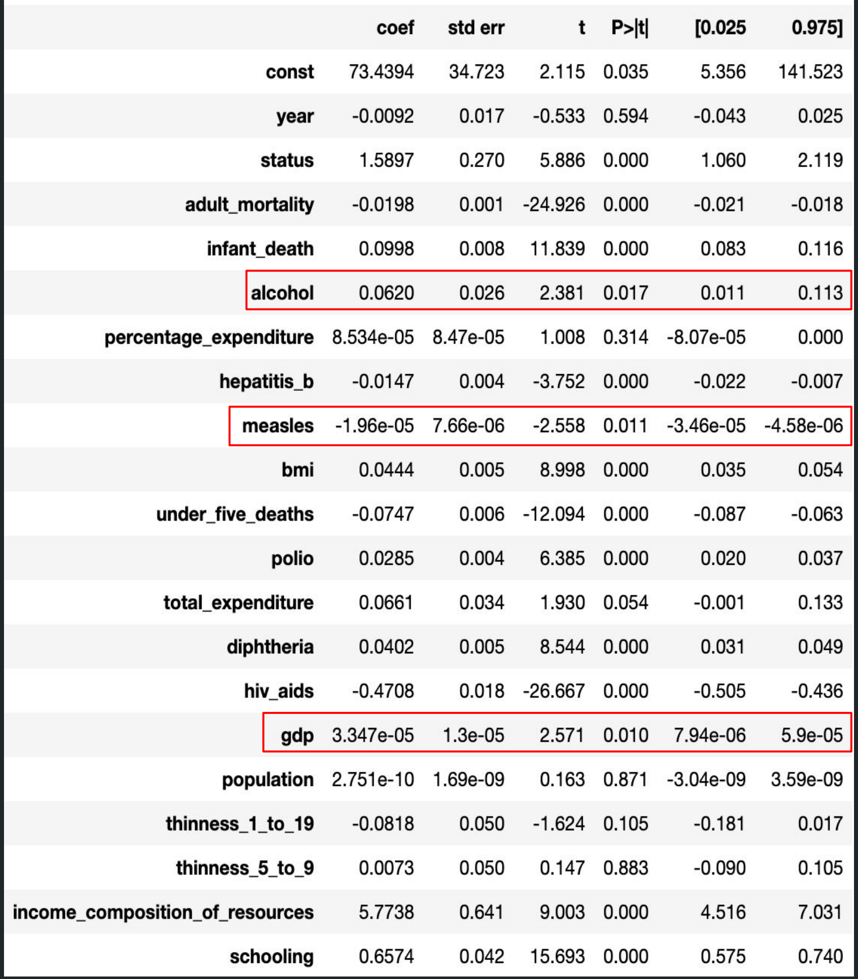
Linear Regression



OLS Regression Results

Dep. Variable:	life_expectancy	**R-squared:**	0.820
Model:	OLS	**Adj. R-squared:**	0.819
Method:	Least Squares	**F-statistic:**	663.3
Date:	Mon, 12 Dec 2022	**Prob (F-statistic):**	0.00
Time:	17:58:51	**Log-Likelihood:**	-8268.0
No. Observations:	2938	**AIC:**	1.658e+04
Df Residuals:	2917	**BIC:**	1.670e+04
Df Model:	20		
Covariance Type:	nonrobust		

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.57e+10. This might indicate that there are
strong multicollinearity or other numerical problems.



