Analyzing and Predicting Suicide Rates in Europe

By: Shawn McClain

The Problem

- Nearly 800,000 people die by suicide in the world each year, which is roughly one death every 40 seconds.
- On average, there are 129 suicides per day in the U.S, which translates to about 1 suicide every 11 minutes.
- Suicide is the 10th leading cause of death in the U.S

Are other modernized countries seeing spikes in suicides?
What factors are correlated with suicide rates?
Can we build a model to address the issue?

Source: https://save.org/about-suicide/suicide-facts/

Clients: Who Might Care?

Governments

Non-Profits

SAVE
SUICIDE AWARENESS
VOICES OF EDUCATION

Healthcare

Potential Factors/ Preliminary Hypothesis

Latitude: Countries in Northern Latitudes are more likely to suffer from Seasonal Affective Disorder which will increase suicide rates in those countries.

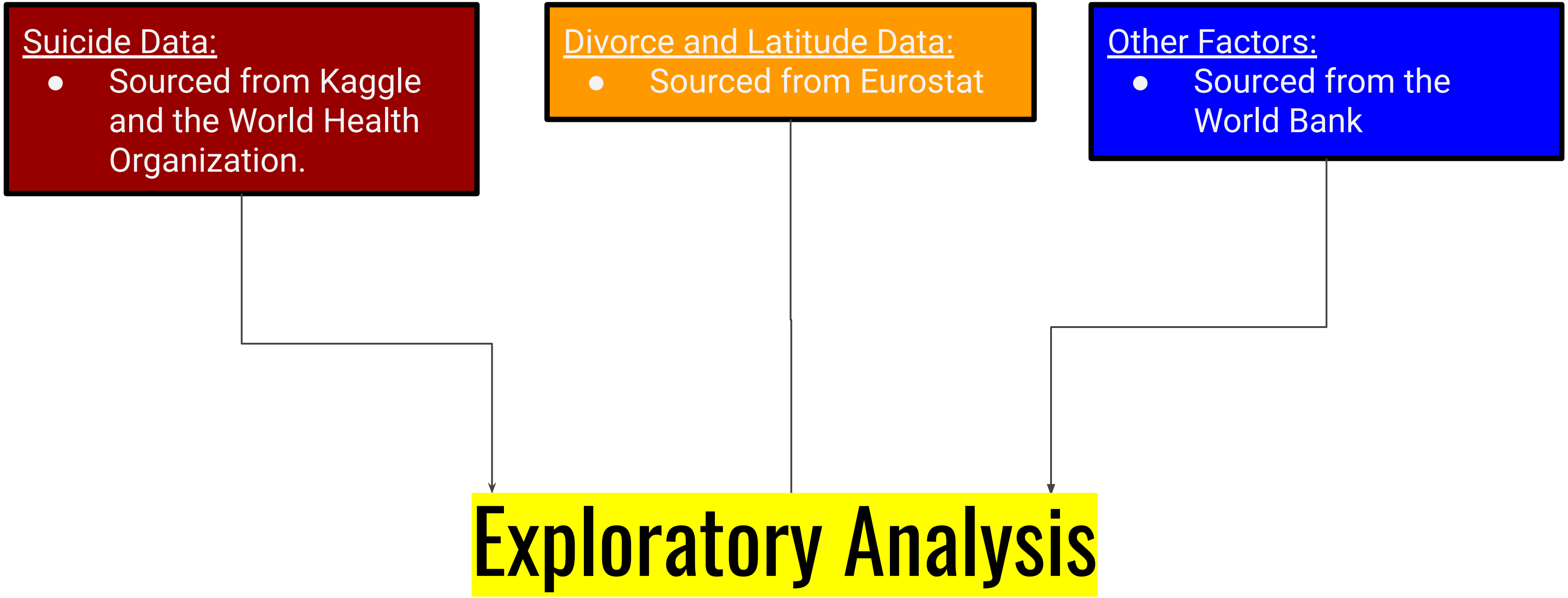
Cell Phone Subscriptions: Countries with higher cell phone subscriptions will suffer from higher rates. Cell phones may make people more anxious.

Year: Suicides will increase over time.

Divorce Rates: Countries with higher divorce rates will have higher rates of suicides. Divorce is more likely to make people depressed.

Electric Power Consumption: Similarly to cell phones, more electricity consumption will yield countries with higher rates of suicide. Higher electric consumption may make people more anxious.

