Oh Baby!

-Data Science Project-

Kao Zi Jian

Zijiankao@gmail.com

GA DAT-13

Mar-Apr 2019

Contents

To Shi An

“Human progress is a march not into the light, but into ever-greater vulnerability”

– *Wang Dong Yue, independent scholar*

1. Introduction

We all have some questions which linger in our heads, refusing to go away. For me, one such is:

"Why is it that people in developed countries have fewer children than their counterparts in less developed ones? Why does greater physical well-being correlate with lower, instead of higher, fertility?"

As a new father, and curious as to the factors influencing fertility in general around the world, **I set out to predict the 2018 Total Fertility Rate (TFR)[1] of any given country, based on a selection of socioeconomic features**

TFR (Informal): Number of children born to a woman over her lifetime

Quick Stats for 2018:[2]

- Highest TFR: 7.15 (NIGER)
- Lowest TFR: 1.22 (TAIWAN)
- SINGAPORE: 1.26

Target Feature: Individual countries' TFR for 2018

- ➔ Continuous variable (float)
- ➔ **NOT** a time-series analysis
- ➔ Max. number of samples: **264** (since 264 countries in the world, according to World Bank)

Independent Features:

- ➔ Socioeconomic data of sample countries
- ➔ Find set of features correlated to TFR
- ➔ Features to be as mutually distinct as possible

[1] Formally, TFR is "the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year." – *World Bank*

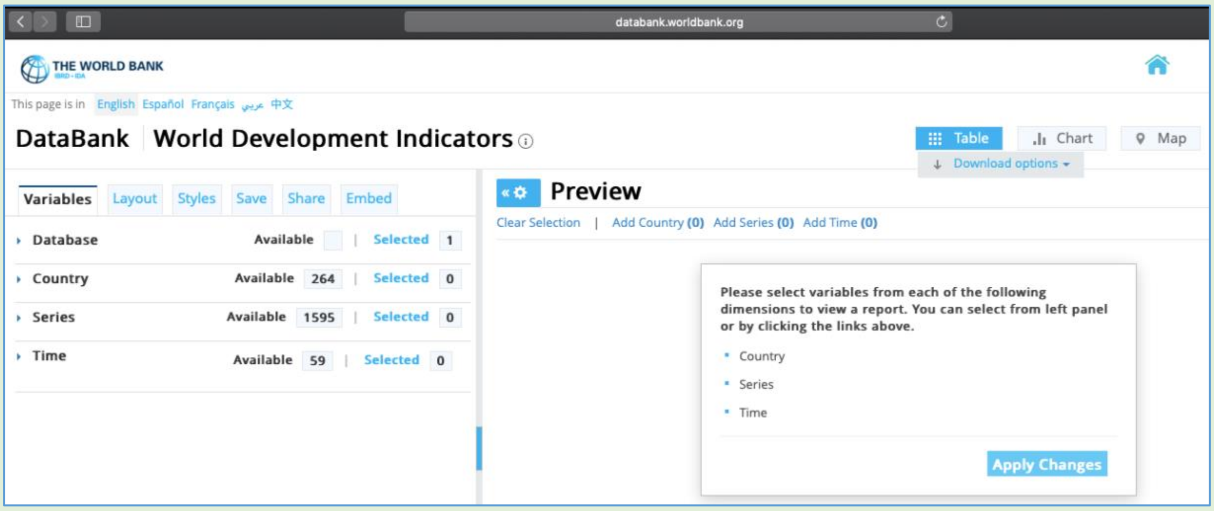
[2] Worldpopulationreview.com

2. Pre-Processing

2.1 Data Assembly

-Source of Data-

Was fortunate to have a one-stop online source for international socioeconomic data in the World Bank's **World Development Indicators** dataset



DataBank | World Development Indicators

Clear Selection | Add Country (0) Add Series (0) Add Time (0)

Database	Available		Selected	1
Country	Available	264	Selected	0
Series	Available	1595	Selected	0
Time	Available	59	Selected	0

Please select variables from each of the following dimensions to view a report. You can select from left panel or by clicking the links above.

- Country
- Series
- Time

Dataset has 3 primary filters:

Filter	Elements	Description
"Countries"	264 countries	Exhaustive set of countries/political regions (including, e.g. "West Bank and Gaza")
"Series"	1595 series	Wide range of data series (e.g. "GDP Per Capita") available across these domains: • Economic • Political • Financial • Environmental • Social • Developmental • Infrastructure
"Time"	59 years	Annual data from 1960 to 2018

Limitations of the dataset:

i) Uneven data availability across Countries/Series/Time

- ➔ e.g. for Series A, Country X has no data, Country Y has data for some years, Country Z data for all years

ii) Some Series not amenable to cross-country comparison

- ➔ Especially for Series involving subjective reporting by interviewees, or those which lack a uniform standard of measurement across countries

- Feature Selection-

Given the small sample size of my dataset (max 264 countries = samples), I could not afford to have irrelevant/weak features. These would introduce additional noise to mislead the model

- ➔ If I were to select all 1595 Series as features, likely to result in high bias and high variance[3]

- ➔ If I were to have just 1 feature, then under-fitting

Decided best to err on the side of fewer features, as downside was relatively smaller

Having settled on a lean model, next question was how to select the features

Since model was lean, decided to handpick the features, based on the following criteria:

- ➔ Decent correlation with target (TFR)
- ➔ Data available from most countries
- ➔ Collectively, the features should not overlap too much with each other

[3] If sample size was large, there would then be low bias, high variance (i.e. "over-fitting")

Since different features/combinations of features met the criteria to varying extents, much recursive weighing/judgment was needed