	coef	std err	t	P>|t|	[0.025	0.975]
const	73.4394	34.723	2.115	0.035	5.356	141.523
year	-0.0092	0.017	-0.533	0.594	-0.043	0.025
status	1.5897	0.270	5.886	0.000	1.060	2.119
adult_mortality	-0.0198	0.001	-24.926	0.000	-0.021	-0.018
infant_death	0.0998	0.008	11.839	0.000	0.083	0.116
alcohol	0.0620	0.026	2.381	0.017	0.011	0.113
percentage_expenditure	8.534e-05	8.47e-05	1.008	0.314	-8.07e-05	0.000
hepatitis_b	-0.0147	0.004	-3.752	0.000	-0.022	-0.007
measles	-1.96e-05	7.66e-06	-2.558	0.011	-3.46e-05	-4.58e-06
bmi	0.0444	0.005	8.998	0.000	0.035	0.054
under_five_deaths	-0.0747	0.006	-12.094	0.000	-0.087	-0.063
polio	0.0285	0.004	6.385	0.000	0.020	0.037
total_expenditure	0.0661	0.034	1.930	0.054	-0.001	0.133
diphtheria	0.0402	0.005	8.544	0.000	0.031	0.049
hiv_aids	-0.4708	0.018	-26.667	0.000	-0.505	-0.436
gdp	3.347e-05	1.3e-05	2.571	0.010	7.94e-06	5.9e-05
population	2.751e-10	1.69e-09	0.163	0.871	-3.04e-09	3.59e-09
thinness_1_to_19	-0.0818	0.050	-1.624	0.105	-0.181	0.017
thinness_5_to_9	0.0073	0.050	0.147	0.883	-0.090	0.105
income_composition_of_resources	5.7738	0.641	9.003	0.000	4.516	7.031
schooling	0.6574	0.042	15.693	0.000	0.575	0.740

Correlation

Life expectancy + Schooling = 0.72

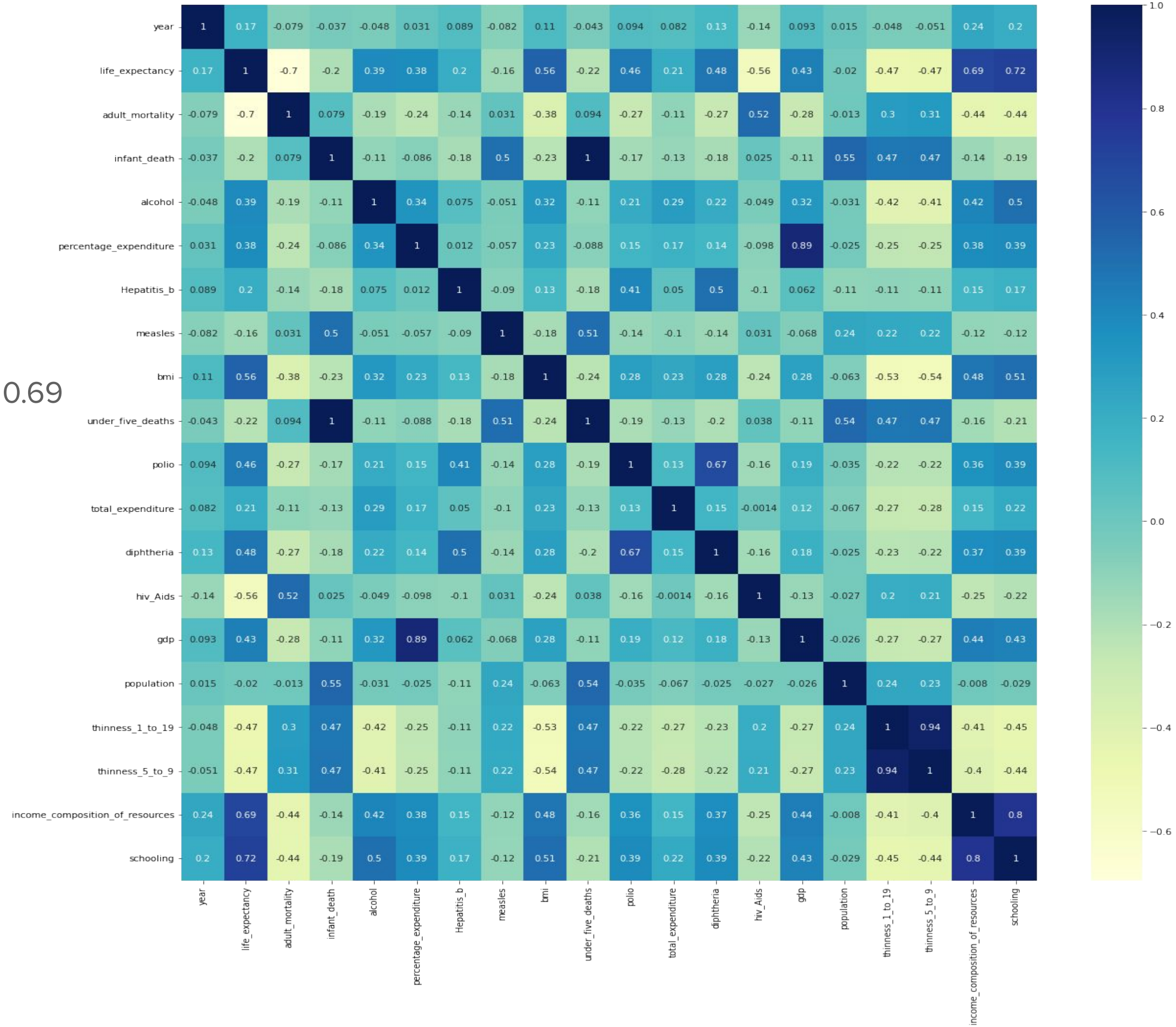
Life expectancy + Income of Resources = 0.69

