Analysis of Human Longevity, Dietary Factors, and National Wealth

By Austin Geary, Scott Cole, and Mohammed Kibria

Motivation

Each member of our team has specific interests in the domain of longevity. But collectively we, like most fortunate humans, love life and want more of it. We proposed to analyze data from countries around the world to determine lifestyle factors that have a high impact on their citizens' lifespans. The questions we sought to answer were:

- Which lifestyle factors most strongly correlate with long (and short) life?
- Which countries have a mix of factors conducive to the longest (and shortest) lifespans?
    - And are there natural clusters of factors and lifespans?
- Can we predict the lifespan of a country based on the values of certain factors?

Data Sources

Data Set: COVID-19 Diet Dataset

- **Description:** Contains various types of food consumed, obesity rates, and undernourished rates for 170 different countries. Specifically, it contains the types of foods and the percentage of calories a country's population generally obtains from them.
- **Location:** https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset
- **Format:** 173kB, CSV
- **Key Variables:** Obesity rate, and percentage of energy intake (kcal) from: cereal (excluding beer), meat and vegetables.
- **Number of Records:** 170
- **Time Period:** While not stated, the data set was compiled in 2021 to study correlations between diets in different countries and how COVID-19 impacted those country's populations. We assume the data represents diets in the year 2020 or 2019.

Data Set: WHO Life Expectancy Data Set

- **Description:** Contains diverse features from the WHO repository that can be used to model life expectancy across countries. Covers 193 countries with features like economic status, adult and infant mortality, alcohol consumption, disease prevalence.
- **Location:** https://www.kaggle.com/kumarajarshi/life-expectancy-who
- **Format:** 326kB, CSV
- **Key Variables:** Life expectancy, year, average liters of alcohol people drink by country.
- **Number of Records:** 2938
- **Time Period:** 2000 - 2015

Data Set: Per Capita GDP

- **Description:** Lists the 2017 per capita GDP for 190 countries both in terms of purchasing power parity and nominal value.
- **Location:** https://www.worldometers.info/gdp/gdp-per-capita/
- **Format:** 5kB, CSV (copy/paste from online table)
- **Key Variables:** Per capita GDP listed in terms of purchasing power parity.

- **Number of Records:** 190
- **Time Period:** 2017

Data Set: Natural Earth, Low Resolution
- **Description:** Geopandas module dataset with geometries for countries of the world.
- **Location:** Within the geopandas module. Code to retrieve is: import geopandas, geopandas.read_file(geopandas.datasets.get_path('naturalearth_lowres'))
- **Format:** 180kB, SHP
- **Key Variables:** iso_a3 (code indicating country identity) and map geometry by country.
- **Number of Records:** 177
- **Time Period:** 2021

Data Manipulation Methods

Aggregation - During the project proposal phase, our plan was to find datasets that span many countries over a number of years, so that we could analyze not only the differences in life expectancy and other factors on a per country basis, but also see how these variables changed over time. This approach proved impractical as the number of years involved was too small. Additionally, it was difficult to gauge in what sense the life expectancy at birth in a given year aligns with the consumption of dairy in that year, for instance. Instead, for each of our datasets we took all data from 2010 onwards, and then grouped by country, aggregating the mean of all numerical variables. We then proceeded with the assumption that these means were representative of these countries in recent times.

Missing values - Our general approach was to drop rows that contained any missing values after the aforementioned aggregation. Aggregate functions will generally ignore missing values and calculate from the data that is not missing - this allows us to include more countries in our analysis. Afterward, there was a subset of countries that had one or two missing values, but those values were in columns that were later determined to be unimportant. We dropped those unimportant columns before dropping the rows, in exchange for more countries.

Joining - In joining the data, we had the clear advantage that all of our data was ultimately organized by country, and so could be joined with certainty. In practice, all of our datasets had different references to various countries, e.g. “Iran” v. “Islamic Republic of Iran.” We standardized references to countries in two ways. First, for datasets that did not have an ISO 2 (International Standard Country Code - 2 letters) or ISO 3 code associated with each country, we had to edit the country names manually to a standard spelling. The standard spelling was determined by a dictionary we found on GitHub mapping country names to ISO 2 codes. This standardization proved very helpful in joining said datasets to datasets that did have ISO codes for countries, given that we could use our dictionary to map country names to ISO 2 codes, and then if necessary use another dictionary we found on GitHub to map ISO 2 codes to ISO 3 codes. Through these methods, we were able to merge all target datasets into our final analytical dataset.

