Joshua Cano

Food Consumption Habits and Their Effects on a Nations Health Indicators

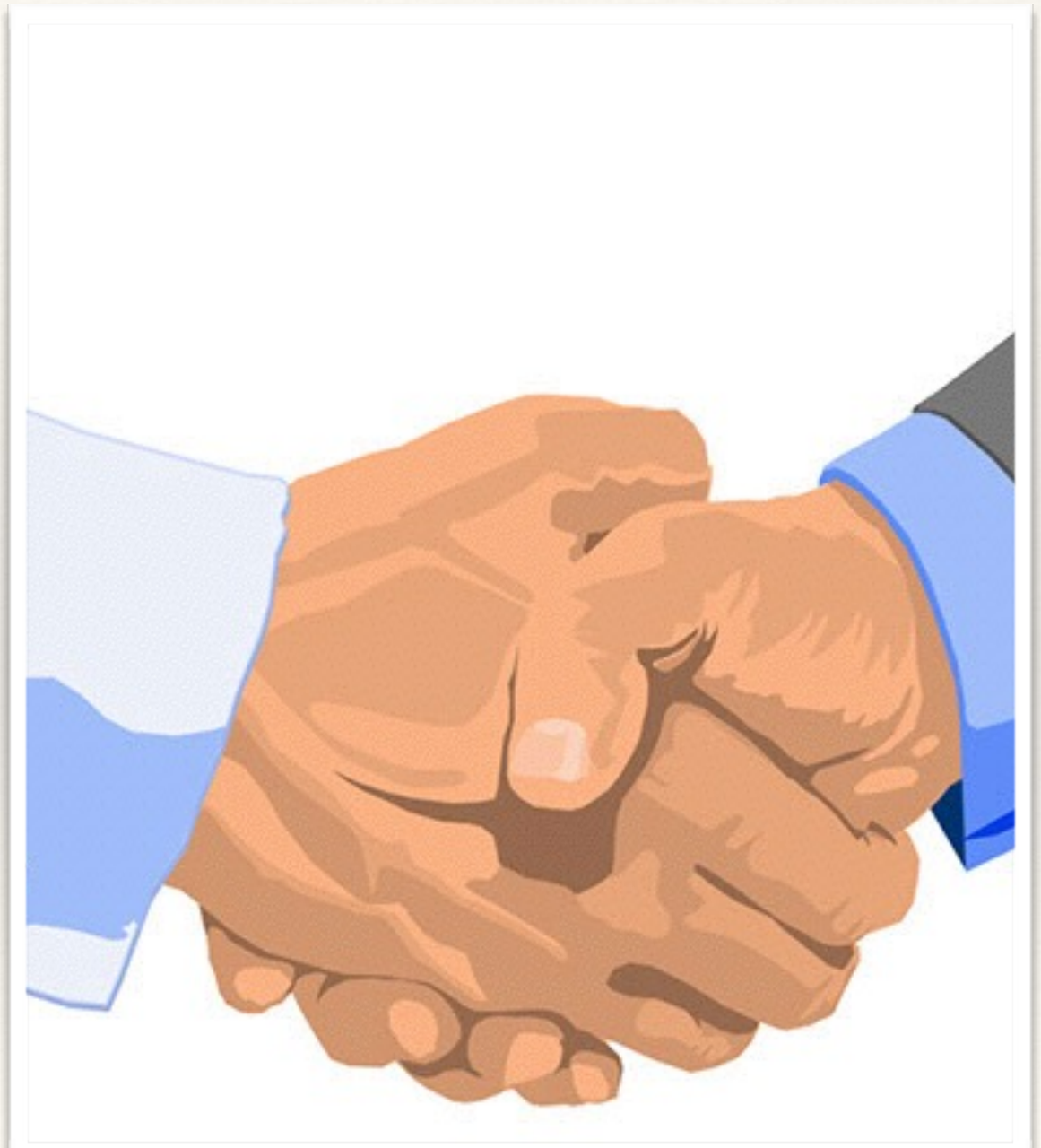
Hypothesis

- A Nation's sugar consumption has a direct correlation to health and other infrastructure indicators

- Are there any interesting relationships that might come out as a result of this exploring this question?

Goals

- Interpretation is primary
- Predictability is secondary



Datasets

2 datasets were used in the study, the Global Consumption Database and the World Development Indicators Dataset

Global Consumption Dataset

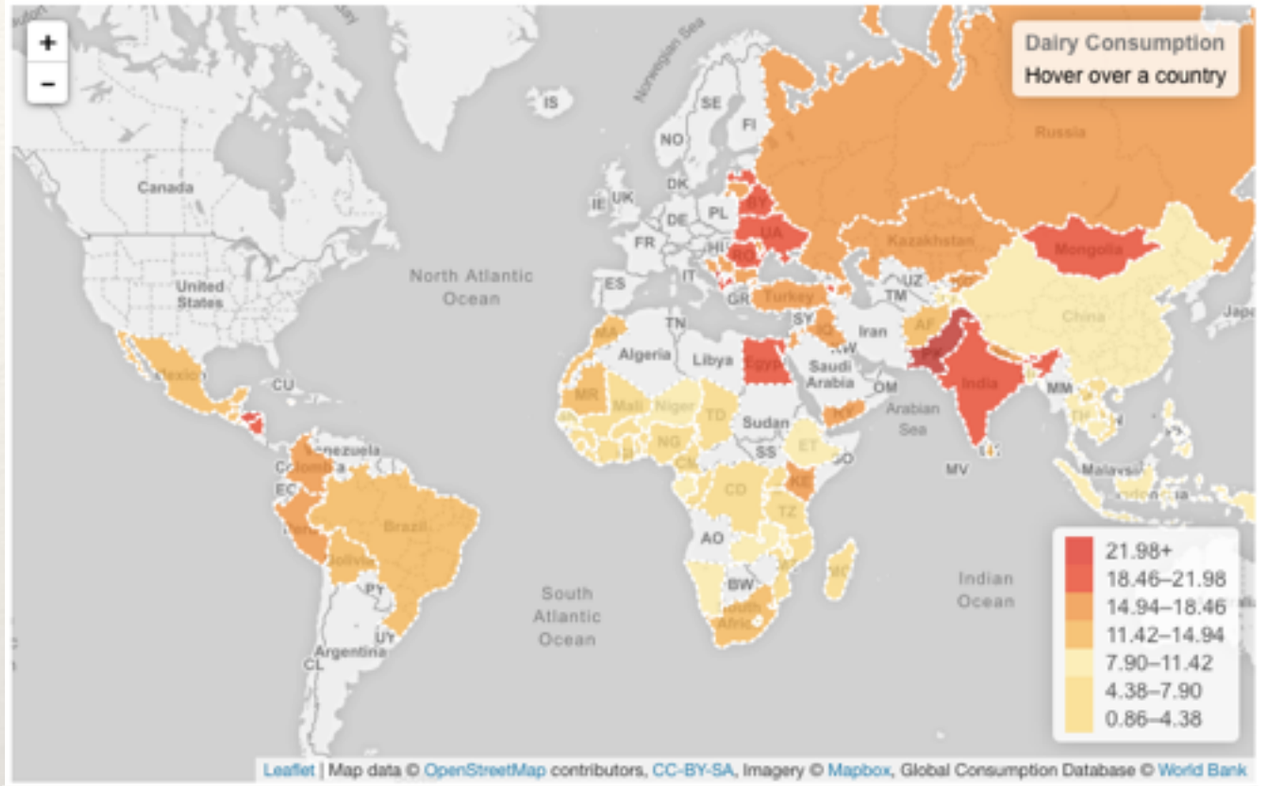
- 2010 breakdown of each Nation's per capita household consumption share broken out by product/service
- i.e. in Afghanistan, pasta is 0.4% of overall hh consumption budget, rent is 8%, cereals/flour products excluding pasta is 17.6%, In Colombia, pasta is 0%, rent is 18% and cereals/flour is 0.0001%
- 2016, World Bank

