Global Suicide Data and Prediction

BY DANIEL ARTZ

Kaggle – Suicide Rates Overview from 1985 to 2015

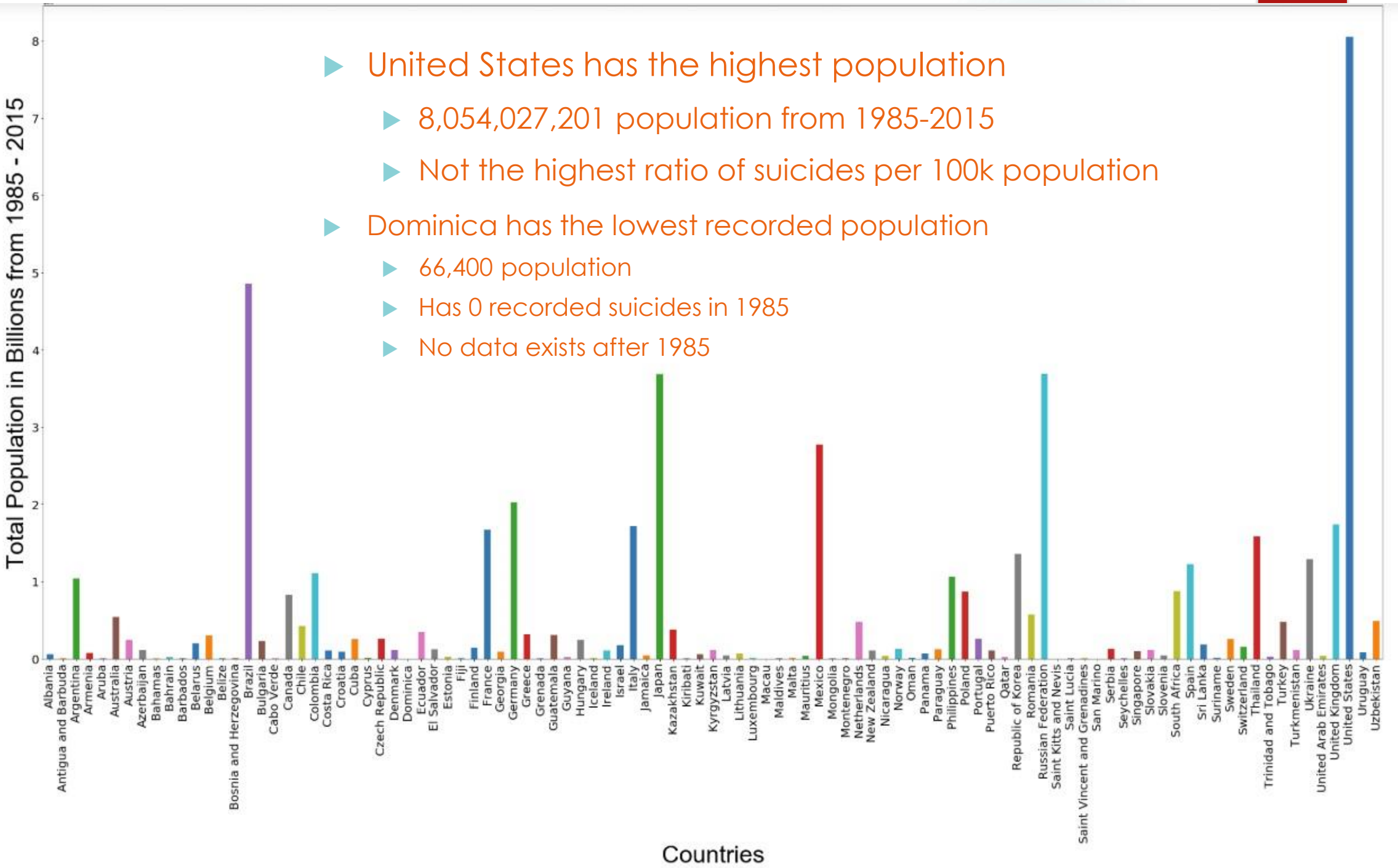
- When looking through datasets this one stood out over the others we reviewed.
  - The data was clean and easy to interpret.
  - Minimal missing values.
  - The features included in the dataset gave a variety of choices for target values.
  - Allowed for use of single label and multi label prediction models.
- Is an interesting topic to research that people usually avoid.