Feature Selection - After joining our datasets by the ISO codes mentioned in the *Data Manipulation Methods* section, we had to select which features were most usable for our

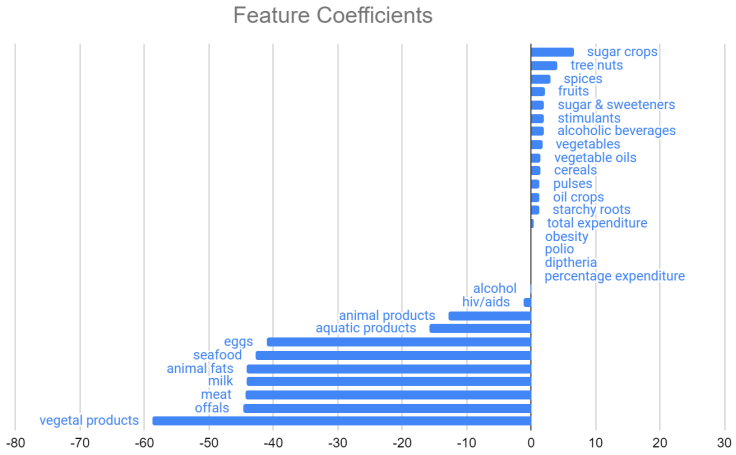
analyses. Our goal was to keep as many countries as possible to have a richer data set to analyze. So we leaned towards maximizing the number of countries and minimizing the number of missing values in a given feature. To do this, we used the *.isna()* method to count up the number of missing values in each column. We noticed two countries that had no life expectancy data, and since this was a key feature for our analyses, we dropped these two countries. Other columns had multiple missing values: ['schooling', 'undernourished', 'income composition of resources', 'thinness 5-9 years', 'thinness 1-19 years', 'population_x', 'hepatitis b', 'bmi]. To preserve the maximum number of countries, we removed these features as they did not provide more insight into our analyses. Features like 'BMI' and 'Income composition of resources' were already captured in other features like 'Obesity' and 'GDP per Capita' so we felt comfortable with leaving these out.

Analysis and Visualization

Linear Regression

One of the questions that we sought to answer in the course of our analysis was whether we were able to predict the lifespan of a country based on values of certain variables. To this end we developed a regression model with the aim of optimizing for a high coefficient of determination score on a withheld "test set" of data. Our methodology involved two strategies that, combined, would select for the minimum number of important features to have a strong, general model.

The first strategy, called Recursive Feature Elimination (RFE), essentially involves searching for a specified number of features from the full set that are most important to the model. The least important feature is removed first, followed by a refitting of the model, and then repeating that process until the specified number is reached.The second strategy involved simply looping through the full number of features and passing each in as a specified number for RFE, and keeping that model that scored highest on the test data set. The best regression model is found below:



Number of features: 29

linear model intercept (b): 2917.669

R-squared score (training): 0.833

R-squared score (test): 0.809

This model selects 29 of the most important features out of the original 37. It should be noted that we removed a few features ahead of time that we thought were a form of data leakage, such as 'adult mortality' and 'infant deaths'.