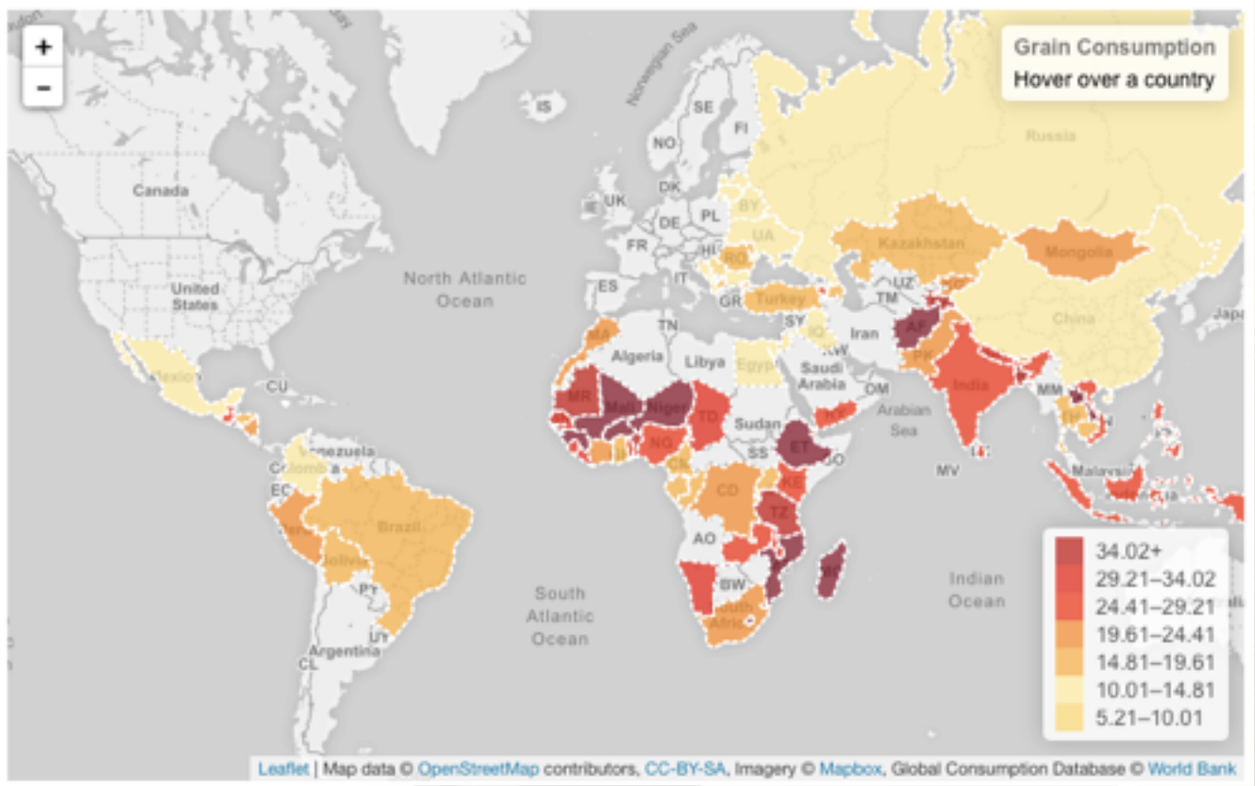
World Development Indicators

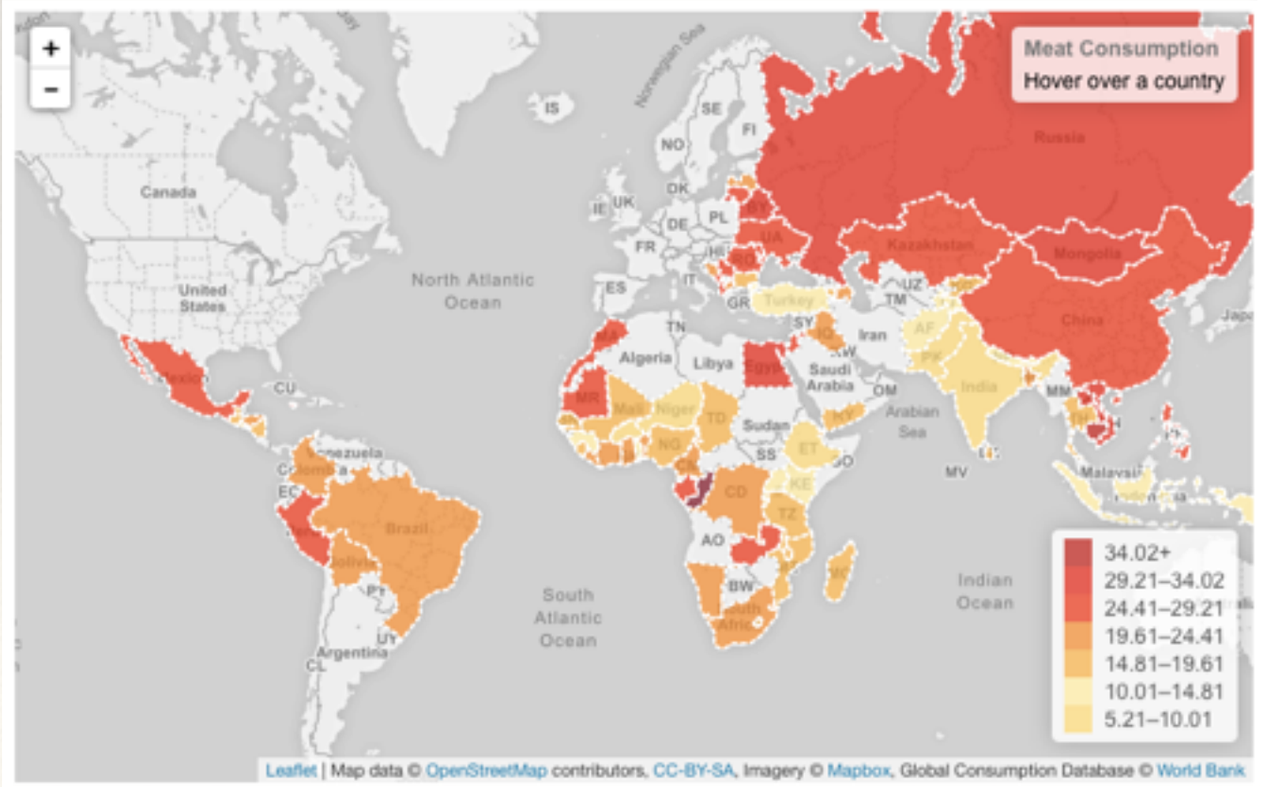
- Collection of 800 development indicators for all countries, separated into 20 categories
- Data for the past 30 years over all categories: Environment, Health, Finance, Climate Change, etc
- Several missing data points for particular countries over certain years
- 2015, World Bank