Data Extraction



Data Cleaning and Wrangling

A Brief Overview of Important Steps

Column and Row Manipulation w/ Stacking

In [12]:
```
countryDf = countryDf.reset_index()
```

In [13]:
```
countryDf= countryDf.set_index(['Country Name', 'Series_name'])
```

In [14]:
```
countryDf.columns.name= 'Year'
#Gave seperate columns year a group name
```

In [15]:
```
countryDf= countryDf.stack()
countryDf= countryDf.unstack('Series_name')
#moving series factors to unique columns instead of rows
```

In [16]:
```
countryDf.head()
```

Out[16]:

	Series_name	Access to electricity (% of population)	Birth rate, crude (per 1,000 people)	CO2 emissions (metric tons per capita)	Current health expenditure per capita (current US$)	Death rate, crude (per 1,000 people)	con
Country Name	Year						
Afghanistan	2000	NaN	48.021	0.037234781	NaN	11.718	
	2001	NaN	47.505	0.037846136	NaN	11.387	
	2002	NaN	46.901	0.047377324	16.24954214	11.048	
	2003	NaN	46.231	0.050481336	17.49073742	10.704	
	2004	NaN	45.507	0.038410043	20.92708722	10.356	

My columns were originally stacked into one column and I needed years stacked into one columns.

Lambda Functions to Replace and Filter Data

```
In [10]: countryDf= countryDf.apply(lambda x: x.replace('..',np.nan))
         #replace '..' with a nan value
```

Null values in this dataframe are marked as a string '..', I used a lambda function to replace them with numpy NaN so we could work with missing data easier.

```
In [20]: europels=['Albania', 'Andorra', 'Armenia', 'Austria', 'Azerbaijan', 'Belarus',
         #list of european countries
```

```
In [21]: europeDf = countryDf[countryDf['Country Name'].map(lambda x: x in europels)]
         europeDf.reset_index(drop=True, inplace=True)
         #creating new euro df from euro list
```

```
In [22]: europeDf.head()
```

Out[22]:

Series_name	Country Name	Year	Access to electricity (% of population)	Birth rate, crude (per 1,000 people)	CO2 emissions (metric tons per capita)	Current health expenditure per capita (current US$)	Death rate, crude (per 1,000 people)	Electric power consumption (kWh per capita)
0	Albania	2000	100.0	16.436	0.978	75.531	5.914	1449.647
1	Albania	2001	100.0	15.590	1.053	81.946	5.879	1351.231
2	Albania	2002	100.0	14.790	1.230	89.858	5.891	1578.166
3	Albania	2003	100.0	14.048	1.413	113.584	5.952	1469.265
4	Albania	2004	100.0	13.381	1.376	151.981	6.061	1797.525

I want to filter for only European countries, so I create a list and use a lambda to map the full dataframe by my Europe Countries and save that to a new dataframe.

Merging Dataframes and Column Creation

In [21]:
```
EUdivorce['mean']= EUdivorce.iloc[:][EUdivorce.columns].mean(axis=1)
#created avg of all years for each country
```

In [22]:
```
EUdivorce
```

Out[22]:

country_name	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	mean
Belgium	2.8	2.8	3.3	3.0	2.7	2.5	2.3	2.2	2.2	2.2	2.1	2.0	2.508
Bulgaria	2.0	2.2	1.9	1.6	1.5	1.4	1.6	1.5	1.5	1.5	1.5	1.5	1.642

A new column **'mean'** is created by taking the mean of all indices and all of the columns (years) for each index.

```
eurolats.head()
#adding latitude
```

Out[23]:

	latitude	name
0	42.546	Andorra
1	23.424	United Arab Emirates
2	33.939	Afghanistan
3	17.061	Antigua and Barbuda
4	18.221	Anguilla

In [24]:
```
eurolats = eurolats[eurolats['name'].isin(europels)]
```

In [25]:
```
eurolats.rename(columns={'name':'country_name'},inplace=True)
```

In [26]:
```
eurolats.set_index('country_name',inplace=True)
```

In [27]:
```
EUdivorce= pd.merge(EUdivorce,eurolats, how='left',on='country_name')
EUdivorce.head()
```

Out[27]:

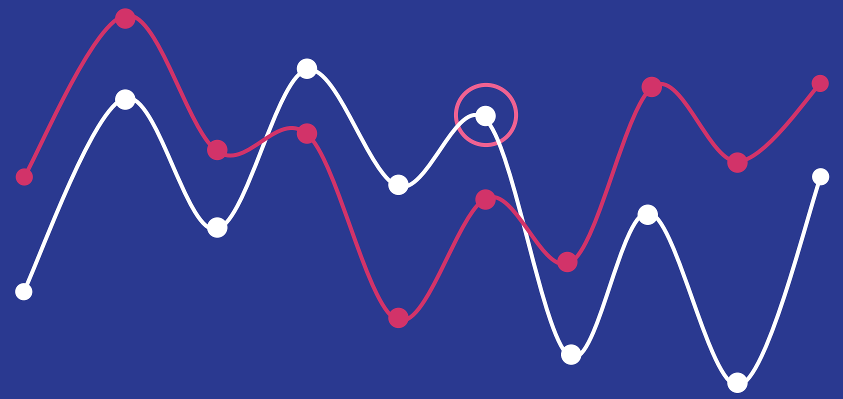
country_name	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	mean	latitude
Belgium	2.8	2.8	3.3	3.0	2.7	2.5	2.3	2.2	2.2	2.2	2.1	2.0	2.508	50.504
Bulgaria	2.0	2.2	1.9	1.6	1.5	1.4	1.6	1.5	1.5	1.5	1.5	1.5	1.642	42.734
Czech Republic	3.1	3.0	3.0	2.8	2.9	2.7	2.5	2.7	2.5	2.5	2.4	2.4	2.708	49.817

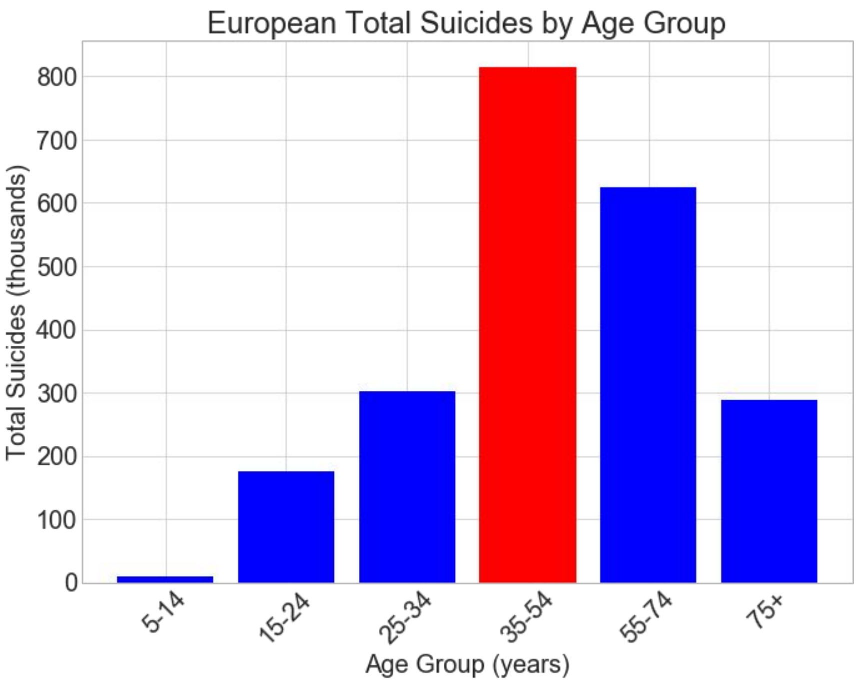
A new dataframe is added for the latitudes for each country.

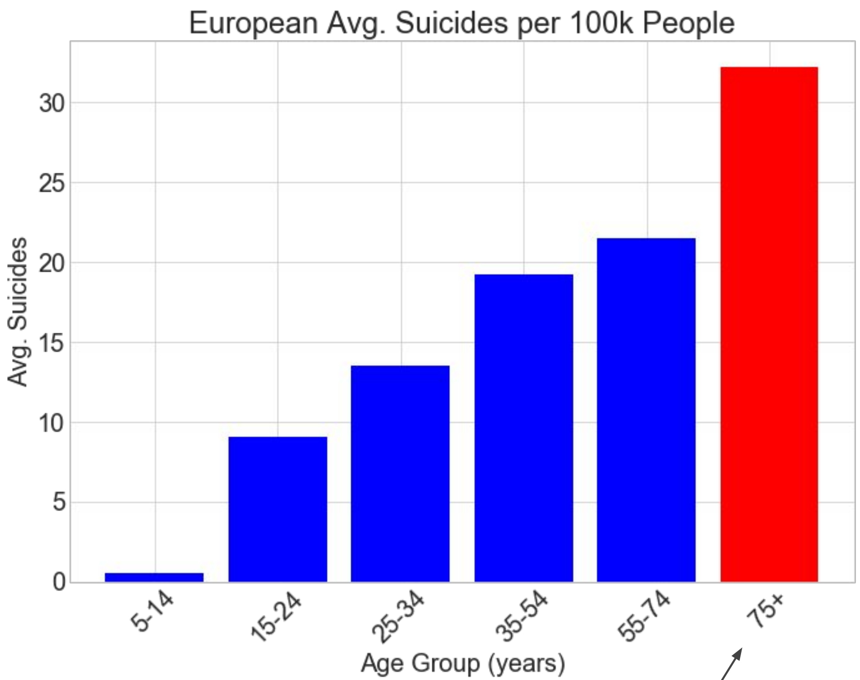
My Europe list is applied to the dataframe.