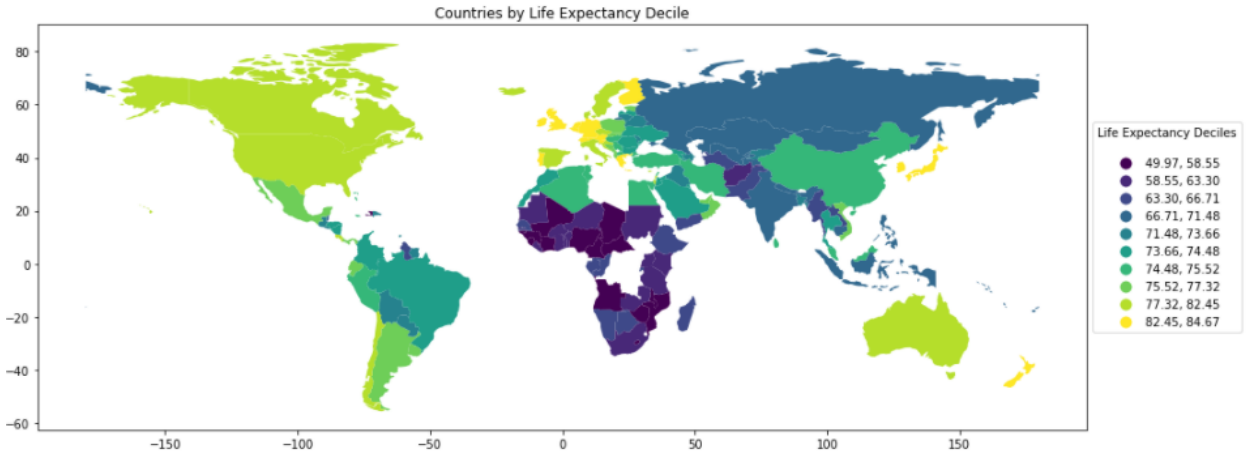
After some consideration and further correlation analysis, we believe that most of the coefficients above do not reflect their genuine effect on lifespan and therefore this model will require either more data, more careful tuning, and/or controlling for confounders to become useful. For instance, vegetable oil is shown to have a positive coefficient, but as we later discovered this was an example of Simpson's paradox, with per capita gdp as a confounder. In fact, as we will show many of these features are confounded by wealth, which makes the model fitting somewhat erroneous. Our dataset consisted of 160 countries, which is likely not a large enough sample to reliably make use of 29 features, especially since most of these features are likely correlated and even collinear. The model is likely learning how these variables track with wealth, and using that to predict against the test set.

Even though this model scored high on a withheld test set, it might not have given us the actual information we are interested in, which is ultimately a question of causality. If we rephrased our question about prediction, it would take the form: "Can we predict the lifespan of a country whose values for each of the features was a random dice roll?" This framing of the question would clarify the need to remove confounders. Also, if we collected data about 10,000 cities rather than only 160 countries, the model might pick up on the nuances of each feature's effect on lifespan.

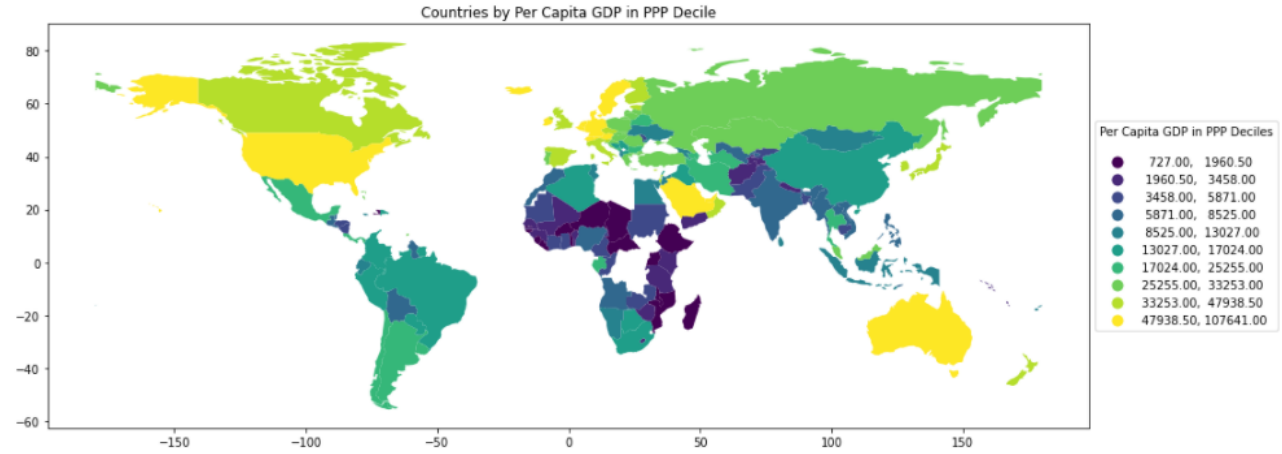
Wealth as a Confounding Variable

At the start of our investigation, we hypothesized that wealth would be a confounding variable between life expectancy and many of the variables we were studying. We based this hypothesis on the many studies we have heard publicized in the news highlighting the disparity in life expectancy between lower-income and higher-income populations in the United States. For wealth to be a strong confounding variable however, we had to check three aspects of our data: 1. Is there large variation in life expectancy across countries? 2. Is there large variation in wealth across countries? 3. Are the trends in these two variables correlated across countries?