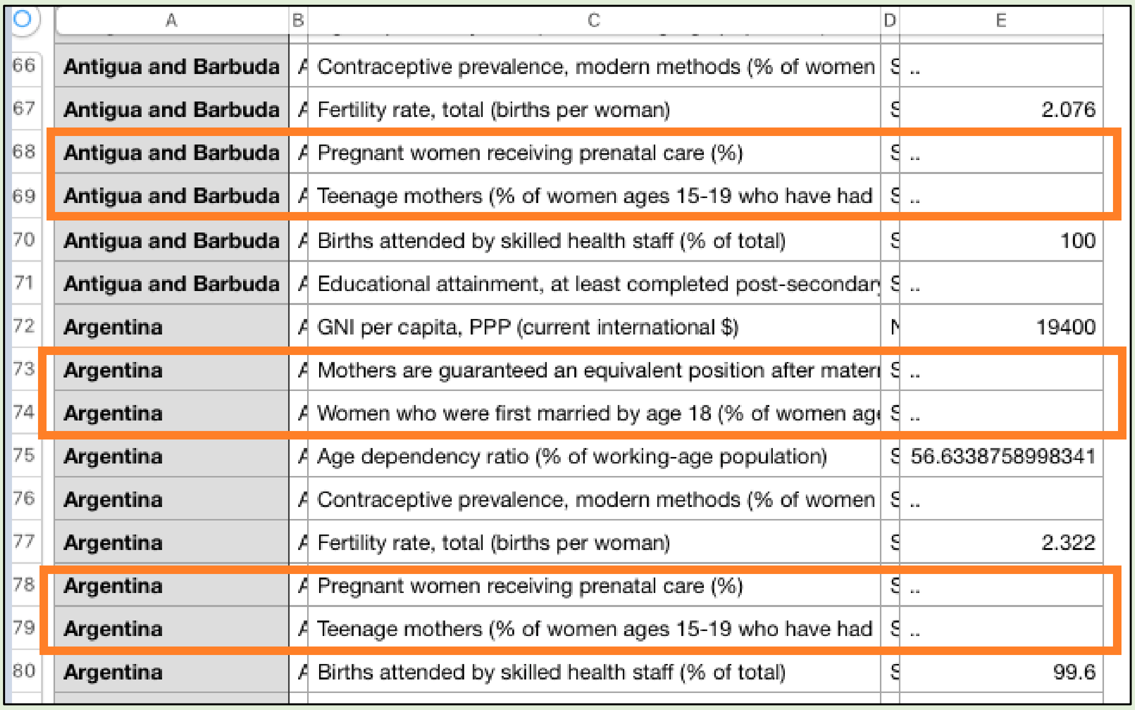
Example

Which is superior—

- Series A: good target correlation, poor data availability across countries
- Series B: weaker target correlation, stronger data availability

On the face of it, Series A wins. Problem, however, is that the poor data availability meant a choice between:

- ➔ Eliminating the countries lacking the data, thus reducing the number of sample countries (from small to smaller); **OR**
- ➔ Preserving the sample size and imputing values to those countries lacking data



	A	B	C	D	E
66	**Antigua and Barbuda**	A	Contraceptive prevalence, modern methods (% of women	S	..
67	**Antigua and Barbuda**	A	Fertility rate, total (births per woman)	S	2.076
68	**Antigua and Barbuda**	A	Pregnant women receiving prenatal care (%)	S	..
69	**Antigua and Barbuda**	A	Teenage mothers (% of women ages 15-19 who have had	S	..
70	**Antigua and Barbuda**	A	Births attended by skilled health staff (% of total)	S	100
71	**Antigua and Barbuda**	A	Educational attainment, at least completed post-secondar	S	..
72	**Argentina**	A	GNI per capita, PPP (current international $)	N	19400
73	**Argentina**	A	Mothers are guaranteed an equivalent position after materi	S	..
74	**Argentina**	A	Women who were first married by age 18 (% of women ag	S	..
75	**Argentina**	A	Age dependency ratio (% of working-age population)	S	56.6338758998341
76	**Argentina**	A	Contraceptive prevalence, modern methods (% of women	S	..
77	**Argentina**	A	Fertility rate, total (births per woman)	S	2.322
78	**Argentina**	A	Pregnant women receiving prenatal care (%)	S	..
79	**Argentina**	A	Teenage mothers (% of women ages 15-19 who have had	S	..
80	**Argentina**	A	Births attended by skilled health staff (% of total)	S	99.6

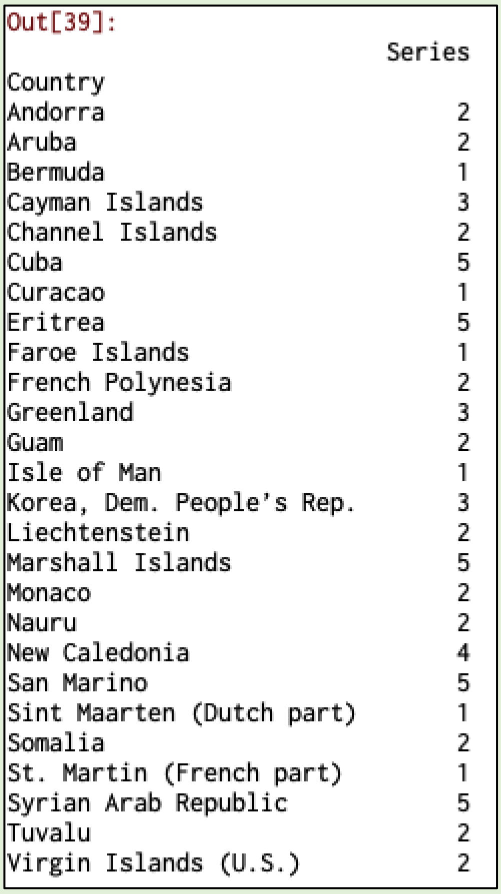
Neither was easy to stomach:

- ➔ Reducing an already small sample size; OR
- ➔ Imputing values based on a small sample size (very dangerous, especially with unsophisticated imputation (e.g. simple average of all countries with data) and/or for series where the majority of countries are lacking data)

Therefore, I tended to be more sympathetic to Type B, over Type A, Series, as data availability was proving to be a high priority

Feature selection was thus a protracted affair to find an optimal combination of features which best met my search criteria.

Sample of countries dropped along the way— mostly small, minor countries:



```
Out[39]:
                                 Series
Country
Andorra                               2
Aruba                                 2
Bermuda                               1
Cayman Islands                        3
Channel Islands                       2
Cuba                                  5
Curacao                               1
Eritrea                               5
Faroe Islands                         1
French Polynesia                      2
Greenland                             3
Guam                                  2
Isle of Man                           1
Korea, Dem. People's Rep.             3
Liechtenstein                         2
Marshall Islands                      5
Monaco                                2
Nauru                                 2
New Caledonia                         4
San Marino                            5
Sint Maarten (Dutch part)             1
Somalia                               2
St. Martin (French part)              1
Syrian Arab Republic                  5
Tuvalu                                2
Virgin Islands (U.S.)                 2
```

Final result: 6 independent features and 179 samples (countries)