The two data frames are **left-joined together** on the key **'country_name'**.

Exploratory Analysis

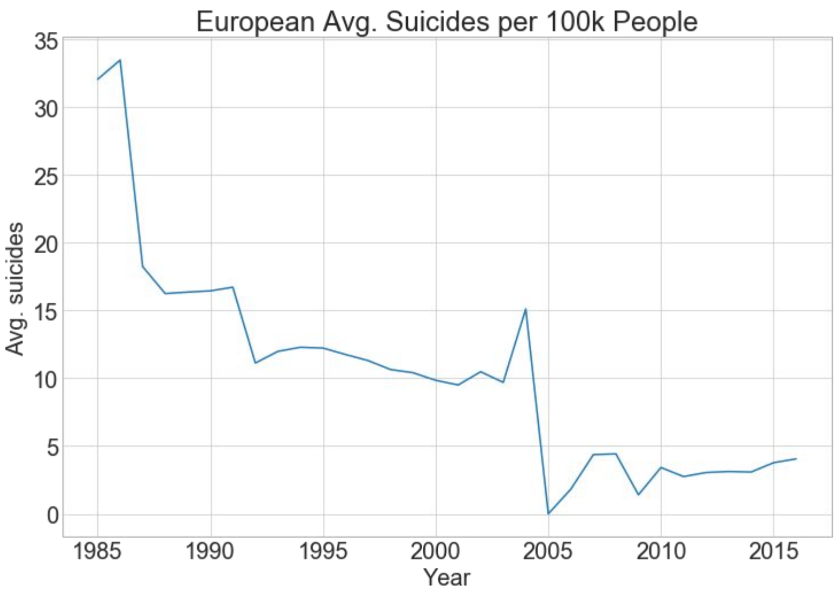


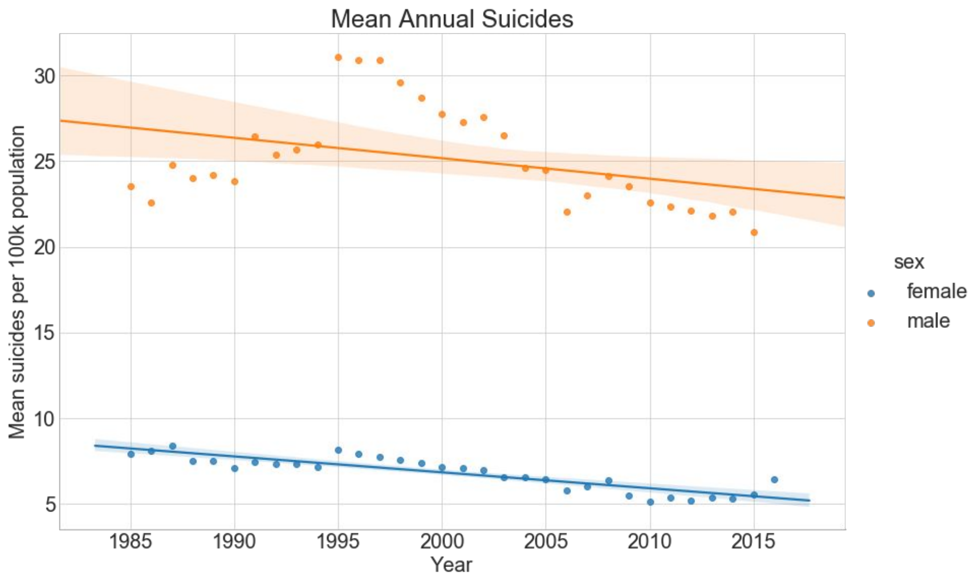




The age group of 35-54 has the highest total suicides (but also the largest population).

The age group of 75+ has the highest per capita rate of suicide.