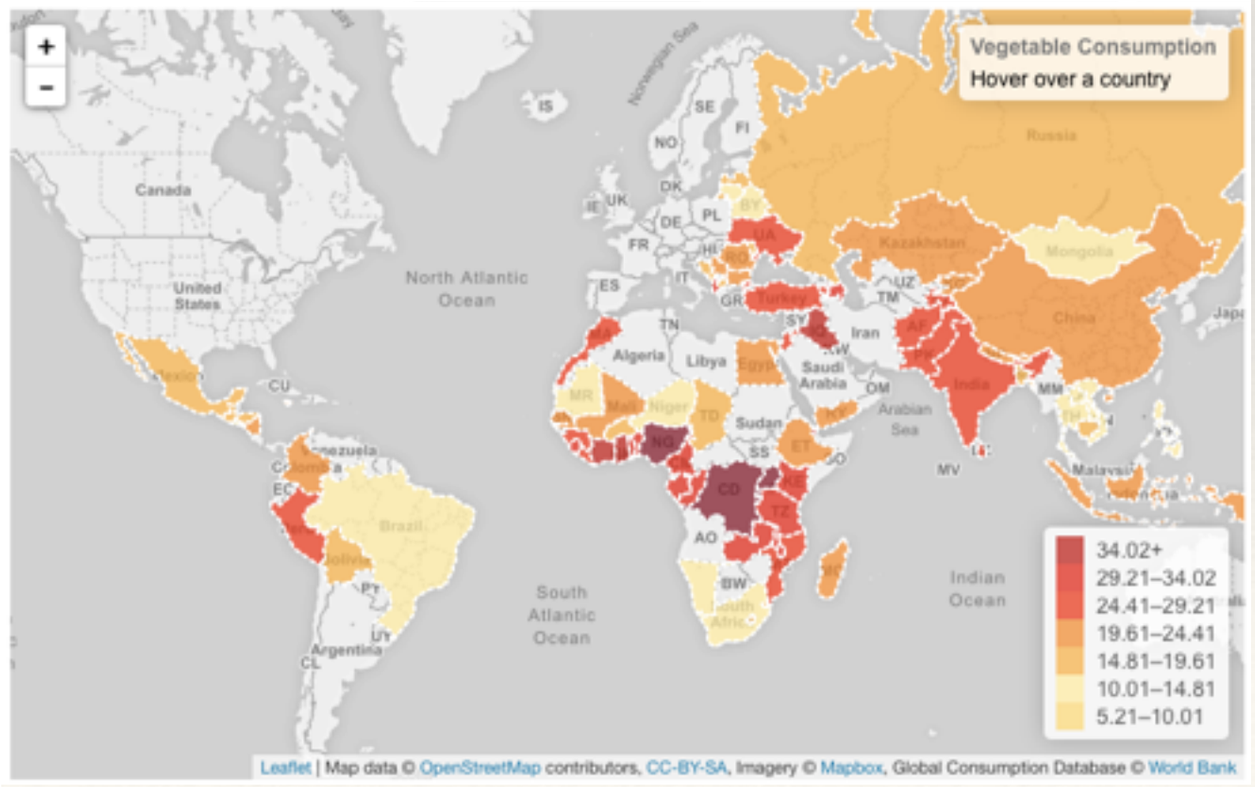
Cleaning & Organizing Data

- For analysis, all food expenditures were aggregated into 6 categories: dairy, meat, grain, vegetable, sugar, beverage, and other (exclusively catering/takeout) and normalized according to food/beverage % of overall consumption
- Tried to focus on health indicators first, and then spread out to environment and poverty indicators
- Identified 30 relevant indicators from WDI dataset, attempted to fill any missing information with the previous years information (presentation focuses on four : life expectancy, birth rate, underweight children, and electricity access)
- On data with a low number of missing items, attempted to use linear regression to fill missing data with error
- Only 89 countries were provided in consumption dataset, reducing my initial observation size considerably
- Transformed WDI data by country and joined WDI dataset to consumption dataset, added human readable names for WDI indicator data


Dairy Consumption
Hover over a country
21.98+
18.46–21.98
14.94–18.46
11.42–14.94
7.90–11.42
4.38–7.90
0.86–4.38
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox, Global Consumption Database © World Bank


Grain Consumption
Hover over a country
34.02+
29.21–34.02
24.41–29.21
19.61–24.41
14.81–19.61
10.01–14.81
5.21–10.01
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox, Global Consumption Database © World Bank


Meat Consumption
Hover over a country
34.02+
29.21–34.02
24.41–29.21
19.61–24.41
14.81–19.61
10.01–14.81
5.21–10.01
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox, Global Consumption Database © World Bank


Vegetable Consumption
Hover over a country
34.02+
29.21–34.02
24.41–29.21
19.61–24.41
14.81–19.61
10.01–14.81
5.21–10.01
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox, Global Consumption Database © World Bank

New Research Question

- After running a few models on our indicators, sugar has small to zero effect on all WDI indicators, as compared to other food consumption habits
- Dairy and grain have the highest correlation to the WDI indicators in almost all cases
- Vegetable consumption comes in as a close third and is normally included in dairy and grain as a relevant variable
- Meat consumption plays a smaller role in most indicators but plays a large role when dairy does not
- Since sugar seems to play almost no role, what can we surmise from the other consumption factors?