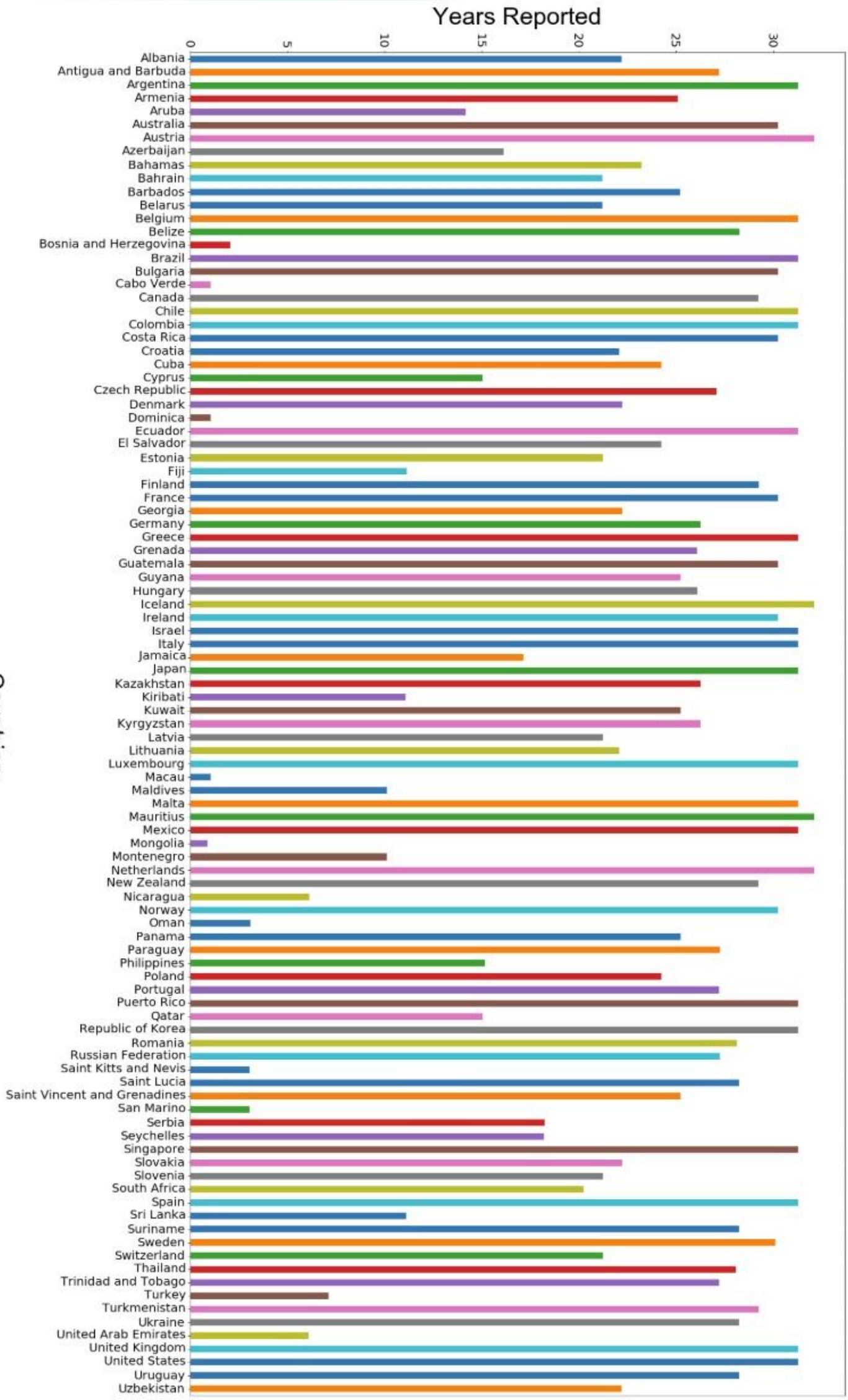
Data Overview

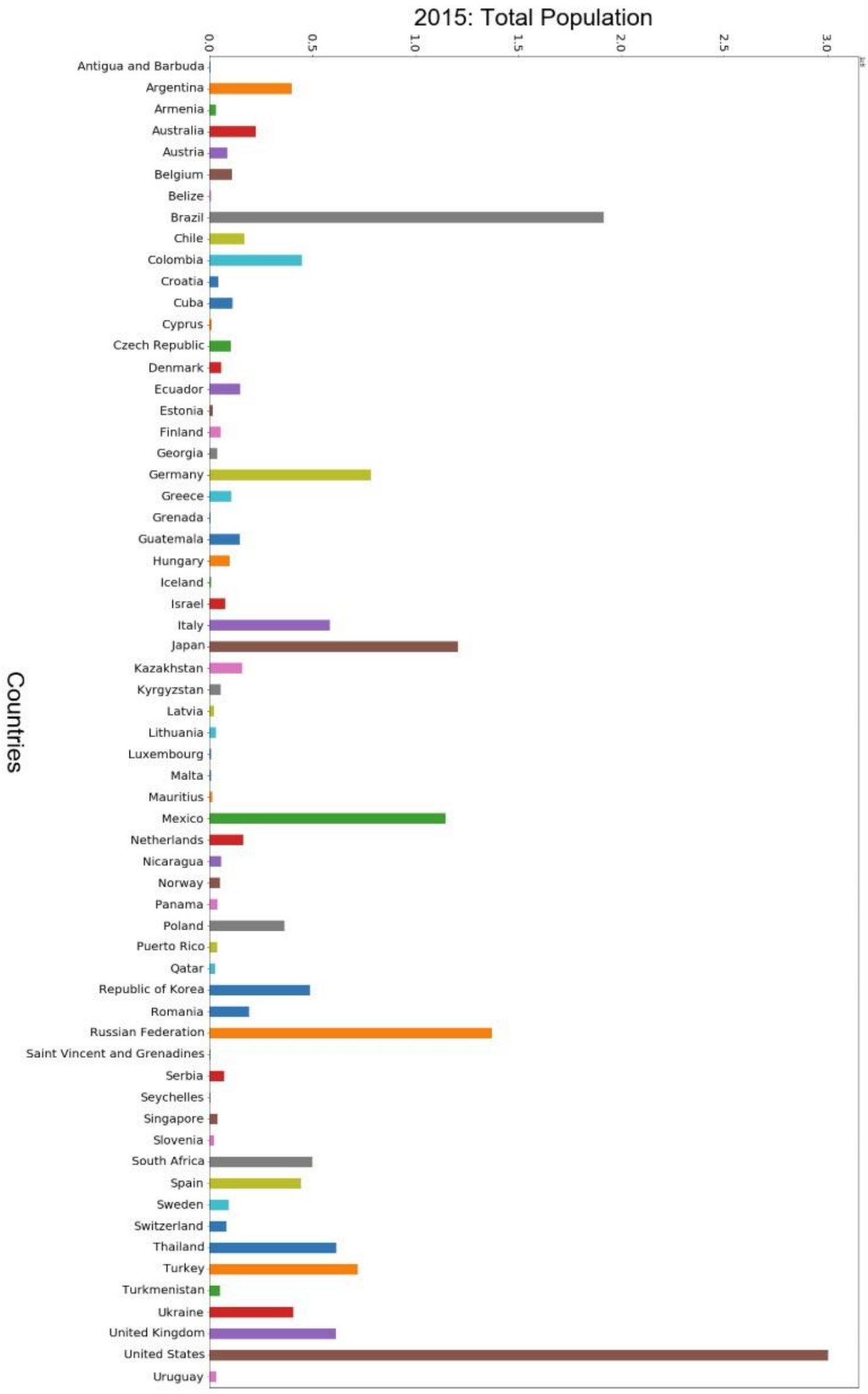
	country	year	sex	age	suicides_no	population	suicides_100k_pop	country_year	gdp_for_year	gdp_per_capita	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	2.156625e+09	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	2.156625e+09	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	2.156625e+09	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	2.156625e+09	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	2.156625e+09	796	Boomers
5	Albania	1987	female	75+ years	1	35600	2.81	Albania1987	2.156625e+09	796	G.I. Generation
6	Albania	1987	female	35-54 years	6	278800	2.15	Albania1987	2.156625e+09	796	Silent
7	Albania	1987	female	25-34 years	4	257200	1.56	Albania1987	2.156625e+09	796	Boomers
8	Albania	1987	male	55-74 years	1	137500	0.73	Albania1987	2.156625e+09	796	G.I. Generation
9	Albania	1987	female	5-14 years	0	311000	0.00	Albania1987	2.156625e+09	796	Generation X