Below you see a map of all countries represented in our study colored by the decile of life expectancy into which their population falls. Darker colors indicate shorter life expectancy; lighter colors indicate longer life expectancy. The map makes it easy to see the very large variation that exists between countries, with shorter life expectancies concentrated in Africa and longer life expectancies concentrated in Europe, North America and Oceania. The variation in life expectancies is enormous, with the shortest life expectancy being about 50 years and the longest being about 85 years. People in the country with the longest life expectancy (Germany) are expected live 71% longer than those in the country with the shortest life expectancy (Sierra Leone)!



Perhaps less surprising is the fact that there is large variation in wealth between countries. That said, the scale of the disparity in wealth is likely surprising to most readers. When measuring wealth through purchasing power parity of per capita GDP, the richest country (Luxembourg) is 148 times wealthier than the poorest country (Central African Republic).



Comparing the two maps, it's easy to see that patterns in life expectancy across countries do strongly correlate with patterns in wealth distribution, as the maps are colored very similarly. The Pearson correlation coefficient for the correlation between per capita GDP in PPP and life expectancy is 0.72, confirming that the correlation between the two variables is strong.

Correlations by Wealth Bins

In order to control for wealth as a confounder, one branch of our analysis involved dividing the dataset into several wealth bins. This strategy of slicing our data into narrow bands is a lesser version of holding wealth constant. If wealth could truly be held constant across a segment of data, that would entirely remove its ability to confound other factors. By grouping our data into narrow bands, we get the effect of *lessening* wealth's ability to confound other factors. We segmented our data into 5 bins as follows:

- All countries - 160 data points
- Poor countries (per capita gdp < $5,000) - 42 data points

- Middling countries ($5,000 < per capita gdp < $14,000) - 41 data points
- Wealthy countries ($14,000 < per capita gdp < $30,000) - 40 data points
- Rich countries (per capita gdp > $30,000) - 37 data points

Then for each bin, we calculated Spearman's correlation coefficient for lifespan with each of the factors:

	alcohol	percentage expenditure	measles	polio	total expenditure	diphtheria	hiv/aids	alcoholic beverages	animal products	animal fats	cereals - excluding beer
all	0.566	0.602	-0.276	0.538	0.407	0.540	-0.768	0.449	0.763	0.672	-0.556
poor	-0.245	0.165	-0.235	0.431	-0.052	0.362	-0.786	-0.232	0.196	0.149	-0.007
middling	0.122	0.092	-0.292	0.443	0.438	0.422	-0.602	-0.048	0.286	0.178	-0.059
wealthy	0.086	0.368	-0.214	0.021	0.595	0.121	-0.447	-0.107	0.284	0.197	-0.195
rich	0.414	0.385	0.295	-0.010	0.316	-0.001	-0.348	0.317	0.313	0.341	-0.176

	eggs	fish, seafood	fruits - excluding wine	meat	milk - excluding butter	offals	oilcrops	pulses	spices	starchy roots
all	0.734	0.286	0.220	0.618	0.616	0.069	-0.267	-0.405	0.152	-0.370
poor	0.384	0.026	0.047	0.040	0.121	-0.211	-0.164	-0.191	0.147	-0.307
middling	0.392	-0.039	0.301	0.043	0.362	0.176	-0.169	-0.173	0.027	-0.232
wealthy	0.101	0.259	0.081	0.212	0.201	-0.131	0.032	0.233	0.223	-0.330
rich	-0.002	0.110	-0.071	0.258	0.162	-0.041	-0.080	-0.467	-0.374	0.391

	stimulants	sugar crops	sugar & sweeteners	treenuts	vegetal products	vegetable oils	vegetables	obesity	population	per capita gdp
all	0.608	-0.082	0.510	0.558	-0.763	0.255	0.474	0.601	-0.074	0.853
poor	0.276	0.341	0.337	-0.006	-0.200	-0.165	0.119	0.138	-0.041	0.471
middling	0.181	0.002	0.228	0.321	-0.286	-0.176	0.306	0.421	-0.168	0.449
wealthy	0.165	-0.096	0.313	0.300	-0.280	-0.120	0.124	0.300	0.006	0.235
rich	0.077	0.141	-0.059	0.237	-0.313	0.021	0.142	-0.203	0.253	0.206

Strategy Confirmed

Per capita GDP correlates to the lifespan of a country's citizens with a coefficient of .85 which is very strong. However, when countries are divided into narrow bands of wealth, the correlation drops to a moderate .2-.47. This indicates that wealth's ability to drive other correlations will be somewhat diminished.