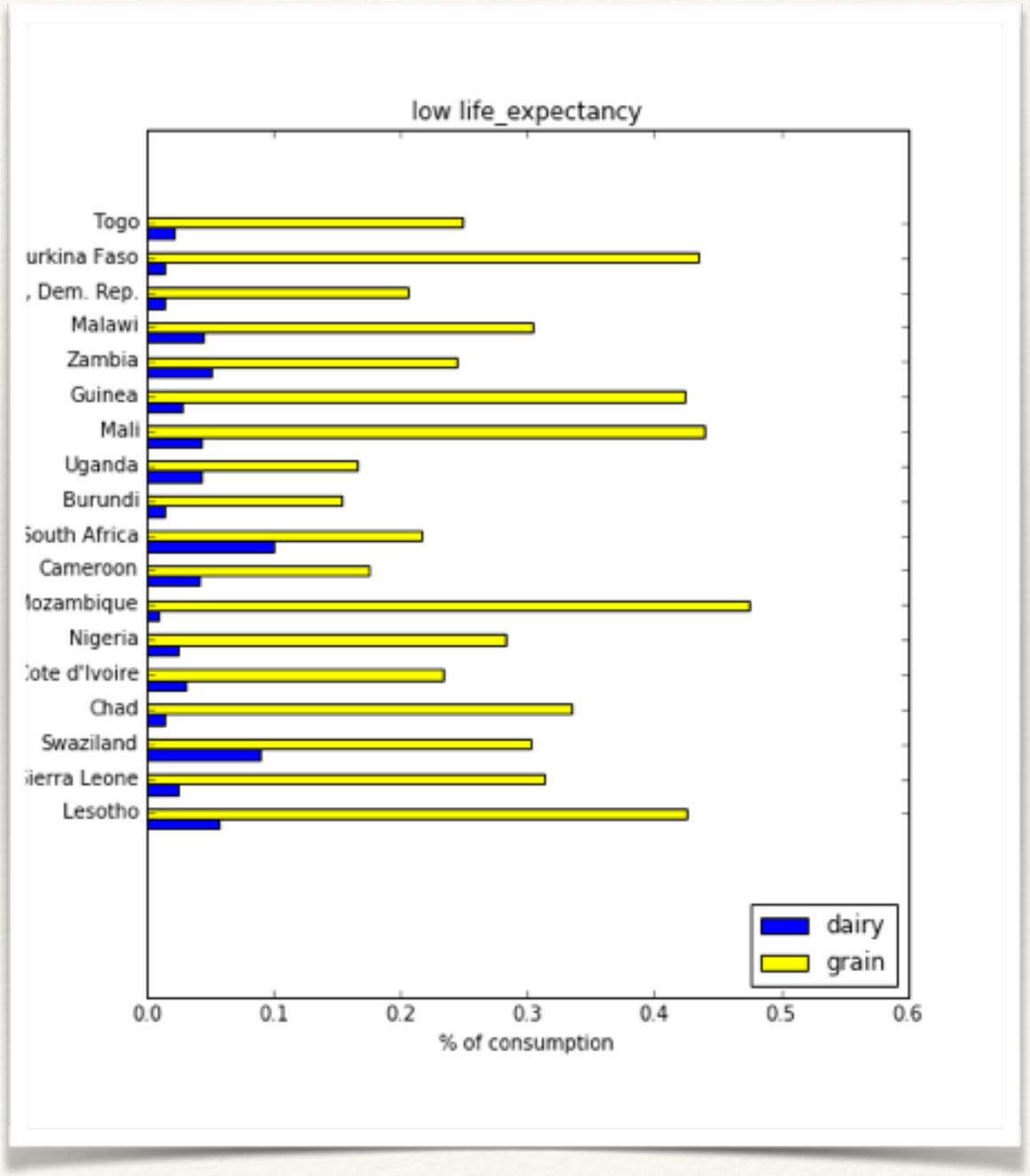
General Approach

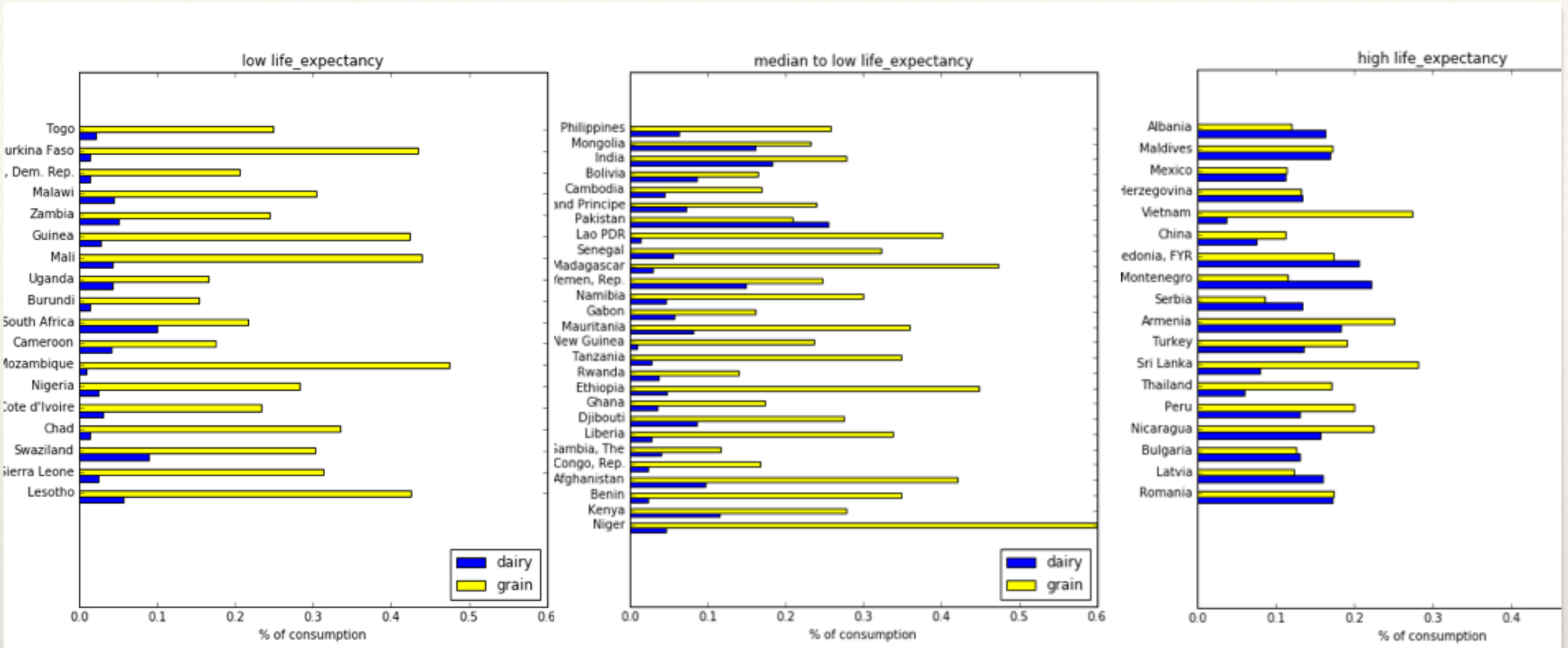
For each output variable, I used 4 models