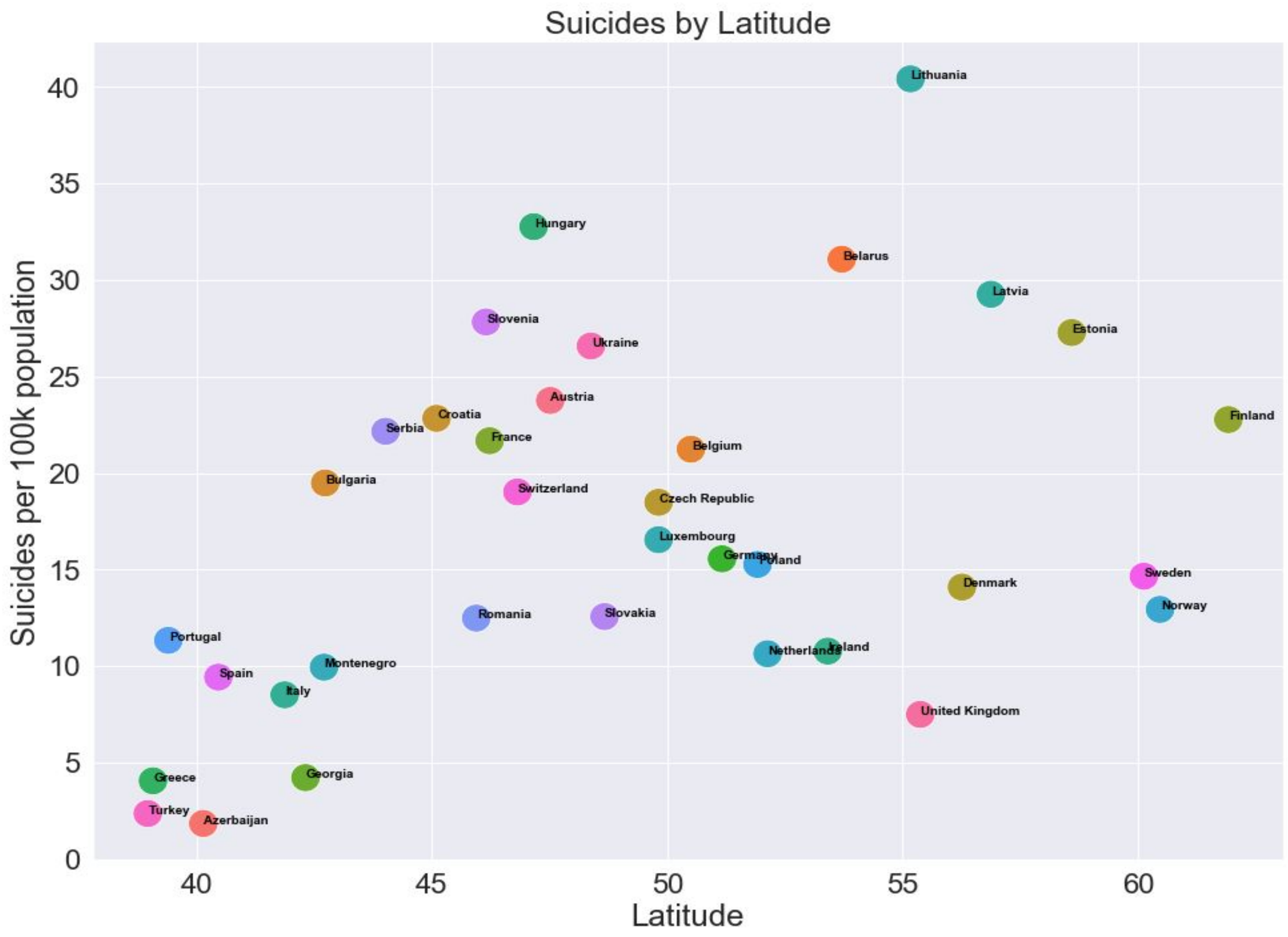




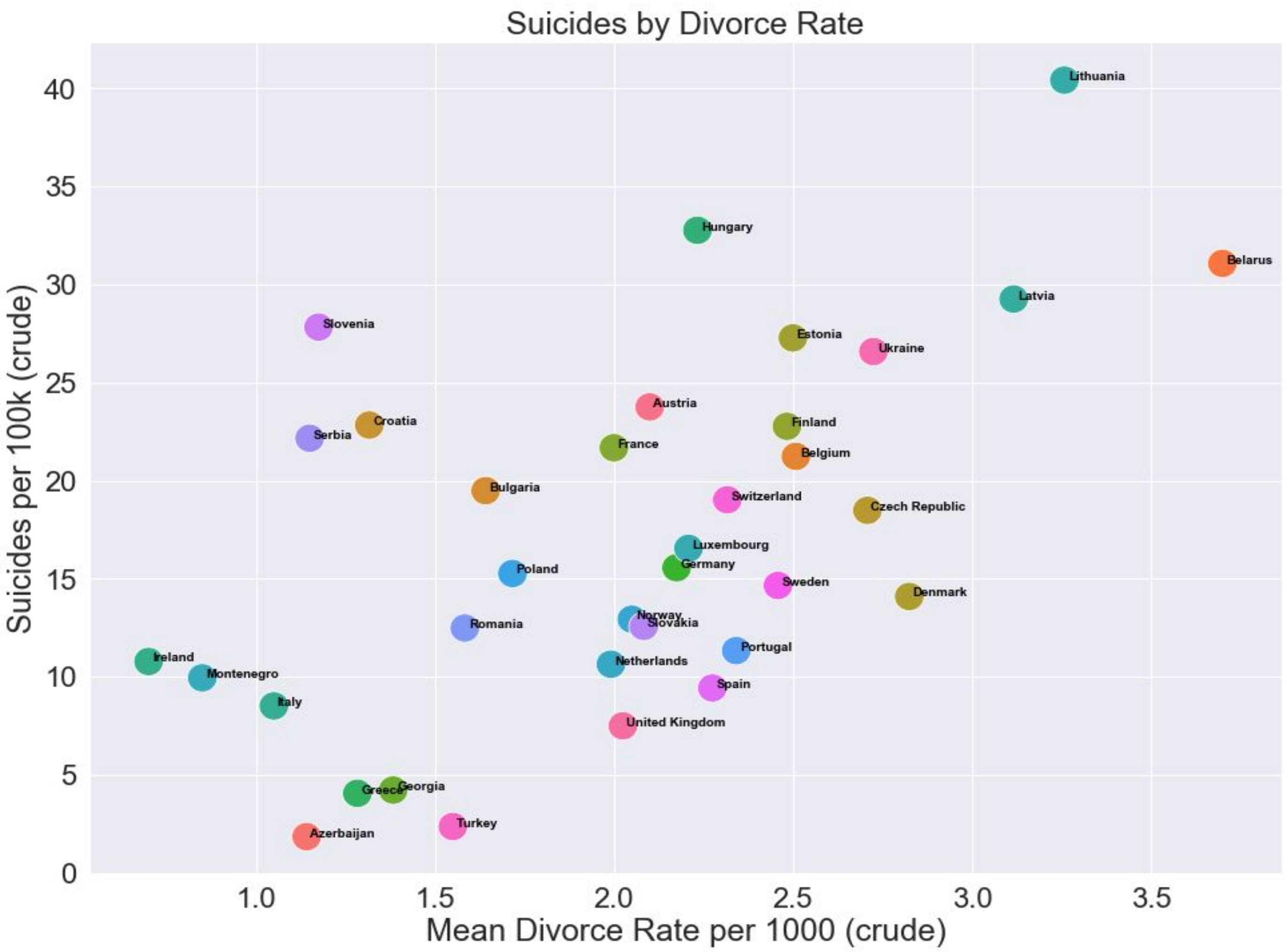
Average suicides are generally trending lower in Europe (as of 2015)

Suicides for men and women are trending lower in Europe, but men are over 3x likely to die from suicide than women.