2.2 Data Cleaning

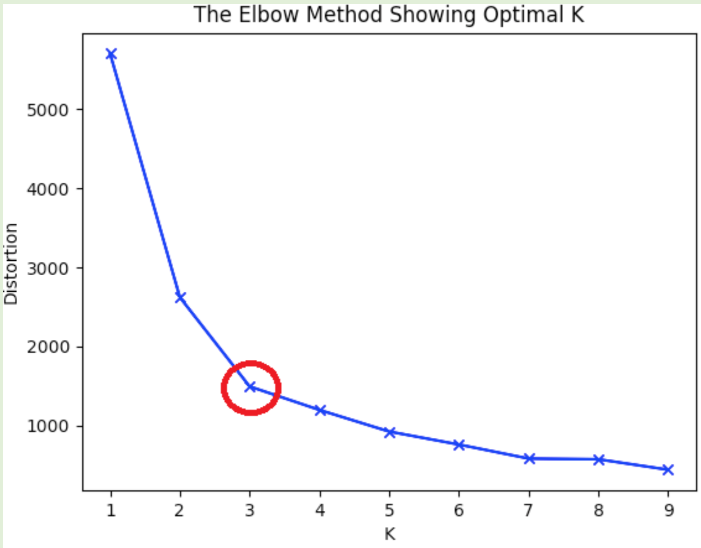
-Data Imputation-

My final dataset for analysis comprised 179 samples (countries) x 6 features. Would have preferred sample count >200, but this was still acceptable. Final task before modelling was to impute the missing values.

Conscious that a simple average was too distortionary, I decided to group the 179 countries into clusters and impute based on the cluster averages. How should they be clustered?

First thought–by geography. However, this would not account for country differences within the same region which were far from the mean. Second, there would be no comparability across regions.

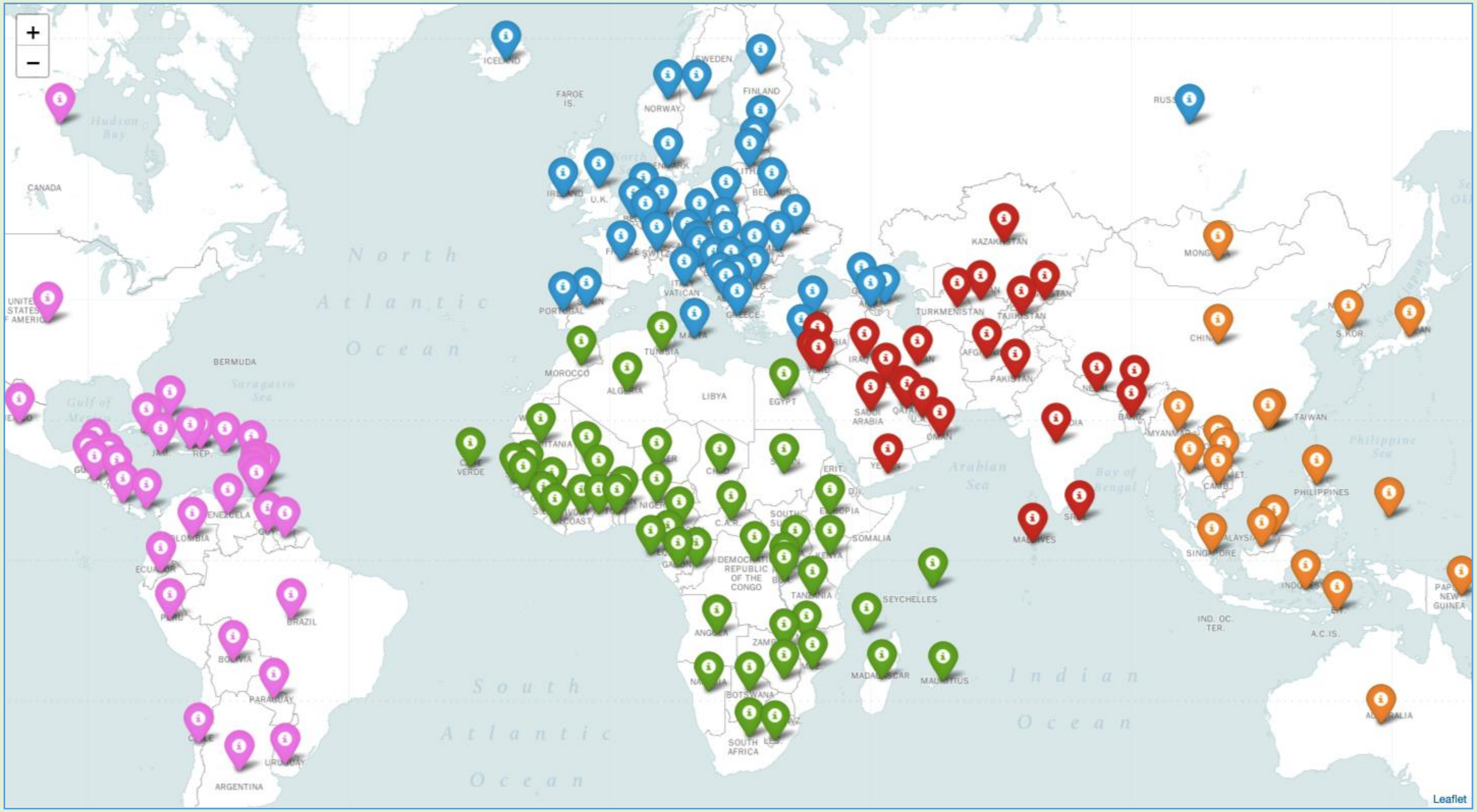
Hence decided to use the Scikit-Learn clustering algorithm to group the countries, based on the 6 features chosen (missing values were imputed by feature-averaging). Though optimal K=3 by analysis, I called 4 clusters in the algorithm in order to isolate a marginal cluster (3 main + 1 marginal).