1. Lasso linear regression to identify the most significant variables
2. Decision tree in an effort to improve results
3. Random forest with tuned for max features
4. Boosting with graphing and tuning for number of estimators and depth of trees

Life Expectancy

- Correlated to grain and dairy % of hh consumption
- Grain has a negative effect on life expectancy while dairy has a positive effect
- Adjusted r-squared was 44.9%
- Boosting won on this indicator narrowly beating random forest with a 4.99 to 5.57 victory in overall error
- Max is 76 and min is 47.48

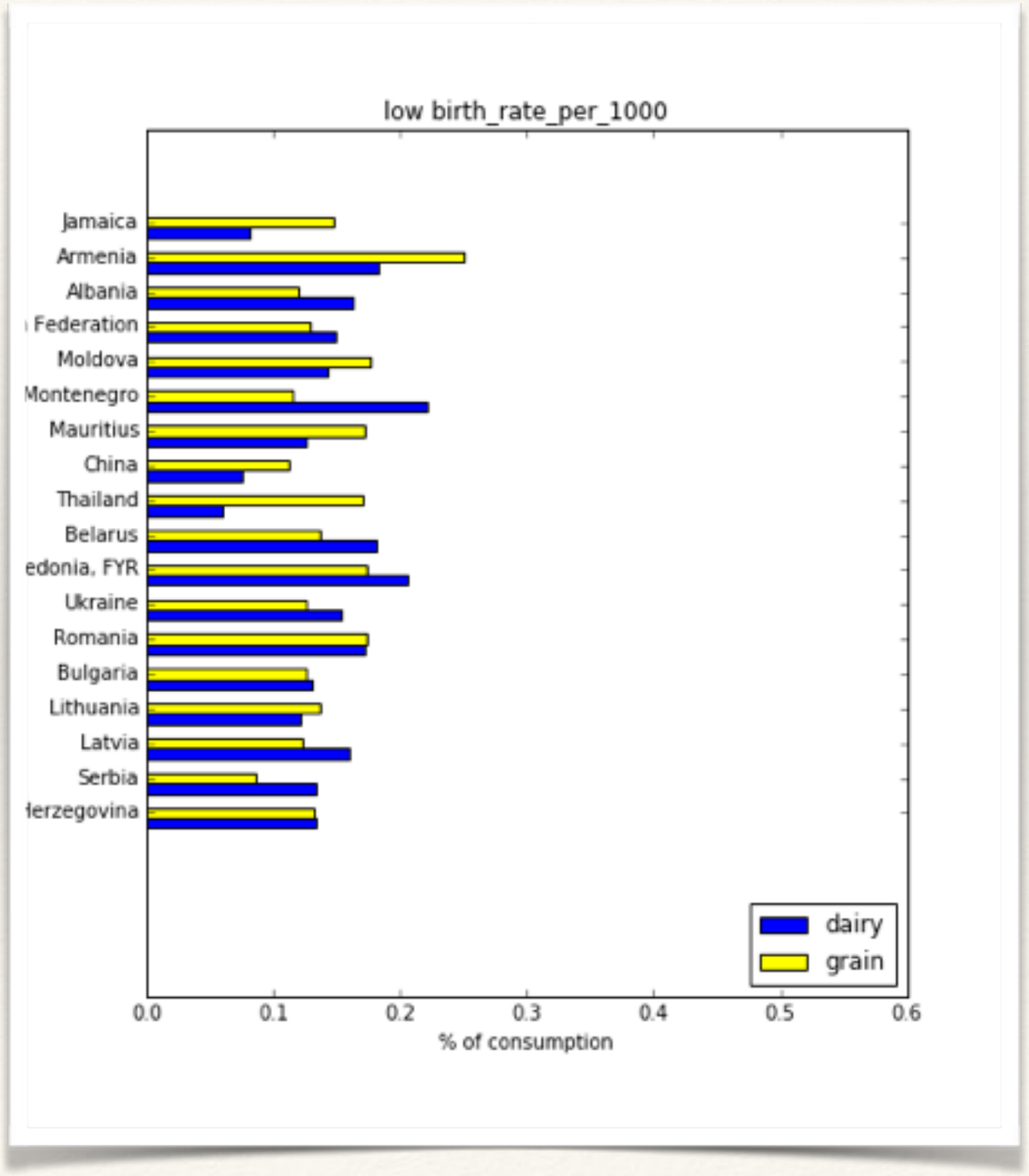


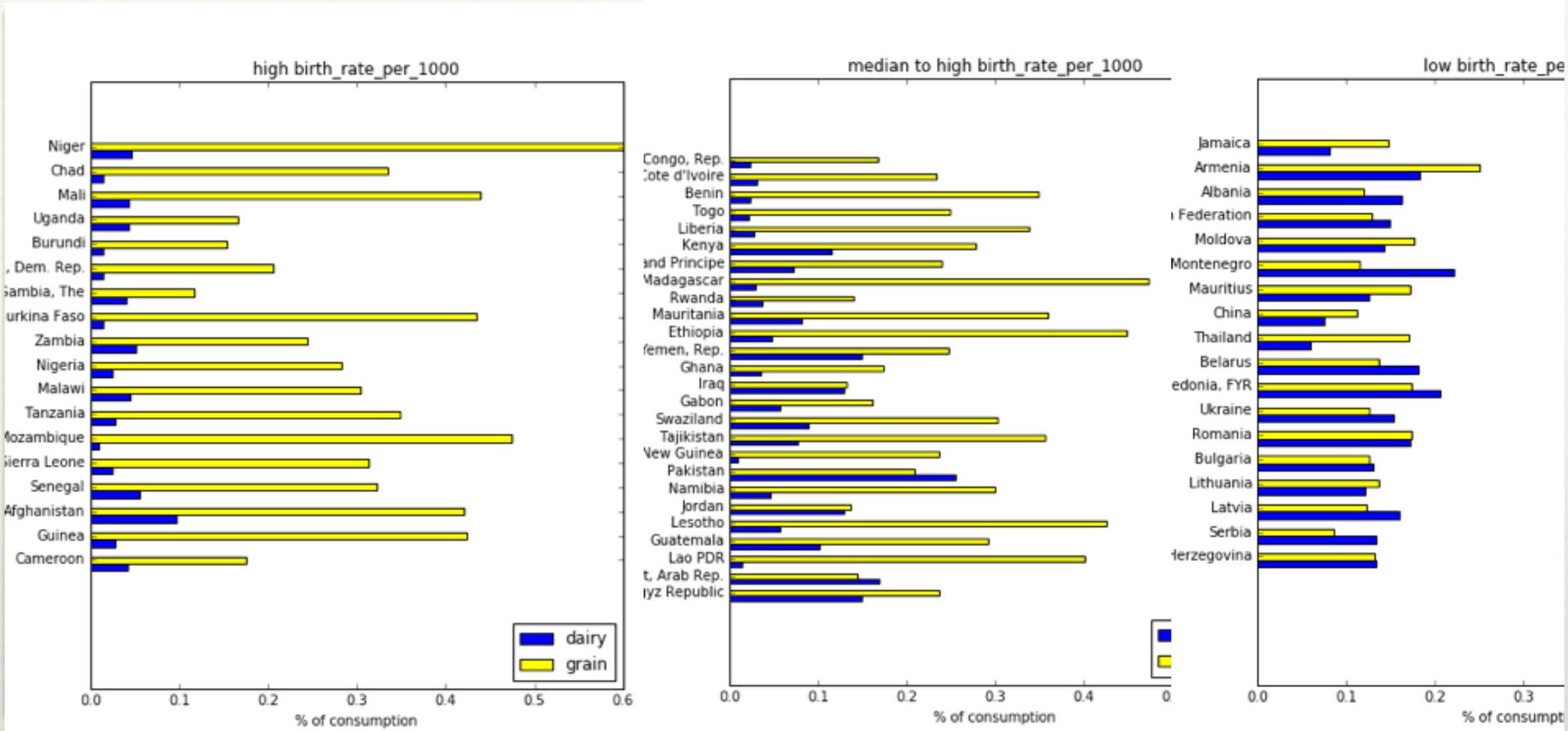


Life Expectancy Graphs

Birth Rate

- Birth rate is closely linked with life expectancy
- In birth rate linear model, grain and vegetable have an equal positive effect on birth rate and dairy has a negative effect
- All three variables seem to have an equal impact
- Boosting model wins over decision tree: 5.88 to 6.5 error rate.