Simpson's Paradox

Simpson's paradox occurred in a few variables. For example, consumption of vegetable oils has a moderately positive correlation with lifespan when all countries are considered. When broken out by wealth however, it actually has a negative valence for each segment, except for rich countries where there is very little correlation at all. This suggests that wealthy countries tend to consume more vegetable oils and live longer, but that the consumption of vegetable oils itself does not cause longer life expectancies.

Not all examples involve a switch in the valence of the correlation, but sometimes just a weakening of an apparently strong relationship. The consumption of animal products and animal fats has a very strong correlation to lifespan when all countries are considered. When looking at bins of countries by wealth, this relationship is moderate at best. This situation suggests that wealthier countries consume more meat and live longer, but the eating of the meat itself might only have a small-to-moderate benefit for lifespan.

Diverging Correlates

Lastly, there are situations that may indicate that what is good for a citizen of a poorer country, is negligible or even bad for a citizen of a wealthier country, and visa versa. Consumption of alcohol and alcoholic beverages seems to be good for citizens of “rich” countries, whereas it’s negligible for “wealthy” and “middling”, and downright bad for citizens of “poor” countries. With the prior knowledge that alcohol has been studied in-depth and known to be a comorbidity for most major diseases, it stands to reason that another factor is confounding the relationship between life expectancy and the consumption of alcohol in “rich” countries.

Another notable example is the incidence of HIV/AIDS which follows a gradient pattern from “poor” to “rich” with the former having a stronger negative correlation and the latter a weaker negative correlation to lifespan. This may suggest that the wealthier a country is, the better treatments it has for HIV/AIDS, and therefore the longer those patients can still survive.

Using PCA and K-means Cluster Analyses to Identify Exceptions to Correlations

So far in our investigation, we have seen just how strongly life expectancy is correlated with wealth from a number of perspectives. In order to see if there are exceptions to this trend, we decided to conduct a cluster analysis hypothesizing that perhaps a cluster or two of countries in this analysis would show that the correlation between life expectancy and wealth is not absolute. To not only conduct this analysis but also display it in two dimensions, we first reduced the number of dimensions of our data set using principal component analysis. Starting with the 14 variables listed below - which all have an absolute Pearson correlation with life expectancy of 0.6 or higher - we used PCA to reduce the number of features for our k-mean cluster analysis to just 2. Note that we did not include life expectancy in the cluster analysis because we wanted to see if the countries would cluster in ways that showed new trends in life expectancy. The result of the cluster analysis below shows 10 distinct clusters of countries when grouped by their values along the two principal components produced by our PCA.

Starting variables: ['alcohol','polio','diphtheria','hiv/aids','animal products','animal fats','eggs', 'meat', 'milk - excluding butter','starchy roots','stimulants','sugar & sweeteners','obesity', 'per_capita_gdp_ppp']

When we look at the median values across the 14 variables for each cluster, we notice that the cluster circled in red (cluster 3) does break with most trends in our data set. Specifically, cluster 3 breaks with 9 of the 14 trends. Below are all clusters shown in the graph along with their median values in those 9 variables in order of their median life expectancy. Generally, there is a clear increasing or decreasing trend in these variables. Cluster 3 represents a break from those trends however - e.g. countries in cluster 3 are considerably less wealthy than would be expected given their life expectancy. They have a considerably lower obesity rate, consume fewer animal products and alcohol, etc. This shows that the relationships between these variables and life expectancy are not absolute, and it would be worthwhile to study the countries in cluster 3 more in depth to understand what factors are strongly correlated to their life expectancies if not those studied in our investigation.