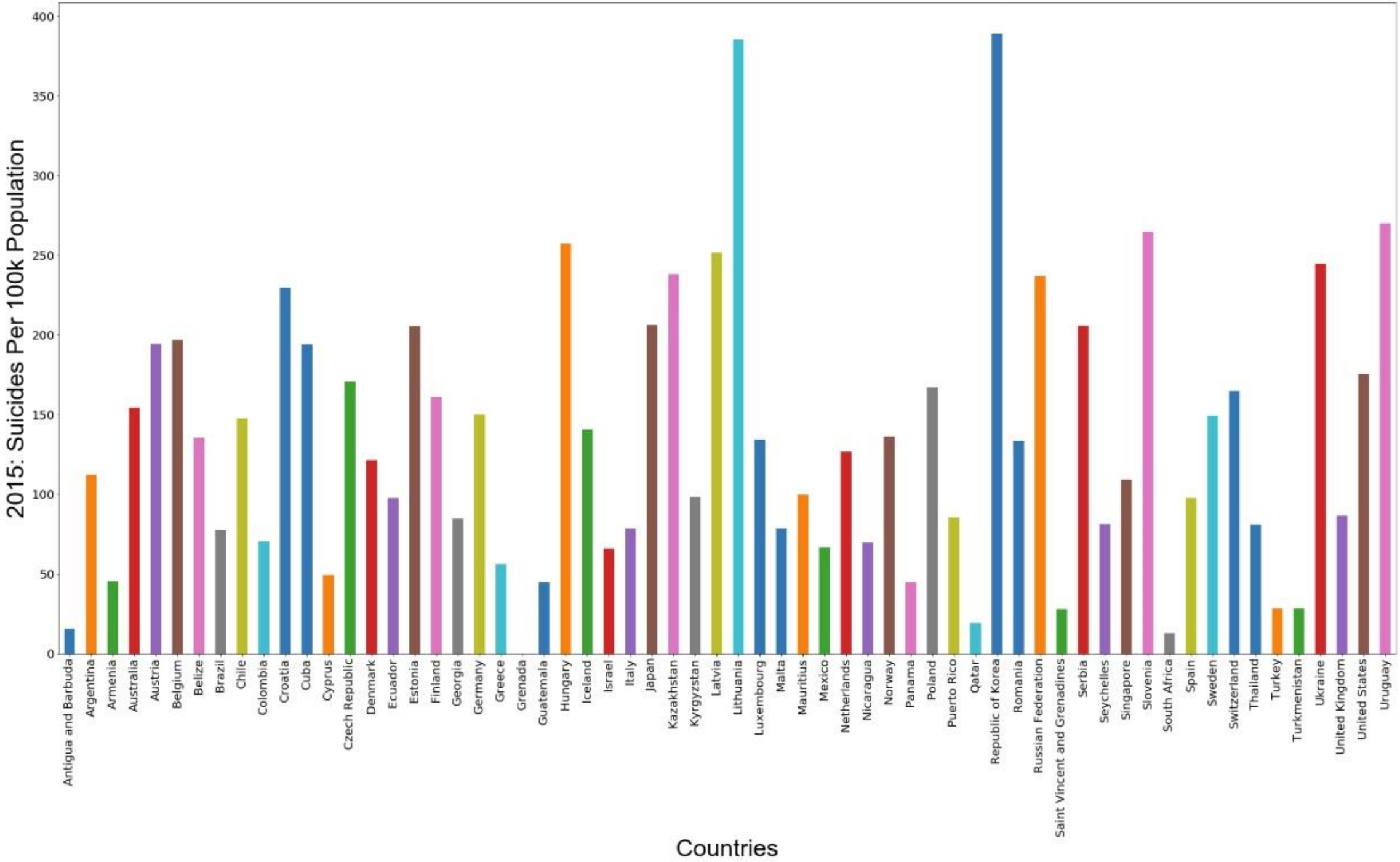
- Removed feature HDI from the original dataset
- Dataset had a large number of strings that needed to be converted