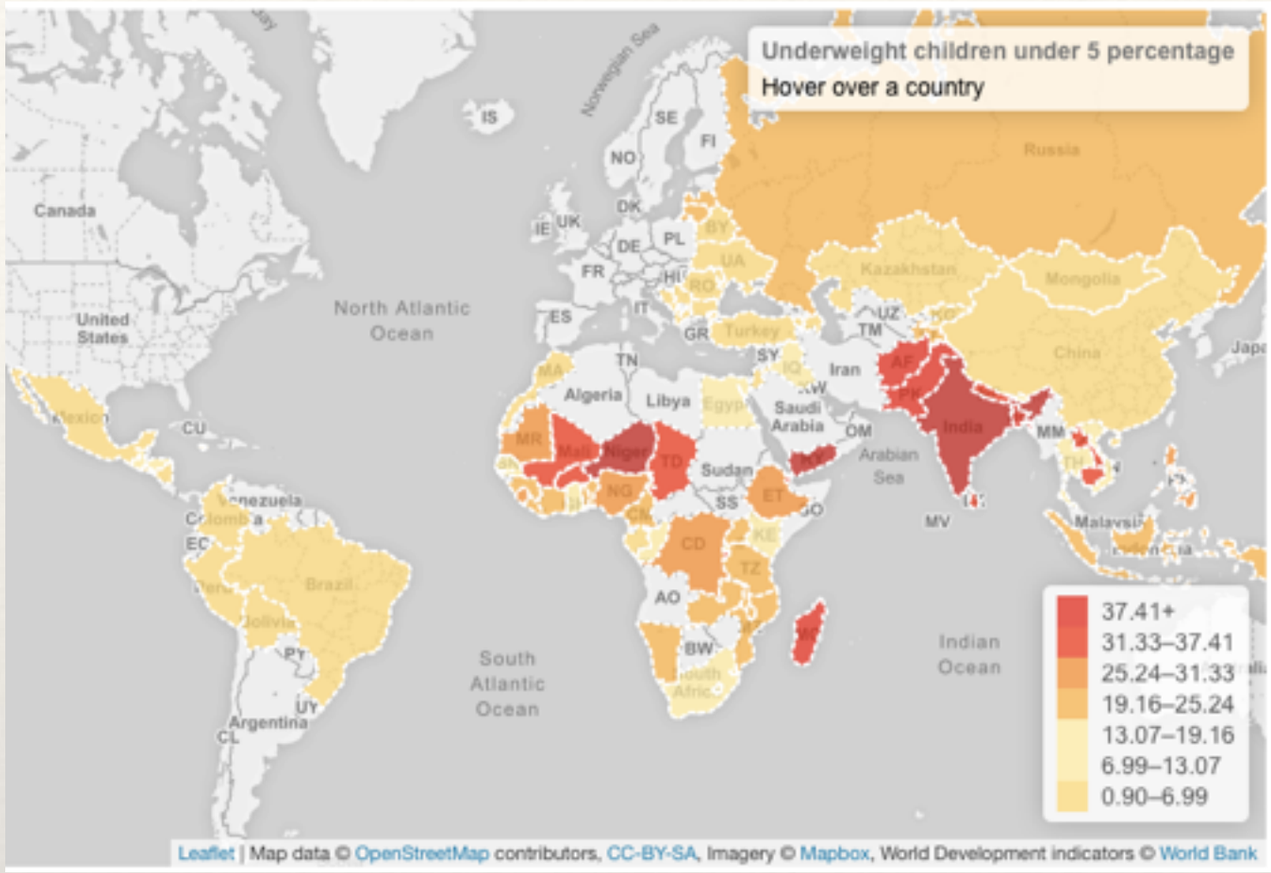


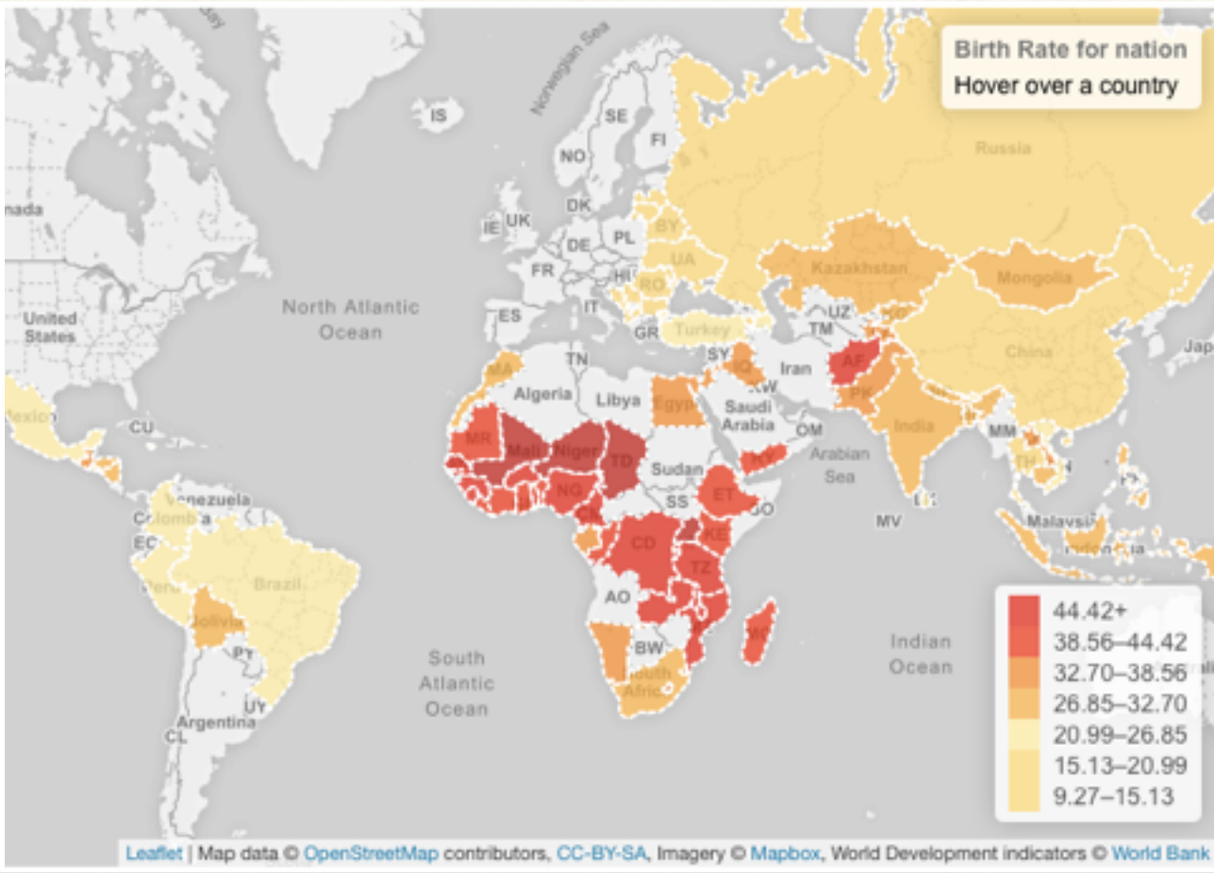


Birth Rate Graphs

Underweight Children < 5

- Underweight is where we start to see some difference between other outputs in this study
- Grain, dairy and meat are the largest factors influencing underweight children
- Grain and dairy have a negative effect on underweight children and meat consumption has a positive effect
- Random forest model won with a 8.84 error compared to linear model error of 9.20