With cluster-based imputation complete, data was now clean and ready for modelling

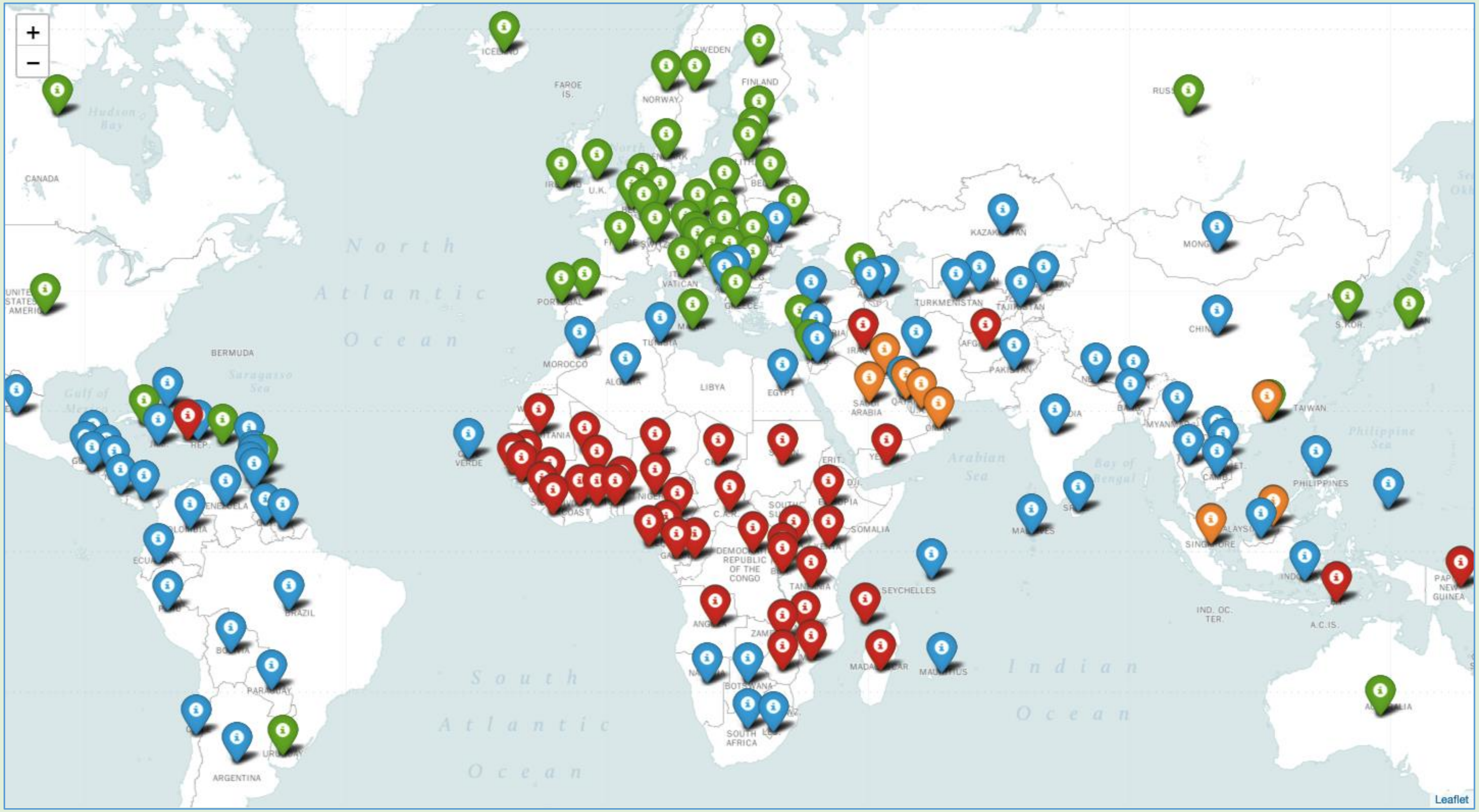
Grouping by Geography for 179 sample countries

[map created using Python Folium library]

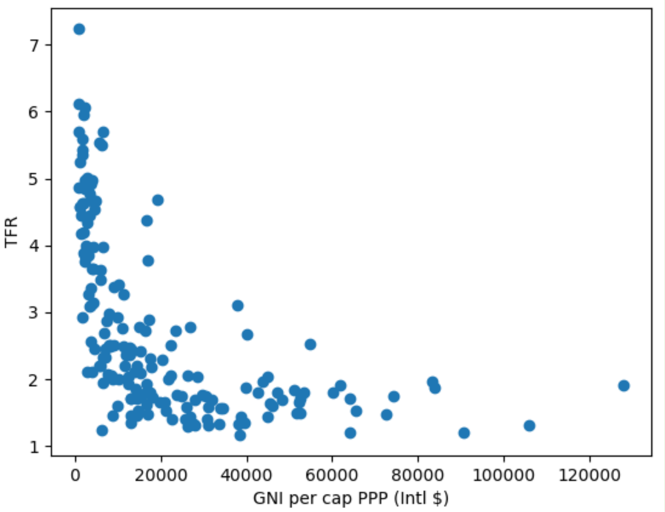
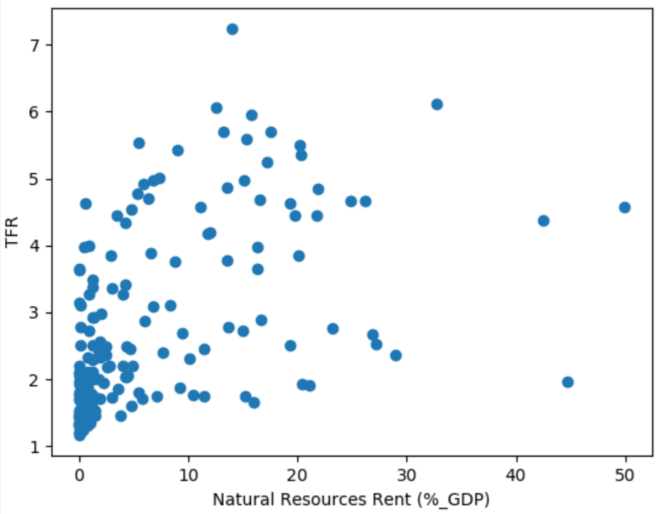
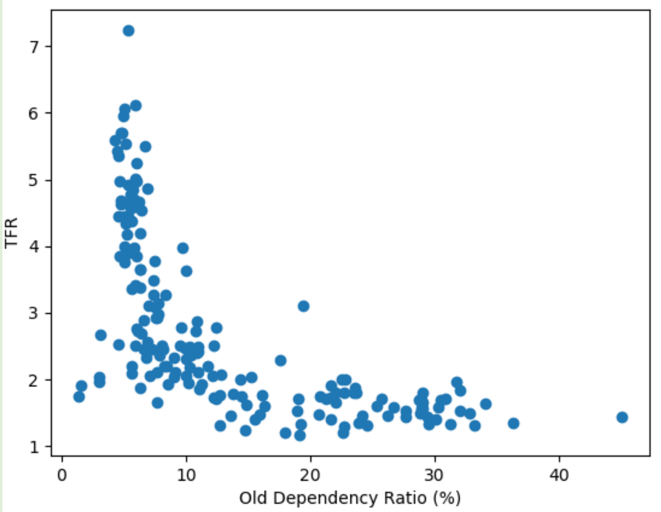