- United States has the highest population
  - 8,054,027,201 population from 1985-2015
  - Not the highest ratio of suicides per 100k population
- Dominica has the lowest recorded population
  - 66,400 population
  - Has 0 recorded suicides in 1985
  - No data exists after 1985




Years Reported
0
5
10
15
20
25
30
Countries
Albania
Antigua and Barbuda
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Barbados
Belarus
Belgium
Belize
Bosnia and Herzegovina
Brazil
Bulgaria
Cabo Verde
Canada
Chile
Colombia
Costa Rica
Croatia
Cuba
Cyprus
Czech Republic
Denmark
Dominica
Ecuador
El Salvador
Estonia
Fiji
Finland
France
Georgia
Germany
Greece
Grenada
Guatemala
Guyana
Hungary
Iceland
Ireland
Israel
Italy
Jamaica
Japan
Kazakhstan
Kiribati
Kuwait
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Macau
Maldives
Malta
Mauritius
Mexico
Mongolia
Montenegro
Netherlands
New Zealand
Nicaragua
Norway
Oman
Panama
Paraguay
Philippines
Poland
Portugal
Puerto Rico
Qatar
Republic of Korea
Romania
Russian Federation
Saint Kitts and Nevis
Saint Lucia
Saint Vincent and Grenadines
San Marino
Serbia
Seychelles
Singapore
Slovakia
Slovenia
South Africa
Spain
Sri Lanka
Suriname
Sweden
Switzerland
Thailand
Trinidad and Tobago
Turkey
Turkmenistan
Ukraine
United Arab Emirates
United Kingdom
United States
Uruguay
Uzbekistan


2015: Total Population
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Countries
Antigua and Barbuda
Argentina
Armenia
Australia
Austria
Belgium
Belize
Brazil
Chile
Colombia
Croatia
Cuba
Cyprus
Czech Republic
Denmark
Ecuador
Estonia
Finland
Georgia
Germany
Greece
Grenada
Guatemala
Hungary
Iceland
Israel
Italy
Japan
Kazakhstan
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Malta
Mauritius
Mexico
Netherlands
Nicaragua
Norway
Panama
Poland
Puerto Rico
Qatar
Republic of Korea
Romania
Russian Federation
Saint Vincent and Grenadines
Serbia
Seychelles
Singapore
Slovenia
South Africa
Spain
Sweden
Switzerland
Thailand
Turkey
Turkmenistan
Ukraine
United Kingdom
United States
Uruguay


2015: Suicides Per 100k Population
0
50
100
150
200
250
300
350
400
Antigua and Barbuda
Argentina
Armenia
Australia
Austria
Belgium
Belize
Brazil
Chile
Colombia
Croatia
Cuba
Cyprus
Czech Republic
Denmark
Ecuador
Estonia
Finland
Georgia
Germany
Greece
Grenada
Guatemala
Hungary
Iceland
Israel
Italy
Japan
Kazakhstan
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Malta
Mauritius
Mexico
Netherlands
Nicaragua
Norway
Panama
Poland
Puerto Rico
Qatar
Republic of Korea
Romania
Russian Federation
Saint Vincent and Grenadines
Serbia
Seychelles
Singapore
Slovenia
South Africa
Spain
Sweden
Switzerland
Thailand
Turkey
Turkmenistan
Ukraine
United Kingdom
United States
Uruguay
Countries