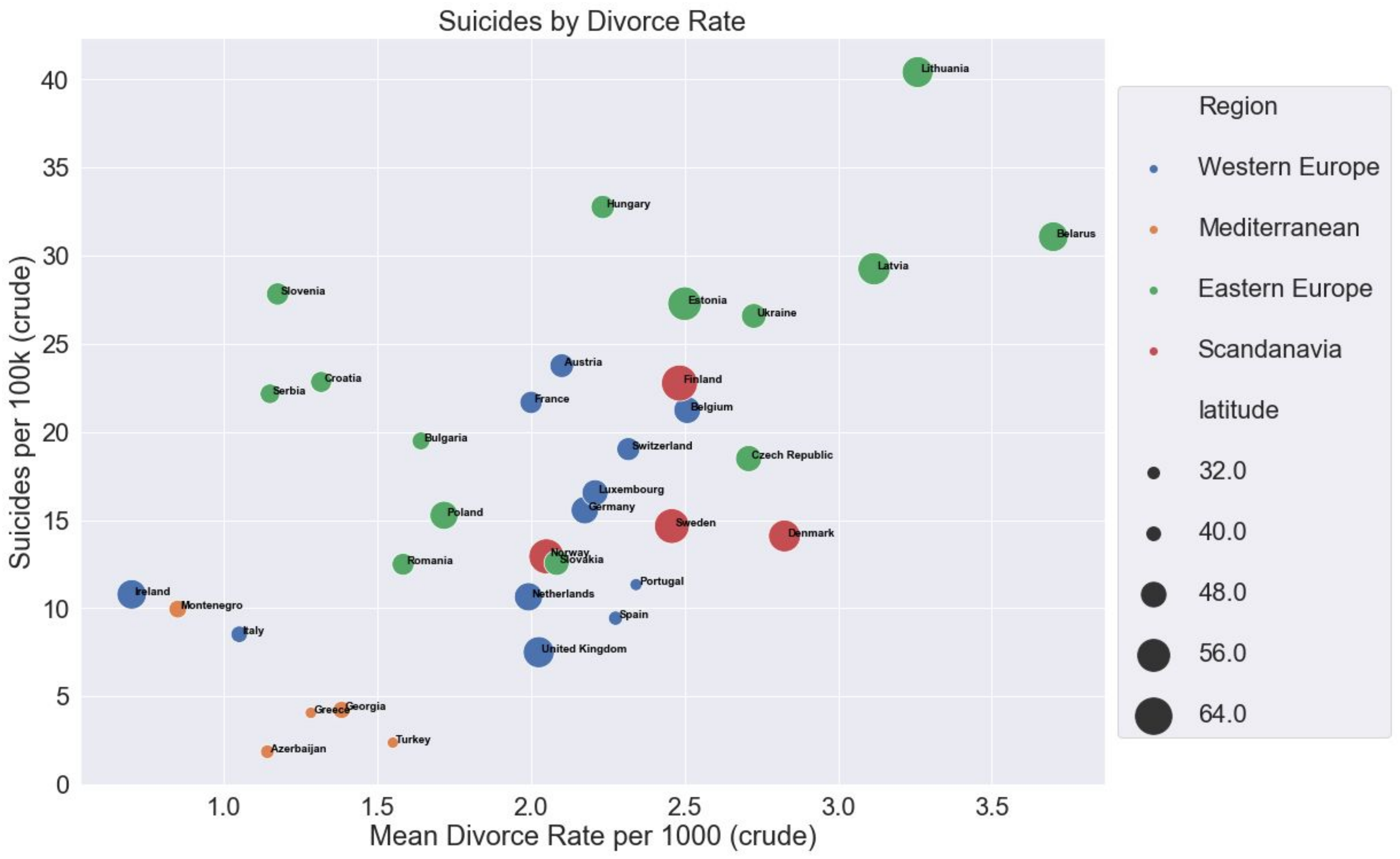
	suicides_no	fraction_of_total
sex		
female	527427	0.238
male	1687876	0.762

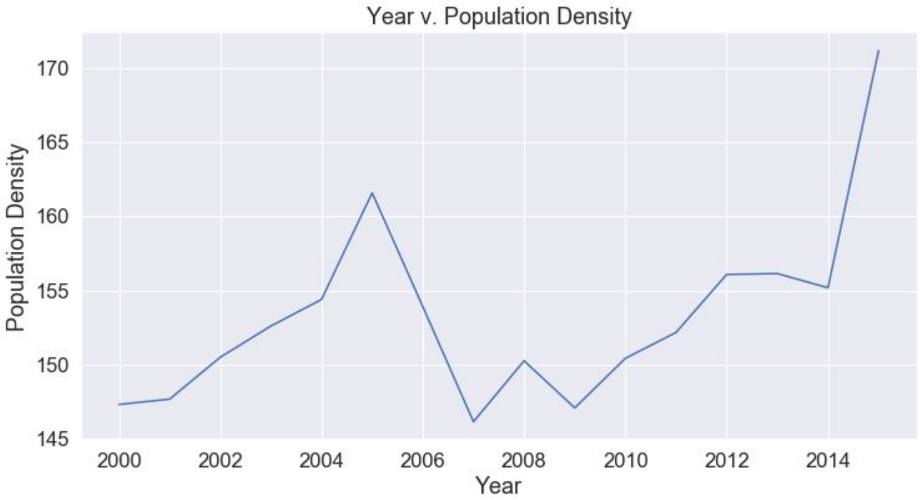

Suicides by Latitude
Suicides per 100k population
Latitude
0
5
10
15
20
25
30
35
40
40
45
50
55
60
Lithuania
Hungary
Belarus
Latvia
Slovenia
Ukraine
Estonia
Austria
Croatia
Serbia
France
Finland
Belgium
Bulgaria
Switzerland
Czech Republic
Luxembourg
Germany
Poland
Sweden
Denmark
Norway
Romania
Slovakia
Portugal
Netherlands
Ireland
Montenegro
Spain
Italy
United Kingdom
Greece
Georgia
Turkey
Azerbaijan