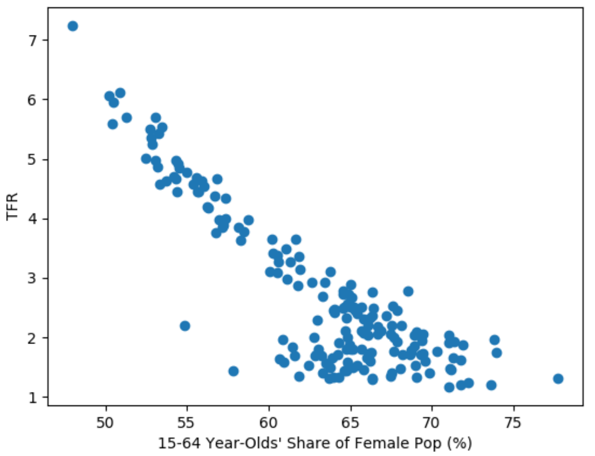
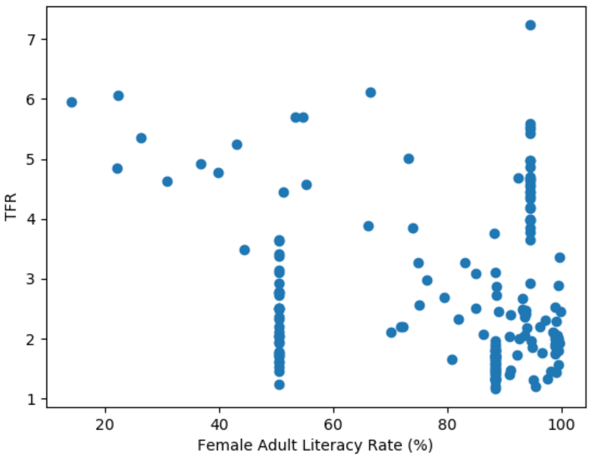
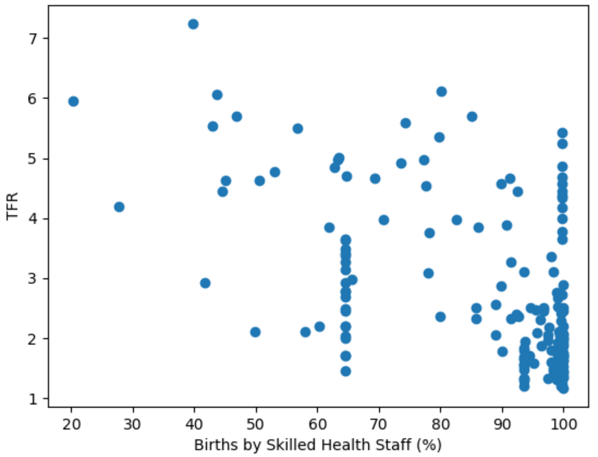
Grouping by KMeans Clustering Algorithm (K=4) for 179 sample countries

[map created using Python Folium library]



Meet the Team

<table>
<tr><th colspan="3">Target: TFR</th></tr>
<tr><th>Feature 1</th><th>Feature 2</th><th>Feature 3</th></tr>
<tr><td><b>Gross National Income Per Capita</b>
<i><b>(@ Purchasing Power Parity, Current International Dollars)</b></i></td><td><b>Natural Resources Rent</b>
<i><b>(% of Gross Domestic Product)</b></i></td><td><b>Old-age Dependency Ratio</b>
<i><b>(Seniors >64 years as % of working population)</b></i></td></tr>
<tr><td>Per capita income for a country’s citizens (including those based abroad), where 1 unit of income (Int’l $) can purchase an equivalent basket of goods domestically as 1USD does in the US. Very suited for cross-country comparisons of economic strength</td><td>Sum of oil, natural gas, coal (hard and soft), mineral, and forest rents (i.e. income earned from the extraction and sale of these resources as % of GDP). Indicative of how much of a country’s GDP is due to a “free lunch”</td><td>Ratio of older dependents—people older than 64—to the working-age population—those ages 15-64. Indicative of the fiscal burden on the working-age (and reproductive) population
<b>NOTE: some target leak with this feature</b></td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
</table>

Target: TFR		
Feature 4	**Feature 5**	**Feature 6**
Share of 15-64 year-olds in Female Population ***(% of Female Population)***	**Adult Female Literacy**[4] ***(% of Adult Female Population)***	**Births Attended by Skilled Health Staff** ***(% of Births)***
Female population between the ages 15 to 64 as a percentage of the total female population **NOTE: some target leak with this feature**	% of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life. Studies have shown this is a crucial factor governing the transition from high to low TFR	% of deliveries attended by personnel trained to care for and advise pregnant women; to conduct deliveries on their own; and to care for new-borns.
		

[4] In case you were wondering why "Prevalence of Contraceptives" does not show up—two reasons: poor data availability (even poorer than Female Literacy) and the fact that Female Literacy subsumes Contraceptives (viz with high literacy and greater knowledge levels, use of contraceptives increases/absence of contraceptives can be mitigated, while with low literacy, even an abundance of contraceptives is pointless)