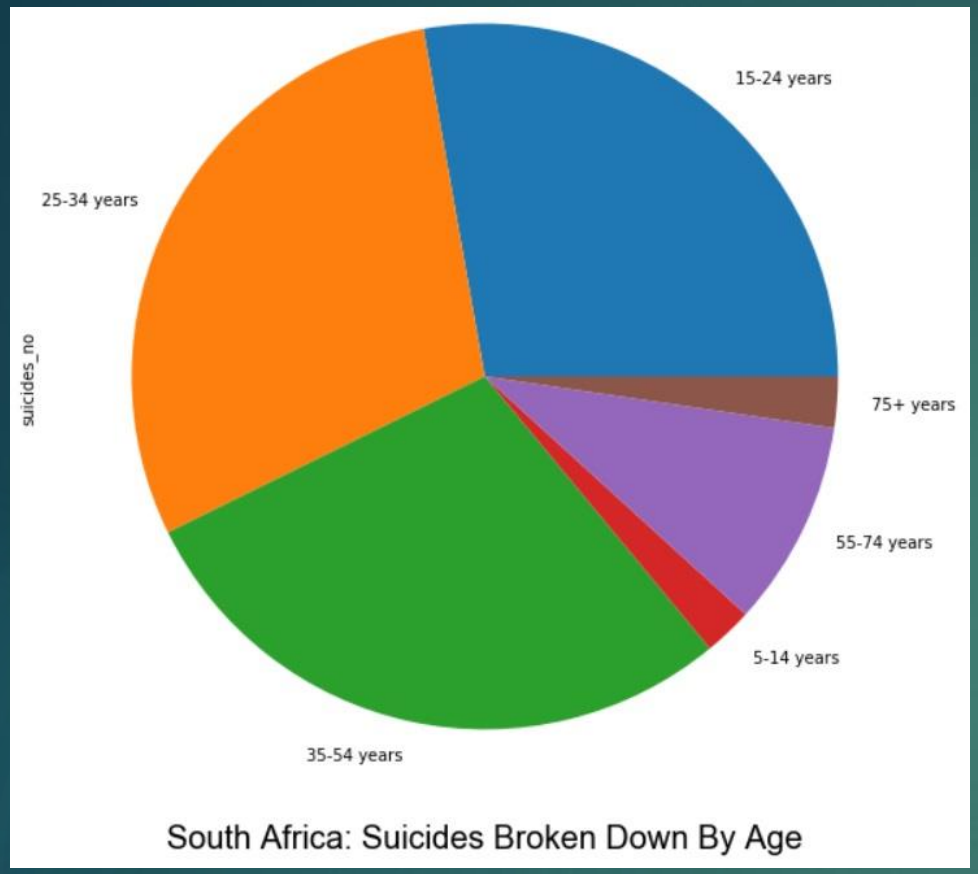
Outliers Chosen

- Chose outliers with the least missing data
- Both outliers had similar total populations in the year 2015
- Republic of Korea (South Korea) – Highest total Suicides/100k Population
  - Total of Suicides per 100k population from 1985-2015 = 9350.45
- South Africa – Lowest (Suicides/100k Population total) > 0 (with most data)
  - Couldn't display statistics on a country with 0 suicides
  - Total of Suicides per 100k population from 1996-2015 = 231.49

Issues With Data Visualization

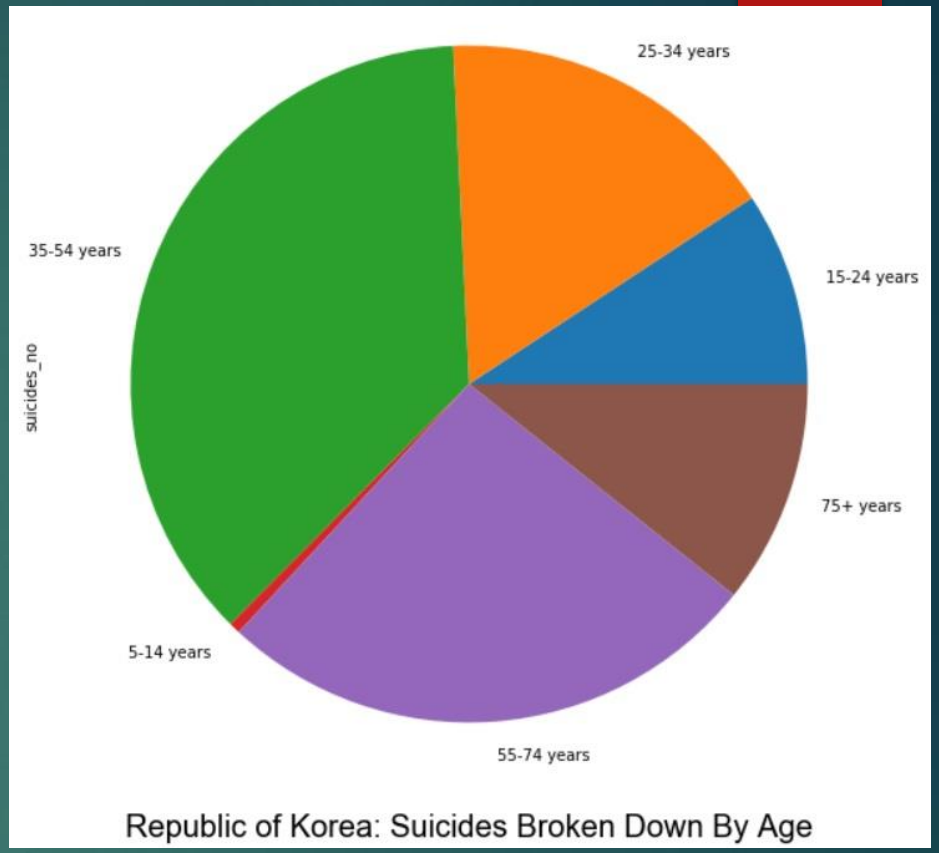
- Using Pandas for the first time took a little time to get used to.
- Determining which features to focus on visualizing took time.
  - Switched between mean suicides/100k population per year to total suicides/100k population
  - The dataset separates the suicides of people of different age, sex and country, making some selections more difficult than others.
- Determining how to choose outliers also took time.
- Some countries were missing numerous rows of data. Most countries had data from 1987 – 2015
  - Made visualization more difficult