Life expectancy + BMI = 0.56

Life expectancy + Diphtheria = 0.48

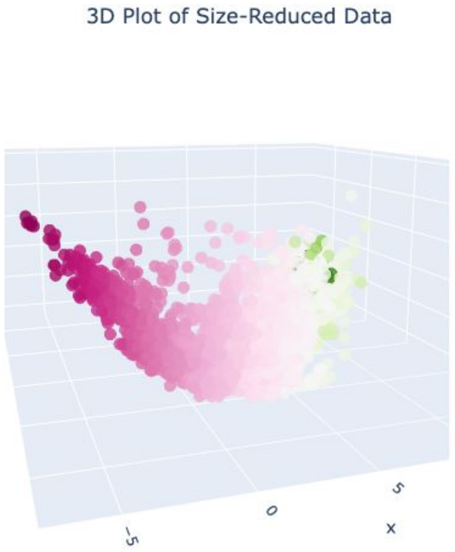
Life expectancy + Polio = 0.46

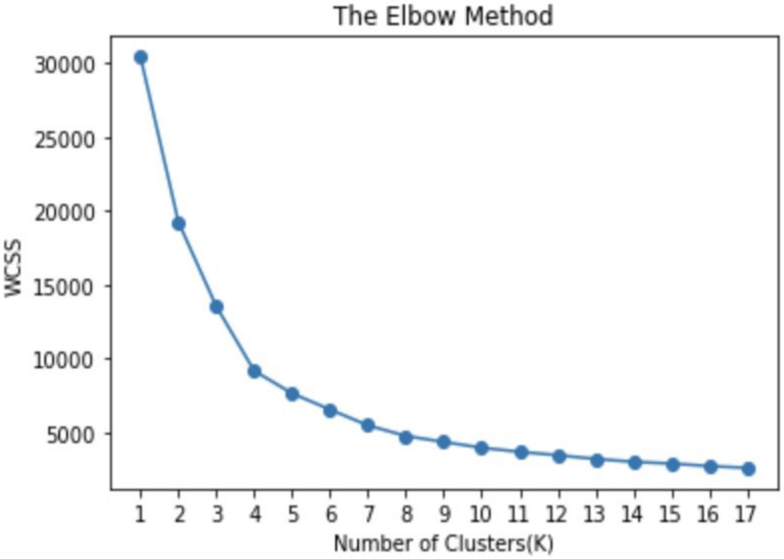
Life expectancy + Population = 0.43



	year	life_expectancy	adult_mortality	infant_death	alcohol	percentage_expenditure	Hepatitis_b	measles	bmi	under_five_deaths	polio	total_expenditure	diphtheria	hiv_Aids	gdp	population	thinness_1_to_19	thinness_5_to_9	income_composition_of_resources	schooling
year	1	0.17	-0.079	-0.037	-0.048	0.031	0.089	-0.082	0.11	-0.043	0.094	0.082	0.13	-0.14	0.093	0.015	-0.048	-0.051	0.24	0.2
life_expectancy	0.17	1	-0.7	-0.2	0.39	0.38	0.2	-0.16	0.56	-0.22	0.46	0.21	0.48	-0.56	0.43	-0.02	-0.47	-0.47	0.69	0.72
adult_mortality	-0.079	-0.7	1	0.079	-0.19	-0.24	-0.14	0.031	-0.38	0.094	-0.27	-0.11	-0.27	0.52	-0.28	-0.013	0.3	0.31	-0.44	-0.44
infant_death	-0.037	-0.2	0.079	1	-0.11	-0.086	-0.18	0.5	-0.23	1	-0.17	-0.13	-0.18	0.025	-0.11	0.55	0.47	0.47	-0.14	-0.19
alcohol	-0.048	0.39	-0.19	-0.11	1	0.34	0.075	-0.051	0.32	-0.11	0.21	0.29	0.22	-0.049	0.32	-0.031	-0.42	-0.41	0.42	0.5
percentage_expenditure	0.031	0.38	-0.24	-0.086	0.34	1	0.012	-0.057	0.23	-0.088	0.15	0.17	0.14	-0.098	0.89	-0.025	-0.25	-0.25	0.38	0.39
Hepatitis_b	0.089	0.2	-0.14	-0.18	0.075	0.012	1	-0.09	0.13	-0.18	0.41	0.05	0.5	-0.1	0.062	-0.11	-0.11	-0.11	0.15	0.17
measles	-0.082	-0.16	0.031	0.5	-0.051	-0.057	-0.09	1	-0.18	0.51	-0.14	-0.1	-0.14	0.031	-0.068	0.24	0.22	0.22	-0.12	-0.12