3. Modelling

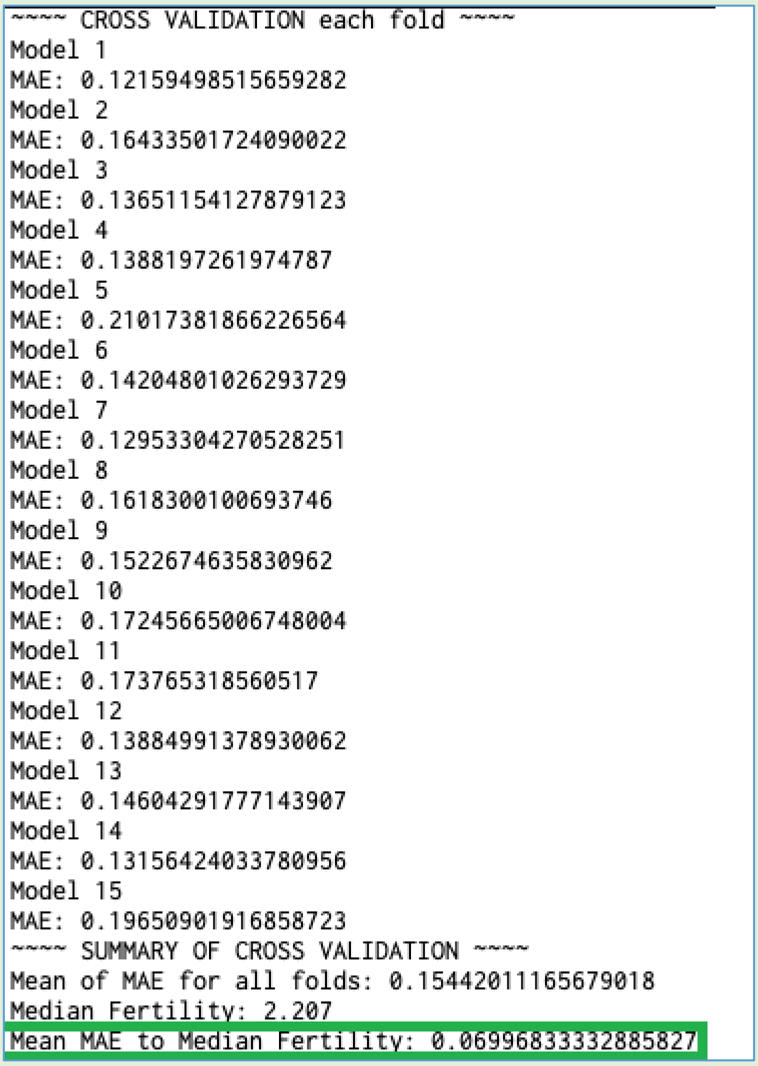
Train-Test split, test_size = 0.2

RepeatedKFold Cross Validation: n_splits=5, n_repeats =3

Metric: Mean Absolute Error (MAE)

3.1 Linear Regression (Scaled)

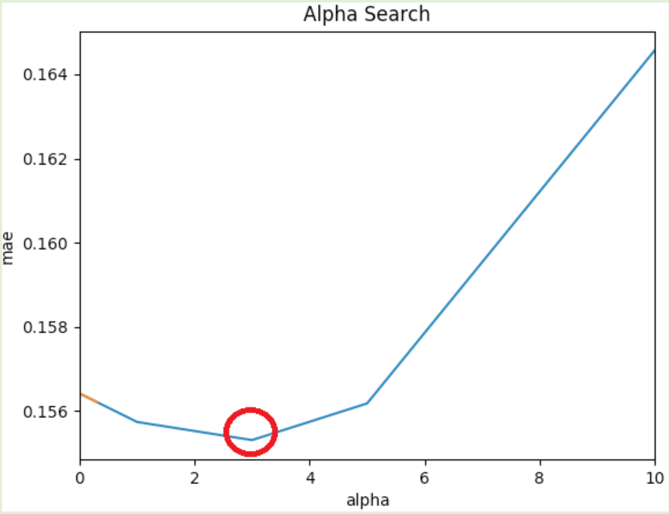
CV Results:

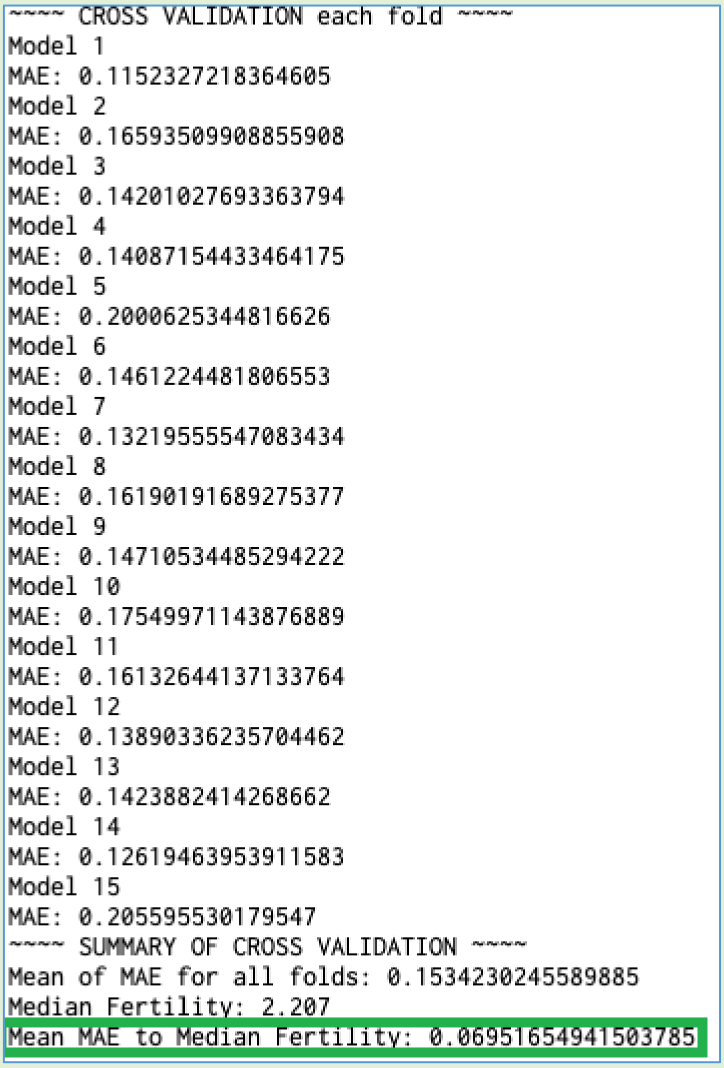


```
~~~~ CROSS VALIDATION each fold ~~~~
Model 1
MAE: 0.12159498515659282
Model 2
MAE: 0.16433501724090022
Model 3
MAE: 0.13651154127879123
Model 4
MAE: 0.1388197261974787
Model 5
MAE: 0.21017381866226564
Model 6
MAE: 0.14204801026293729
Model 7
MAE: 0.12953304270528251
Model 8
MAE: 0.1618300100693746
Model 9
MAE: 0.1522674635830962
Model 10
MAE: 0.17245665006748004
Model 11
MAE: 0.173765318560517
Model 12
MAE: 0.13884991378930062
Model 13
MAE: 0.14604291777143907
Model 14
MAE: 0.13156424033780956
Model 15
MAE: 0.19650901916858723
~~~~ SUMMARY OF CROSS VALIDATION ~~~~
Mean of MAE for all folds: 0.15442011165679018
Median Fertility: 2.207
Mean MAE to Median Fertility: 0.06996833332885827
```

3.2 Regularised Linear Regression

I chose an L1-regularised LR (regularisation moderates outliers), or Ridge Regression [Ridge(alpha=3)]