Age Groups: Total Suicides from 1985 - 2015



South Africa: Suicides Broken Down By Age

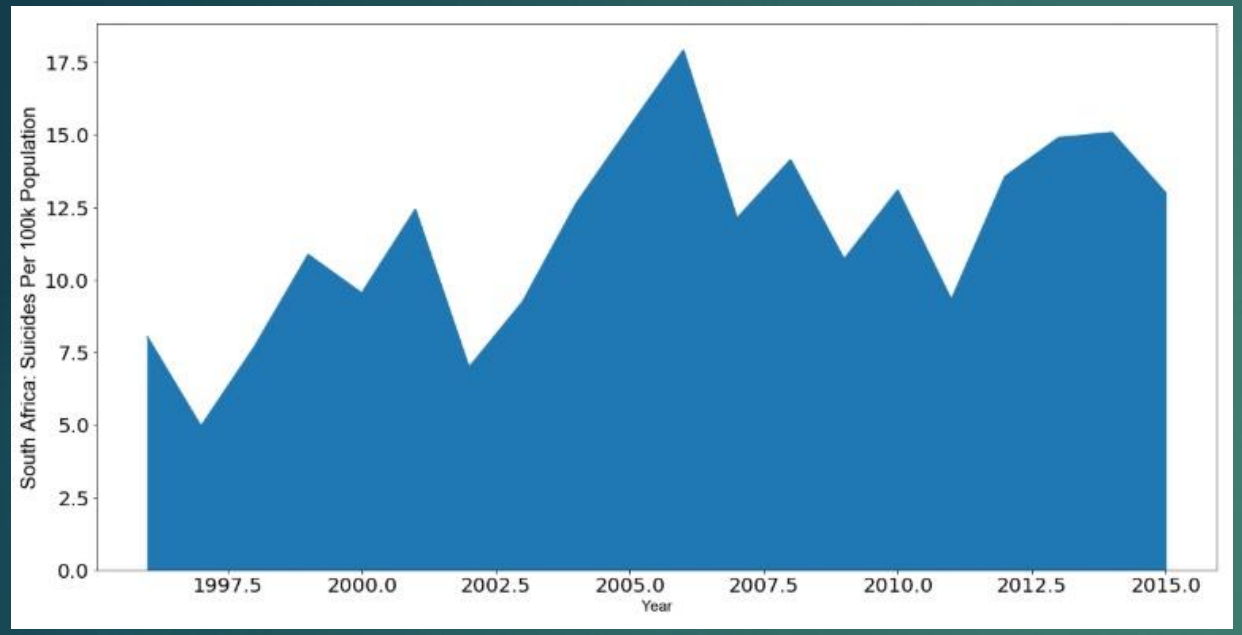
Age	#	%
5-14:	163	2.2%
15-24:	2034	27.8%
25-34:	2160	29.5%
35-54:	2102	28.7%
55-74:	691	9.4%
75+:	171	2.3%

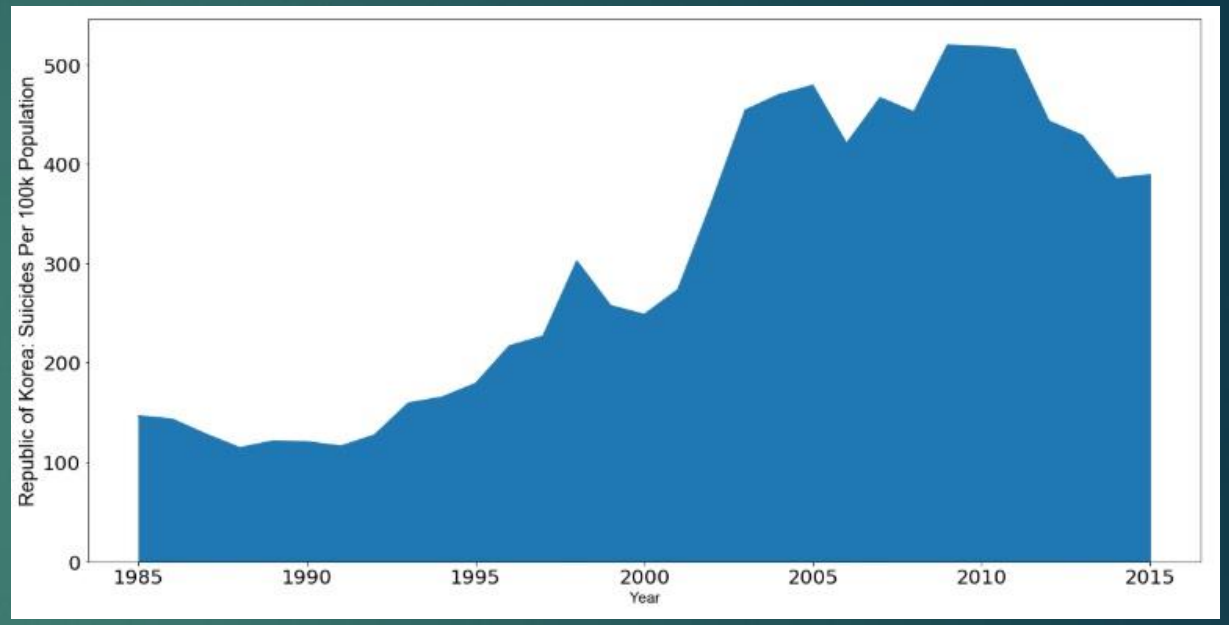


Republic of Korea: Suicides Broken Down By Age

Age	#	%
5-14:	1428	.5%
15-24:	24243	9.3%
25-34:	43167	16.5%
35-54:	96292	36.8%
55-74:	68574	26.2%
75+:	28026	10.7%