bmi	0.11	0.56	-0.38	-0.23	0.32	0.23	0.13	-0.18	1	-0.24	0.28	0.23	0.28	-0.24	0.28	-0.063	-0.53	-0.54	0.48	0.51
under_five_deaths	-0.043	-0.22	0.094	1	-0.11	-0.088	-0.18	0.51	-0.24	1	-0.19	-0.13	-0.2	0.038	-0.11	0.54	0.47	0.47	-0.16	-0.21
polio	0.094	0.46	-0.27	-0.17	0.21	0.15	0.41	-0.14	0.28	-0.19	1	0.13	0.67	-0.16	0.19	-0.035	-0.22	-0.22	0.36	0.39
total_expenditure	0.082	0.21	-0.11	-0.13	0.29	0.17	0.05	-0.1	0.23	-0.13	0.13	1	0.15	-0.0014	0.12	-0.067	-0.27	-0.28	0.15	0.22
diphtheria	0.13	0.48	-0.27	-0.18	0.22	0.14	0.5	-0.14	0.28	-0.2	0.67	0.15	1	-0.16	0.18	-0.025	-0.23	-0.22	0.37	0.39
hiv_Aids	-0.14	-0.56	0.52	0.025	-0.049	-0.098	-0.1	0.031	-0.24	0.038	-0.16	-0.0014	-0.16	1	-0.13	-0.027	0.2	0.21	-0.25	-0.22
gdp	0.093	0.43	-0.28	-0.11	0.32	0.89	0.062	-0.068	0.28	-0.11	0.19	0.12	0.18	-0.13	1	-0.026	-0.27	-0.27	0.44	0.43
population	0.015	-0.02	-0.013	0.55	-0.031	-0.025	-0.11	0.24	-0.063	0.54	-0.035	-0.067	-0.025	-0.027	-0.026	1	0.24	0.23	-0.008	-0.029
thinness_1_to_19	-0.048	-0.47	0.3	0.47	-0.42	-0.25	-0.11	0.22	-0.53	0.47	-0.22	-0.27	-0.23	0.2	-0.27	0.24	1	0.94	-0.41	-0.45
thinness_5_to_9	-0.051	-0.47	0.31	0.47	-0.41	-0.25	-0.11	0.22	-0.54	0.47	-0.22	-0.28	-0.22	0.21	-0.27	0.23	0.94	1	-0.4	-0.44
income_composition_of_resources	0.24	0.69	-0.44	-0.14	0.42	0.38	0.15	-0.12	0.48	-0.16	0.36	0.15	0.37	-0.25	0.44	-0.008	-0.41	-0.4	1	0.8
schooling	0.2	0.72	-0.44	-0.19	0.5	0.39	0.17	-0.12	0.51	-0.21	0.39	0.22	0.39	-0.22	0.43	-0.029	-0.45	-0.44	0.8	1