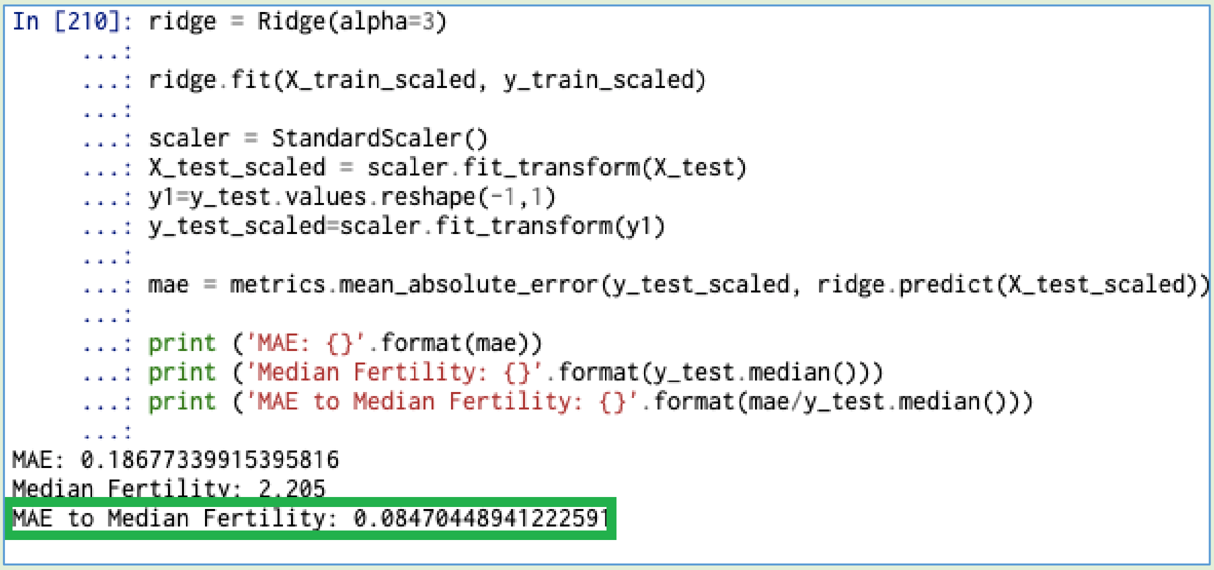




```
~~~~ CROSS VALIDATION each fold ~~~~
Model 1
MAE: 0.1152327218364605
Model 2
MAE: 0.16593509908855908
Model 3
MAE: 0.14201027693363794
Model 4
MAE: 0.14087154433464175
Model 5
MAE: 0.2000625344816626
Model 6
MAE: 0.1461224481806553
Model 7
MAE: 0.13219555547083434
Model 8
MAE: 0.16190191689275377
Model 9
MAE: 0.14710534485294222
Model 10
MAE: 0.17549971143876889
Model 11
MAE: 0.16132644137133764
Model 12
MAE: 0.13890336235704462
Model 13
MAE: 0.1423882414268662
Model 14
MAE: 0.12619463953911583
Model 15
MAE: 0.205595530179547
~~~~ SUMMARY OF CROSS VALIDATION ~~~~
Mean of MAE for all folds: 0.1534230245589885
Median Fertility: 2.207
Mean MAE to Median Fertility: 0.06951654941503785
```

4. Testing

4.1 Testing the Holdout Set



```
In [210]: ridge = Ridge(alpha=3)
     ...:
     ...: ridge.fit(X_train_scaled, y_train_scaled)
     ...:
     ...: scaler = StandardScaler()
     ...: X_test_scaled = scaler.fit_transform(X_test)
     ...: y1=y_test.values.reshape(-1,1)
     ...: y_test_scaled=scaler.fit_transform(y1)
     ...:
     ...: mae = metrics.mean_absolute_error(y_test_scaled, ridge.predict(X_test_scaled))
     ...:
     ...: print ('MAE: {}'.format(mae))
     ...: print ('Median Fertility: {}'.format(y_test.median()))
     ...: print ('MAE to Median Fertility: {}'.format(mae/y_test.median()))
     ...:
MAE: 0.18677339915395816
Median Fertility: 2.205
MAE to Median Fertility: 0.08470448941222591
```

4.2 Passage to India

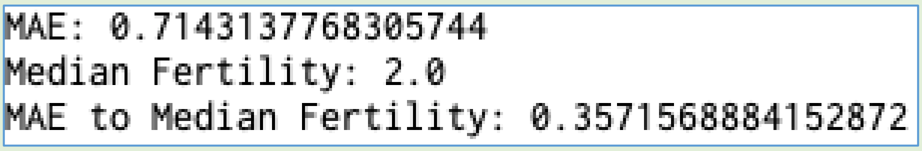
As a final piece, I wanted to see how well the model fared when it came to intra-country TFR predictions. Naturally, I looked for a big country. Enter India.

This time there was no one-stop data source. Data for each feature had to culled from multiple sources. For some features, data corresponding to the feature description could not be found, and I had to use close substitutes instead. In the case of one particular feature lacking data ("Natural Resources Rent"), I had to use a blanket value (=the country value) for each sample state.

After data pre-processing, I ended up with a testable set of 17 states (out of a possible 29). Not too bothered as these 17 covered most of India.

Hoped to see that the MAE as % of median TFR in the India test set was close to that obtained for the country test (i.e. **8.5%**).

Result:



```
MAE: 0.7143137768305744
Median Fertility: 2.0
MAE to Median Fertility: 0.3571568884152872
```

Turns out, much worse than the country result (36% vs 8.5%).