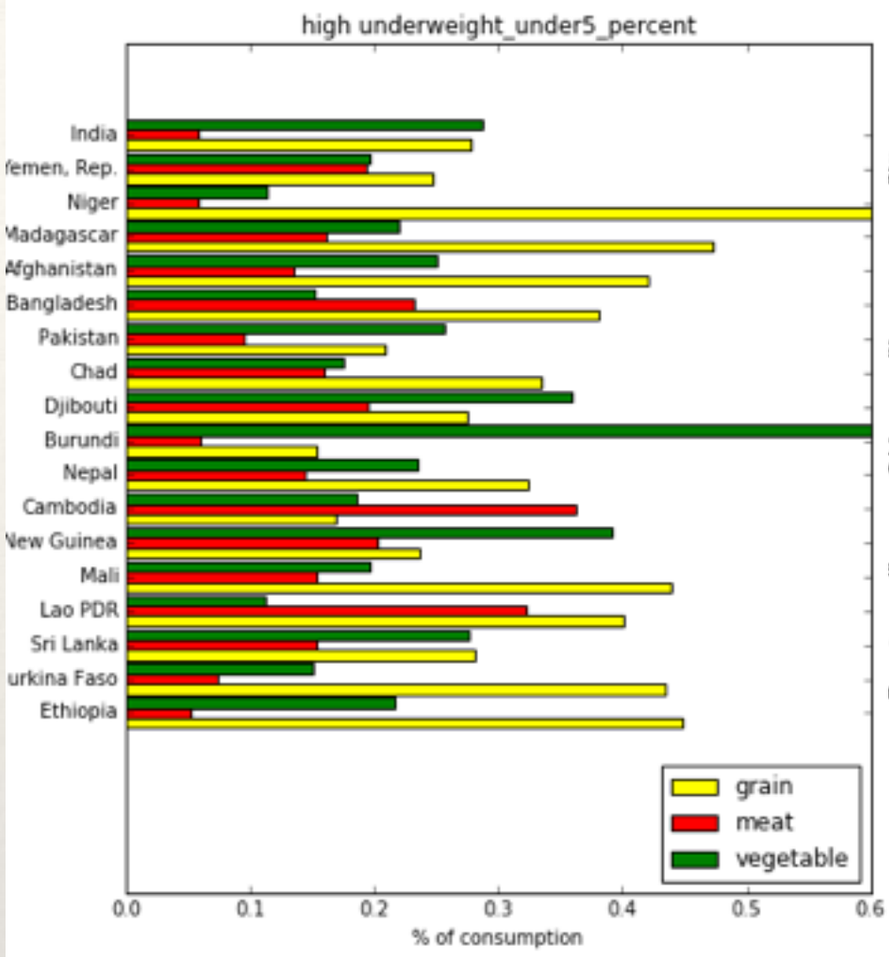


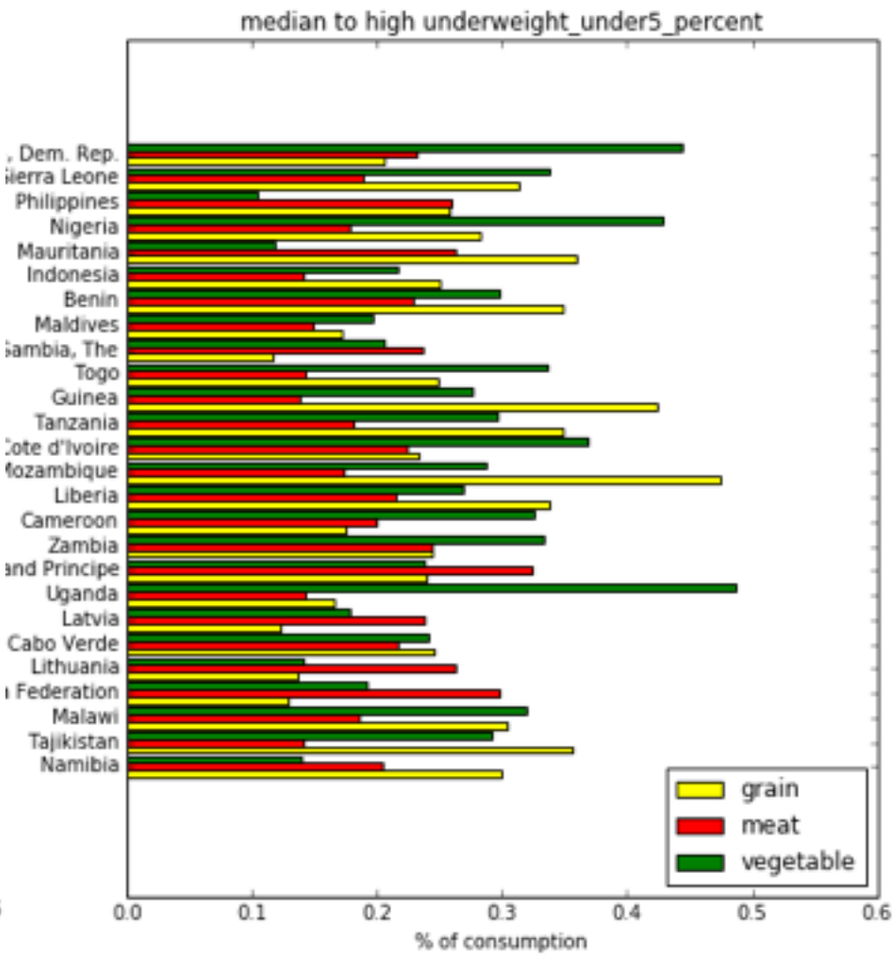


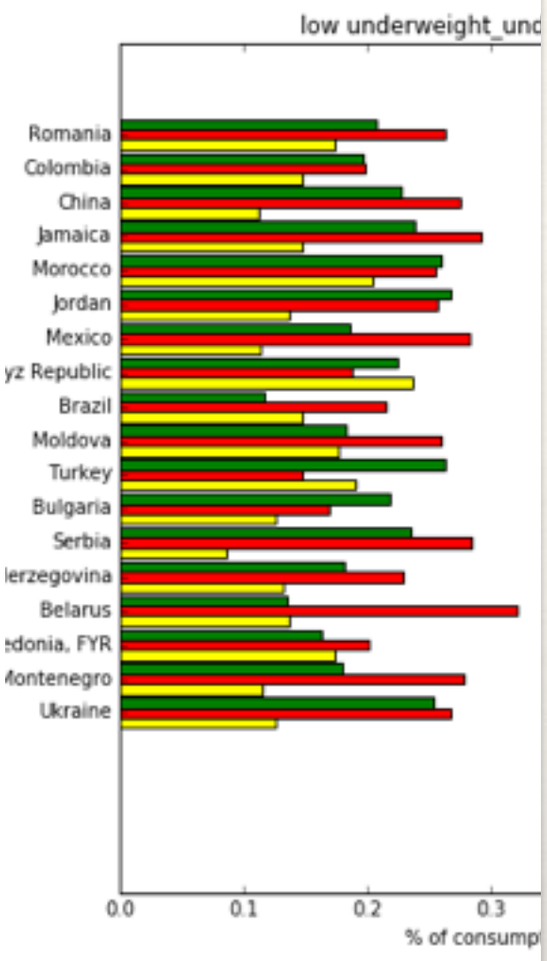
Underweight

Birthrate

Heat Map Differences







Underweight Graphs

Output Correlations

	Homicides	Food_production_index	youth literacy	birth rate	elec. access	underweight	power outages in month	mortality rate	life_expectancy
Homicides	1.000000	-0.150024	0.131891	-0.051898	0.069591	-0.198737	-0.105792	-0.076063	0.011441
Food Production Index	-0.150024	1.000000	-0.397732	0.483412	-0.460679	0.315836	0.111610	0.422071	-0.370076
youth literacy	0.131891	-0.397732	1.000000	-0.759531	0.748767	-0.554746	-0.308871	-0.810196	0.658405
birth rate	-0.051898	0.483412	-0.759531	1.000000	-0.849756	0.540792	0.345953	0.874506	-0.821171
elec access	0.069591	-0.460679	0.748767	-0.849756	1.000000	-0.574614	-0.219976	-0.828720	0.823523
underweight	-0.198737	0.315836	-0.554746	0.540792	-0.574614	1.000000	0.502235	0.585396	-0.451636
power outages	-0.105792	0.111610	-0.308871	0.345953	-0.219976	0.502235	1.000000	0.352317	-0.235327
mortality rate	-0.076063	0.422071	-0.810196	0.874506	-0.828720	0.585396	0.352317	1.000000	-0.908088
life expectancy	0.011441	-0.370076	0.658405	-0.821171	0.823523	-0.451636	-0.235327	-0.908088	1.000000

Findings

- There is no correlation between sugar and the nations health indicators I inspected
- Most countries with long life expectancy normally have a relatively high dairy intake
- In a nation where hh spending a considerable amount on grain then health indicators will be adversely effected
- Youth literacy is highly correlated to electricity access and life expectancy
- Birth rate is highly correlated with underweight children, power outages and child mortality rate

Findings

- Africa dominates the findings and it would be interesting to segment data to see how models perform without using Africa
- Population under poverty line would be a good variable to include
- More data would be ideal, only 89 countries from global consumption dataset
- There are many, many other variables in the dataset that play a much bigger role in the relationship and could provide a more accurate prediction of these outputs in the future
- The consumption dataset is fairly fine grained and would be interesting to do a study on an even smaller window (i.e. How do nations that consume more poultry do in relation to others, etc)
- Interesting facts from global consumption dataset: spending patterns—the shares of spending people allocate to different items—are remarkably similar across all levels of total spending. The most notable differences are in food and beverages, whose share decreases as total spending goes up, and in transport, whose share increases the most as spending rises.

All maps located at

http://bit.ly/1qiAt75

Thanks!