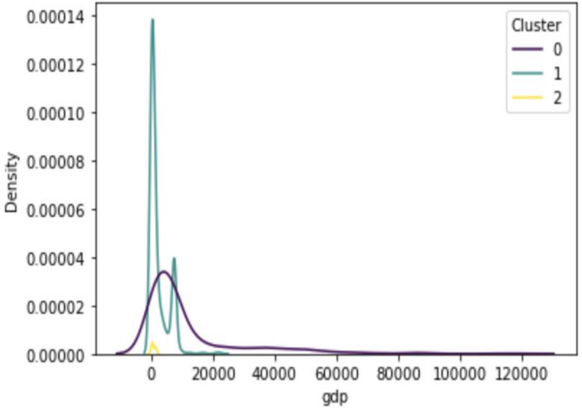
PCA & Elbow Method

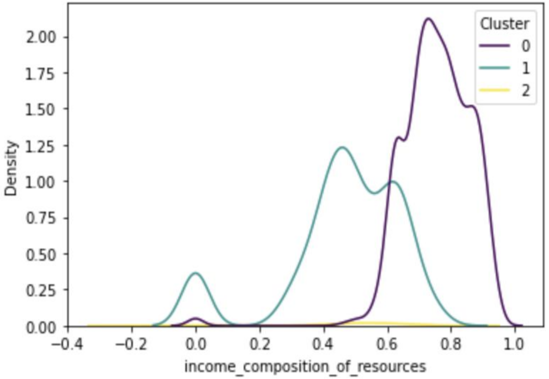


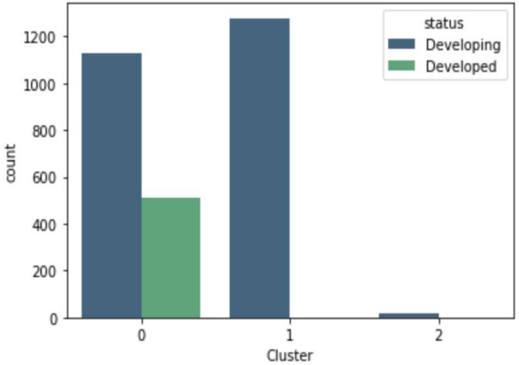


K-Means Clustering

Cluster 0 = High life expectancy
Cluster 1 = Moderate life expectancy
Cluster 2 = Low life expectancy



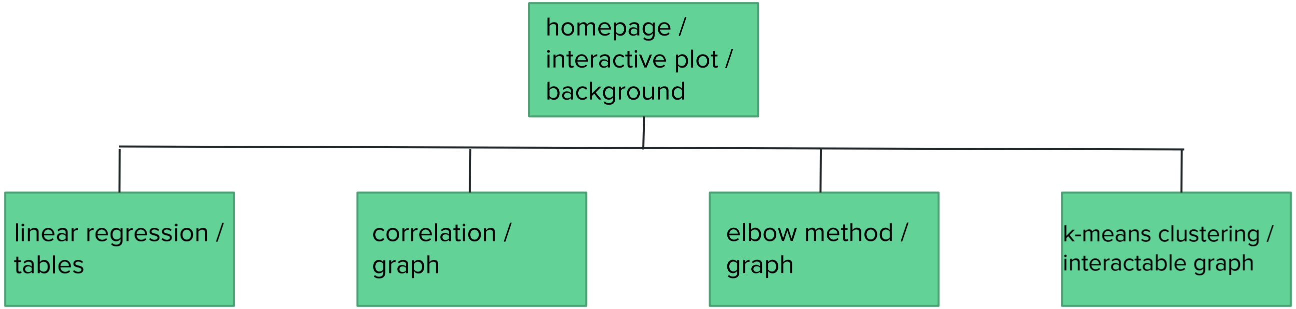




Server API and Web Front-end

Flask

web development,
one drop at a time



homepage / interactive plot / background

linear regression / tables

correlation / graph

elbow method / graph

k-means clustering / interactable graph

DEMO…

Thank you!