cluster	alcohol	animal products	animal fats	eggs	meat	milk - excluding butter	stimulants	obesity	per_capita_gdp_ppp	life expectancy
5	1.164	5.96080	0.16245	0.05680	3.09950	0.78785	0.06435	7.05	3699.5	54.250000
0	1.274	3.52080	0.18390	0.06990	1.81290	0.66190	0.04690	6.80	2247.0	59.383333
7	0.672	3.76700	0.31300	0.08300	1.15830	0.88060	0.06180	7.10	2434.0	62.133333
4	3.090	8.52880	0.93630	0.29920	3.02890	2.98590	0.17890	15.00	8361.0	71.633333
9	7.262	10.70420	1.37830	0.62640	4.58110	2.49090	0.28570	26.10	13109.0	73.650000
3	0.693	5.67590	0.36595	0.30955	2.10930	1.36445	0.09570	17.05	8196.5	73.866667
1	2.935	9.25515	0.84355	0.42220	3.60545	3.12410	0.30915	23.10	15173.5	74.266667
6	7.602	13.14190	1.17730	0.63890	5.72570	4.33500	0.38310	23.50	23522.0	75.200000
2	10.086	15.41110	2.99970	0.63360	5.71430	4.90720	0.50370	25.60	40797.0	81.916667
8	10.013	17.57840	4.02310	0.93120	6.30305	5.29820	0.91210	22.75	54839.0	81.991667

The countries in cluster 3 include the following: Algeria, Bangladesh, Cambodia, Egypt, Gambia, Guatemala, Honduras, Iran, Jordan, Kiribati, Malaysia, Mauritius, Morocco, Nicaragua, Sao Tome and Principe, Senegal, Sri Lanka, Tajikistan, Thailand, Tunisia, Turkey, Yemen. It's worth noting that 13 of these 22 countries are predominantly Muslim, and perhaps some aspect of Islamic societies is contributing to these discrepancies.

Factors Correlated with Longevity for Long-lived Countries

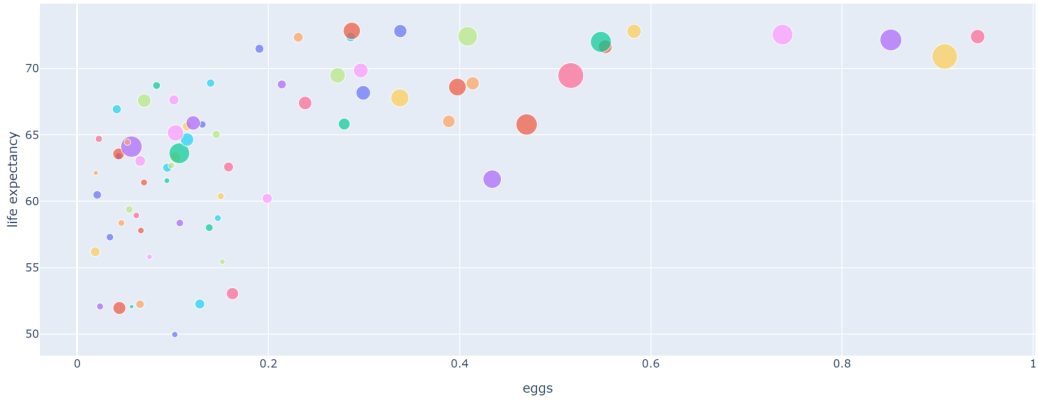
We next wanted to focus explicitly on the strongest correlates with life expectancy for 4 groups of countries to see how those correlates varied between the groups: top and bottom 25% of countries with respect to life expectancy, and the top and bottom 25% of countries with respect to wealth.

The top 25% of countries with respect to life expectancy were essentially countries with life expectancies above 75 years, and so we first subset our data to just these countries. A Pearson correlation analysis between life expectancy and the other features in our dataset reveals that percentage expenditure on healthcare, per capita GDP, and average liters of alcohol consumed are the strongest correlates for this group of countries.