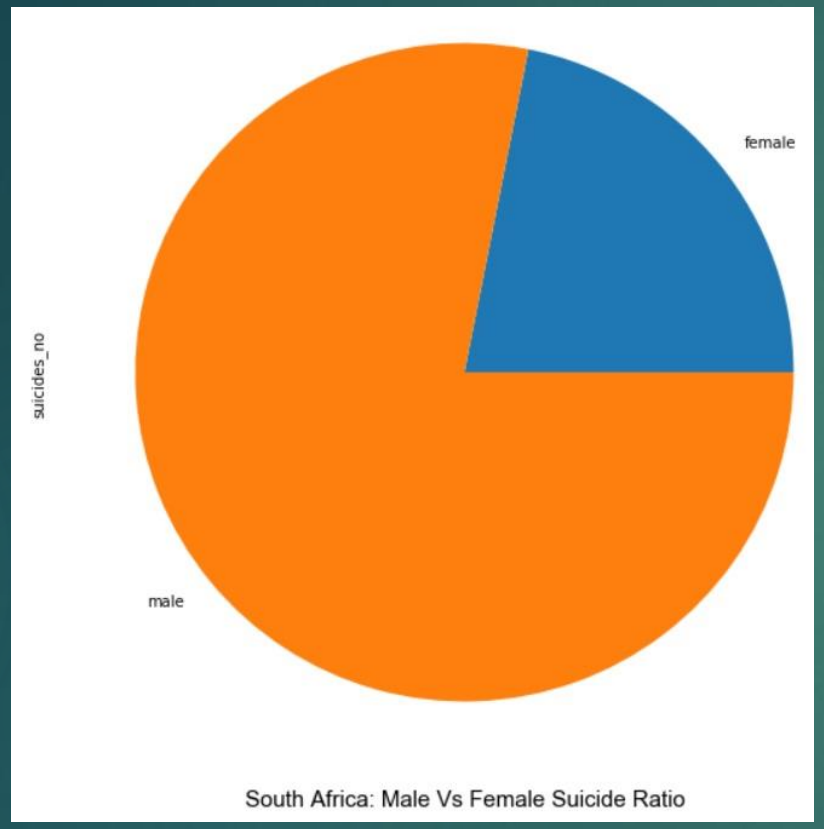
Suicides per 100k population



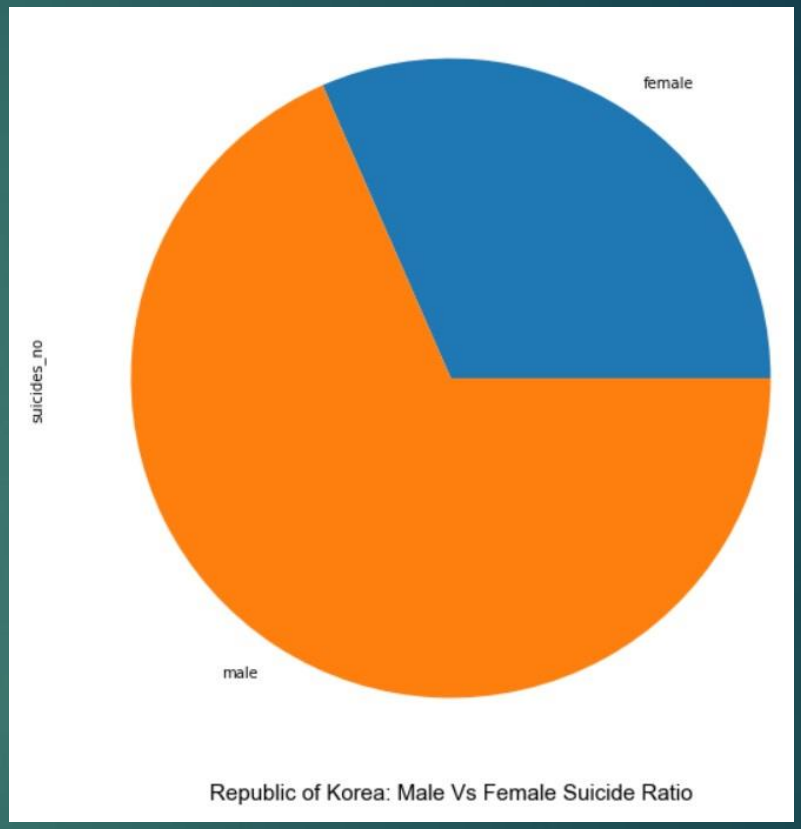


- Steady increase in total suicides per year in South Korea
  - Similar population to South Africa
  - Suicides decrease from 2010-2014
  - Suicide numbers begin to trend upwards by 2015
- South Africa displays mild fluctuations in suicides per 100k population from 1995 - 2015

Total suicides based on sex



Sex	#	%
Male:	5719	78.1%
Female:	1602	21.9%



Sex	#	%
Male:	179115	68.4%
Female:	82615	31.6%

Feature Engineering

- Used a correlation table to remove features with a influence on the target.
  - Substantially increased accuracy across all tested models when predicting the target value "sex".
- Removed the feature "HDI" which was missing over 90% of its values.

	country	year	sex	age
country	1.000000e+00	0.022769	5.217358e-20	1.269341e-04
year	2.276923e-02	1.000000	0.000000e+00	-2.932878e-03
sex	5.217358e-20	0.000000	1.000000e+00	-1.683559e-19
age	1.269341e-04	-0.002933	-1.683559e-19	1.000000e+00
suicides_no	1.185555e-01	-0.004546	-1.446292e-01	-6.599386e-02
population	2.276923e-02	1.000000	0.000000e+00	-2.932878e-03
suicides_100k_pop	5.522414e-02	-0.039037	-3.914965e-01	-1.272520e-01
country_year	9.994466e-01	0.033676	0.000000e+00	8.086930e-05
gdp_for_year	1.635989e-01	0.094511	-5.824804e-19	3.045725e-04
gdp_per_capita	5.334082e-02	0.339134	-1.336263e-18	-7.849172e-04
generation	2.609368e-03	0.236322	0.000000e+00	1.823793e-01

- Dropped features "sex":
  - Population
  - Generation
  - Country_year
- Attempted to remove the feature "year". However, it decreased the accuracy on all tested models.