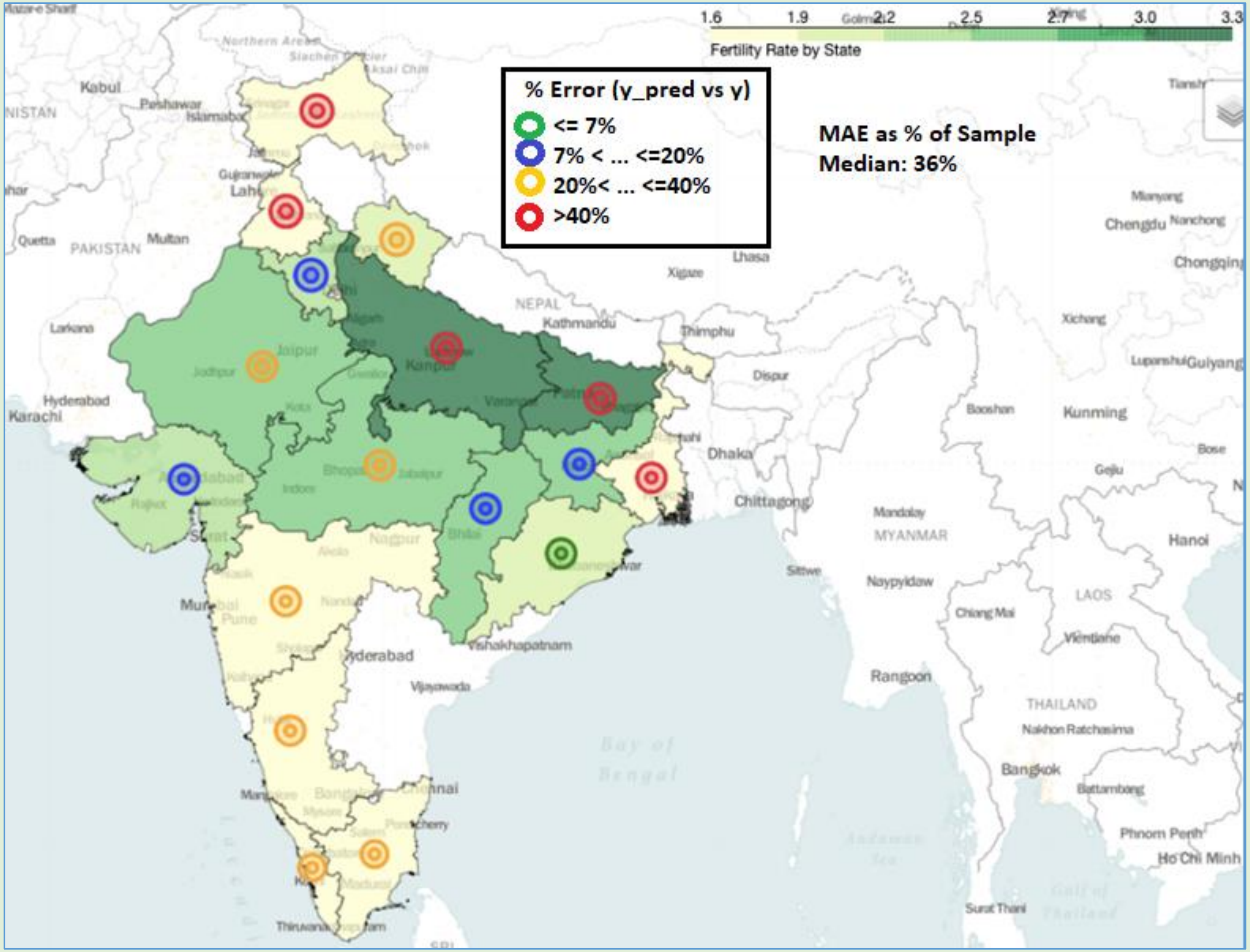
But not unexpected, given the makeshift condition of my India data. Also, very likely that there are influences on TFR which are missing/present in inter-country vs intra-country levels. Hence training a model based on one level (inter-country) is likely to fail to capture the influence of forces specific to the other level (intra-country), resulting in a big divergence in test results

It is obvious that compared to the international regression results, the India results (**predicted by model trained on the former**) have both higher bias and higher variance:

	International	**India**
Test Set Sample Count	35 countries[5]	17 states
Mean Absolute Error as % of Median TFR in Test Set	8.5%	35.7%
Range of % Residuals	(est.) 5 - 10%[6]	2 - 85%

[5] 0.2 x 179 countries

[6] Based on CV on page 15 as proxy for test residuals, which I did not explicitly compute, as I did for India


1.6
1.9
2.2
2.5
2.7
3.0
3.3
Fertility Rate by State
% Error (y_pred vs y)
<= 7%
7% < ... <=20%
20%< ... <=40%
>40%
MAE as % of Sample
Median: 36%
Kabul
Peshawar
Islamabad
Quetta
PAKISTAN
Multan
Larkana
Hyderabad
Karachi
Jaipur
Jodhpur
Kota
Kanpur
Varanasi
Patna
NEPAL
Kathmandu
Thimphu
Lhasa
Dispur
Dhaka
Chittagong
Bhopal
Jabalpur
Indore
Rajkot
Surat
Nagpur
Mumbai
Pune
Hyderabad
Vishakhapatnam
Vijayawada
Chennai
Bengaluru
Mysore
Mangalore
Pondicherry
Madurai
Thiruvananthapuram
Bay of Bengal
MYANMAR
Mandalay
Naypyidaw
Rangoon
THAILAND
Bangkok
Chiang Mai
LAOS
Vientiane
Hanoi
Kunming
Chengdu
Chongqing
Guiyang
Phnom Penh
Ho Chi Minh

5. Thoughts

- ❖ **Data pre-processing is key**. Excellent data (abundant and good quality) more than makes up for any mediocrity in modelling
  - ➔ E.g. with solid data, a linear regression should perform as well as an XGBoost

  On the other hand, if data is scarce/problematic (as most real-world datasets are), then feature engineering/precise modelling must come in

- ❖ Was fortunate in this project, thanks to the World Bank dataset, not to have to do any feature engineering (in addition to feature selection)

- ❖ To take this project further, one might consider:
  - ➔ Running time-series analyses (for a given country or countries) to measure changes in the predictive relationship between features and target over time
  - ➔ Scraping motherhood forums to suss out common pain points to do with having/raising kids (though this analysis would be more developed-world-centric)

6. Sources and Acknowledgments

6.1 Sources

International Regression
https://databank.worldbank.org/data/source/world-development-indicators

India Regression
https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_fertility_rate
http://pbplanning.gov.in/pdf/Statewise%20GSDP%20PCI%20and%20G.R.pdf
https://m.rbi.org.in/Scripts/PublicationsView.aspx?id=18812
https://www.who.int/hrh/resources/16058health_workforce_India.pdf
https://www.statista.com/statistics/620240/old-age-dependency-ratio-by-state-india/
http://niti.gov.in/content/population-number-male-female-rural-urban
indiafacts.in/india-census-2011/literacy-rate-india-2011

6.2 Acknowledgments

Many thanks to:

- ❖ My instructors and classmates at GA DAT-13, for the knowledge and sharing
- ❖ My son, for the inspiration
- ❖ My wife, for your understanding
- ❖ StackOverflow, for your omniscience