Factors Correlated with Longevity for Short-lived Countries

We then subset countries at the bottom 25% of life expectancy, which is under 73 years, and found different features most strongly correlated with life expectancy than we did for the

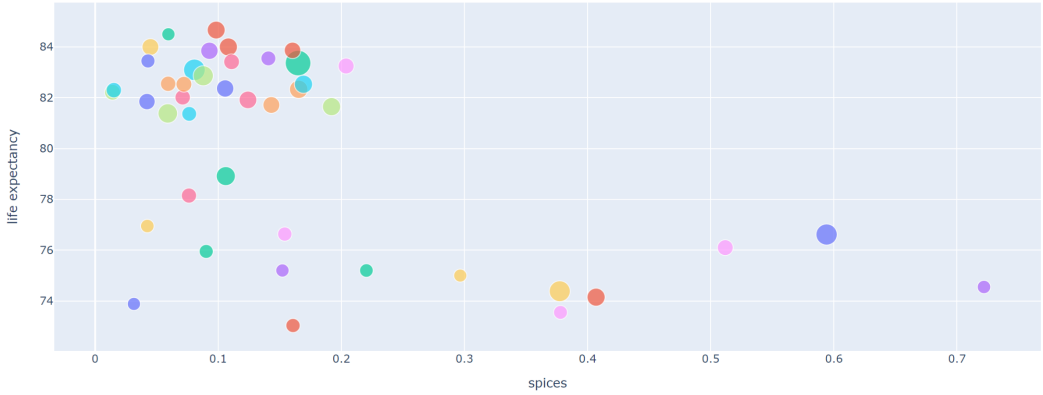
longest-lived countries: prevalence of hiv/aids, adult mortality, and the consumption of eggs. It is particularly worth noting that egg consumption has a positive correlation with life expectancy for countries with life expectancies below 73 years. To delve deeper we can plot egg consumption against life expectancy, but also encode per capita GDP as dot size.



What we can see from this graph is that as egg consumption increases, life expectancy does increase as well, but so does the per capita GDP. A likely reason for this is that per capita GDP is a confounding variable, influencing both egg consumption and life expectancy.

Factors Correlated with Longevity for Wealthy Countries

The top 25% of countries as measured by per capita GDP included countries with per capita GDP above $28,000 per year. The strongest correlates for this group are adult mortality and spice consumption, with countries that consume fewer spices enjoying a longer life expectancy.

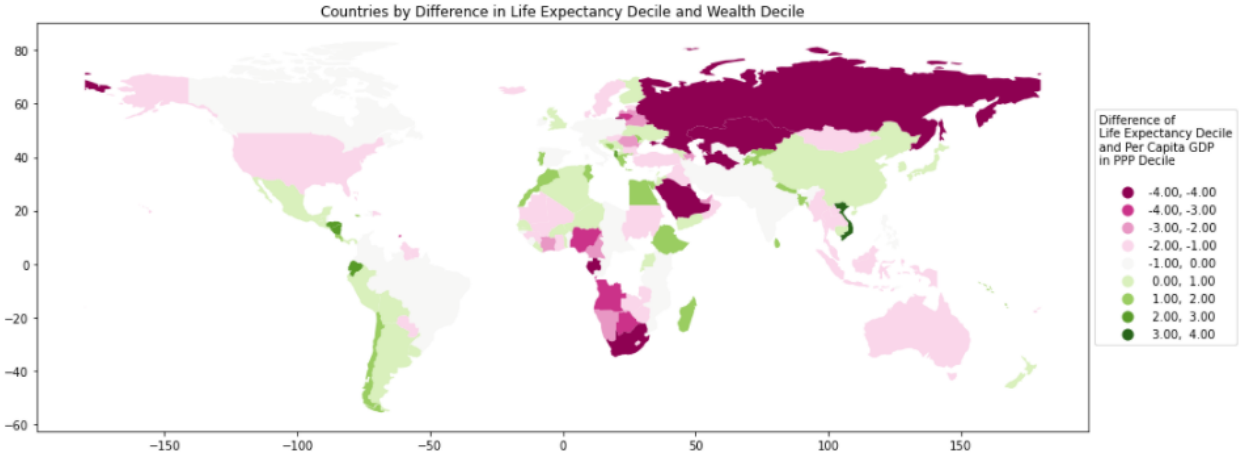


Factors Correlated with Longevity for Poor Countries

The strongest correlates with life expectancy for countries at the bottom 25% of per capita GDP (less than $5,000 per year) include: prevalence of hiv/aids, adult mortality, and the consumption of eggs. These are the same correlates that we observed for countries that have a low life expectancy. This shows further potential evidence of wealth being a confounder for the relationship between life expectancy and the other features in our analysis.