Types of Classifiers

- Logistic Regression
- Supervised Neural Network
- Decision Tree

Target Variables

- Sex
  - Male
  - Female
- Age
  - 5-14
  - 15-24
  - 25-34
  - 35-54
  - 55-74
  - 75+

Logistic Regression

- Attempted numerous models by performing cross validation on different models in a parameter grid.
  - penalty: l2, l1
  - solver: liblinear, saga
  - C: 0.01, .1, 1, 2, 5, 7, 9, 10, 20, 30, 40, 50, 60, 70, 100, 500, 1000
  - Predicting sex
    - Model with highest accuracy: penalty = l2, solver = liblinear, C = .01, CV accuracy = 70.8%, accuracy = 69.9%
  - Predicting age (used one vs rest multi_class parameter)
    - Model with highest accuracy: penalty = l2, solver = liblinear, C = 1000, multi class = ovr, CV accuracy = 39.5%, accuracy = 40.5%

Supervised Learning Neural Network

- Attempted numerous models by performing cross validation on different models in a parameter grid.
  - Hidden layer sizes: (5,), (10,), (100,), (5, 5), (10, 10), (50, 50)
  - activation: relu, logistic
  - alpha: .0001, .01, .05, .10, .2, .6
  - Learning rate: adaptive
  - Solver: lbfgs, adam
  - Predicting sex
    - Model with highest accuracy: hidden layer size = (100,), activation = relu, solver = adam, random state = 1, alpha = .1, CV accuracy = 72.6%, accuracy = 75.9%
  - Predicting age
    - CV showed alpha .2 yielded the highest accuracy. However, I manually tested smaller alphas on the full dataset and an alpha of .002 increased the accuracy from .753 to .771
    - Model with highest accuracy: hidden layer size = (10, 10), activation = relu, solver = adam, random state = 1, alpha = .002, CV accuracy = 51.1%, accuracy = 77.1%

Decision Tree

- Model with the highest accuracy for both prediction of sex and age
- Predictions were more time efficient than Neural Network models
- Attempted numerous models by performing cross validation on different models in a parameter grid.
  - Max depth: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
  - Min samples split: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
  - min_samples_leaf: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
  - Predicting sex
    - Model with highest accuracy: max depth = 15, min samples split = 2, min samples leaves = 2, CV accuracy = 74.5%, accuracy = 81.1%
  - Predicting age (used one vs rest multi_class parameter)
    - Model with highest accuracy: max depth = 15, min samples split = 2, min samples leaf = 2, CV accuracy = 77.8%, accuracy = 86.4%

Decision Tree Prediction Metrics

Sex Prediction Metrics

	precision	recall	f1-score	support
Male	0.82	0.79	0.81	2775
Female	0.80	0.83	0.81	2789
micro avg	0.81	0.81	0.81	5564
macro avg	0.81	0.81	0.81	5564
weighted avg	0.81	0.81	0.81	5564

Consistent f1-score when predicting both male and female

Age Prediction Metrics

	precision	recall	f1-score	support
15-24	0.77	0.80	0.78	930
35-53	0.95	0.93	0.94	916
75+	0.86	0.88	0.87	933
25-34	0.79	0.78	0.79	947
55-74	0.87	0.85	0.86	888
5-14	0.94	0.94	0.94	950
micro avg	0.86	0.86	0.86	5564
macro avg	0.86	0.86	0.86	5564
weighted avg	0.86	0.86	0.86	5564

Much higher f1-score when predicting ages 5-14 and 35-53

Decision Tree Age Prediction Confusion Matrix

		Predicted					
		0	1	2	3	4	5
Actual	0	741	0	0	129	0	60
	1	0	851	0	65	0	0
	2	0	0	824	0	109	0
	3	167	42	0	738	0	0
	4	0	0	129	0	759	0
	5	58	0	0	0	0	892

	precision	recall	f1-score	support
15-24	0.77	0.80	0.78	930
35-53	0.95	0.93	0.94	916
75+	0.86	0.88	0.87	933
25-34	0.79	0.78	0.79	947
55-74	0.87	0.85	0.86	888
5-14	0.94	0.94	0.94	950
micro avg	0.86	0.86	0.86	5564
macro avg	0.86	0.86	0.86	5564
weighted avg	0.86	0.86	0.86	5564

- When looking at the two age ranges with the lowest precision
  - Large occurrence of 25-34 years old being predicted as 15-24 years old
  - Large occurrence of 15-24 years old being predicted as 25-34 years old

Application Area

- Suicide Prevention by flagging people as at risk
- Would be a useful tool for psychologists and/or physicians
- People flagged as at risk patients could be given preventative care

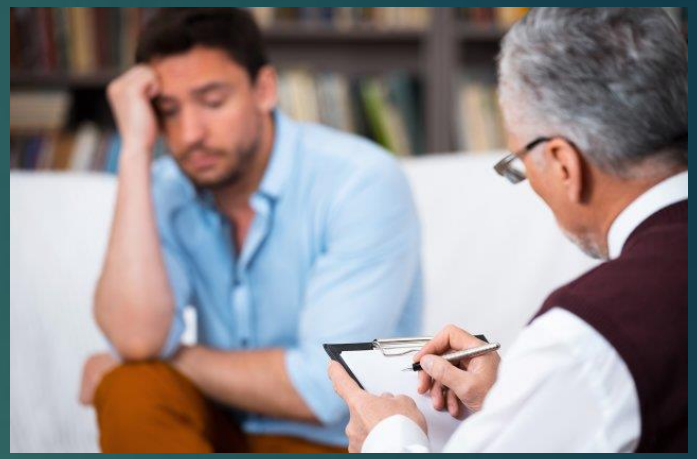


Photo Source: https://money.usnews.com/careers/best-jobs/psychologist

Future Work

- Data Gathering from two main groups:
  - Would need to gather data from medical surveys or by other means
    - Physician office visits or therapist visits
  - People who committed suicide
  - People who died from other causes
- Data would be representative of individuals rather than groups
  - Features would need to be completely different
    - Would need more features than the current data set
    - Would require further research into more appropriate models
      - Decision trees would be inefficient for larger and more complex datasets

Dataset URL

https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016