Highlighting Wealthy Countries with Shorter Lifespans, Poor Countries with Longer Lifespans

We next focused on countries that had shorter or longer life expectancies than their level of wealth would have us expect. To identify these countries, we calculated what decile each country fell into for both life expectancy and wealth, as represented by per capita GDP measured in purchasing power parity. We then subtracted the two measures for each country from each other. Countries with highly positive scores are in much higher deciles for life expectancy than wealth, and so have a higher life expectancy than wealth alone would predict. Countries with highly negative scores have a lower life expectancy than wealth alone would predict.



Countries with strongly lower life expectancies than wealth would predict include Gabon, Kazakhstan, Kuwait, Russia, Saudi Arabia, South Africa, and Turkmenistan. We do not have time in this investigation to explore the factors that might be correlated with these lower than expected life expectancies, but we submit that it would be worthwhile to do so in a future study given that these countries are 4 deciles lower in life expectancy than they are in wealth. A study of possible reasons why this is so could yield important insights into what factors in a country are particularly detrimental to longevity. Similarly, studying the countries that are overperforming with regards to their wealth level could yield important insights into what factors in a country are particularly beneficial for longevity. The only country that is 4 deciles higher in life expectancy than it is in wealth is Vietnam.

Conclusion

In this analysis, we were able to answer all of the questions we posed at the start of our study in full or in part. We were able to fuse several datasets into a final set of features which proved strongly predictive of a country's life expectancy, once combined in a linear regression model. We were able to study countries with long and short life expectancies, and found that the strongest correlates with life expectancy between the two groups did indeed differ entirely, e.g. alcohol is strongly positively correlated with life expectancy in countries with long life expectancies, and in countries with short life expectancies the strongest positive correlate is egg consumption. Permeating our analysis however, is the realization that wealth confounds the relationship between life expectancy and many of the features in our analysis.

We were able to demonstrate these confounding relationships through several steps. The first was producing evidence of the exceptional strength of the correlation between life

expectancy and wealth, as measured by per capita GDP. The second was producing evidence that when conditioning on wealth, many of the correlations between life expectancy and other variables change - by either disappearing, switching valence, or differing significantly in strength across wealth bands. Whether wealth is acting as a confounder, mitigator or collider with these variables is beyond the scope of our study, but our analyses show that it is one of the three in many cases.

The strength of the relationship between life expectancy and wealth brings us to our most interesting findings: there are groups of countries whose life expectancies cannot be reliably predicted by wealth. This includes some countries with middling life expectancies, which we discovered in our PCA and k-means cluster analysis. It also includes some wealthier, shorter-lived countries and poorer, longer-lived countries. Further study of lifestyles and habits in these groups of countries could reveal fascinating ways that people in those countries are prolonging their lives without the benefits afforded to citizens of wealthier countries. It could also reveal particularly harmful habits present in some wealthier countries that people would do well to avoid. In either case, our investigation points to these countries as the best place to extend this study - in so doing, perhaps our lives.

Statement of Work

Scott Cole: I initially found several data sets with social features for inclusion in our analysis, but none of them covered enough countries to keep in our study. I then tried joining some of the data sets other team members had found and produced the data set that we used for our project analysis. My major analytical contributions were the world map choropleths and associated analyses, the PCA cluster analysis, and the outlier analysis showing which countries had significantly higher or lower life expectancies than wealth alone would predict.

Austin Geary: My contribution involved setting up the initial pipelines and aggregation strategy, creating dictionaries in order to map country names to their iso2 and iso3 codes, development of the linear regression model, and the Spearman correlation by wealth bin. In the report, I wrote the Motivation section, most of the Data Manipulation section, and part of the Analysis and Visualization (Linear Regression and Correlations by Wealth Bins).

Mohammed Kibria: I worked on delving into the food dataset as it correlates to different slices of the data; like wealth, poverty, long life expectancy, and short life expectancy. I built some plotly interactive visualizations along with correlation heatmaps to take a look at the data deeper. I used the dataset that the team has agreed upon and also cleaned the data to get the final dataset in the correlation studies. Studying subsets in the dataset was a major part of my contribution along with trying to uncover what this could mean in the broader context of our narrative.