r= 54.5%

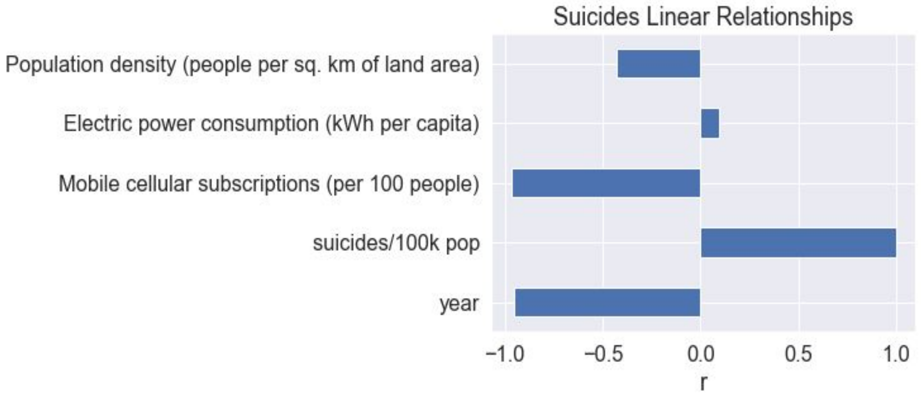

Suicides by Divorce Rate
Suicides per 100k (crude)
Mean Divorce Rate per 1000 (crude)
0
5
10
15
20
25
30
35
40
1.0
1.5
2.0
2.5
3.0
3.5
Lithuania
Hungary
Belarus
Latvia
Slovenia
Estonia
Ukraine
Austria
Croatia
Finland
Serbia
France
Belgium
Bulgaria
Switzerland
Czech Republic
Luxembourg
Poland
Germany
Sweden
Denmark
Norway
Slovakia
Romania
Portugal
Ireland
Netherlands
Montenegro
Spain
Italy
United Kingdom
Greece
Georgia
Turkey
Azerbaijan

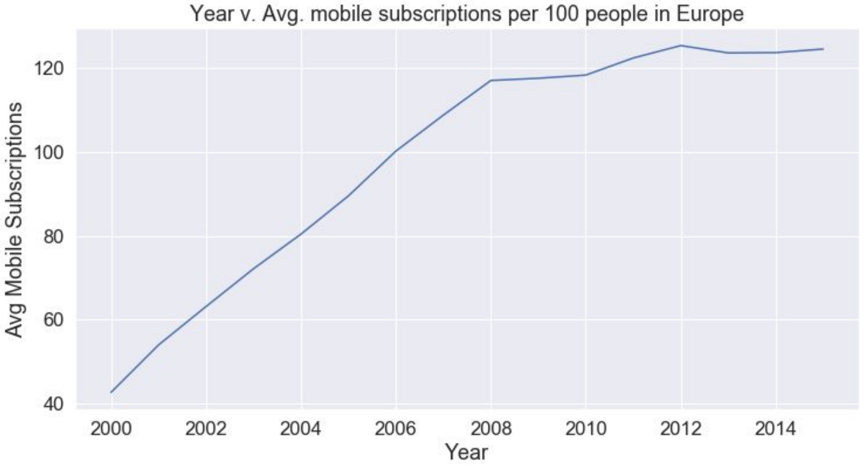
r=57.5%


Suicides by Divorce Rate
Suicides per 100k (crude)
Mean Divorce Rate per 1000 (crude)
0
5
10
15
20
25
30
35
40
1.0
1.5
2.0
2.5
3.0
3.5
Lithuania
Hungary
Belarus
Latvia
Slovenia
Estonia
Ukraine
Austria
Croatia
Finland
Serbia
France
Belgium
Bulgaria
Switzerland
Czech Republic
Luxembourg
Germany
Poland
Sweden
Denmark
Norway
Slovakia
Romania
Portugal
Ireland
Netherlands
Montenegro
Spain
Italy
United Kingdom
Greece
Georgia
Turkey
Azerbaijan
Region
Western Europe
Mediterranean
Eastern Europe
Scandanavia
latitude
32.0
40.0
48.0
56.0
64.0



The largest correlation to suicides are year and Mobile cellular subscriptions.

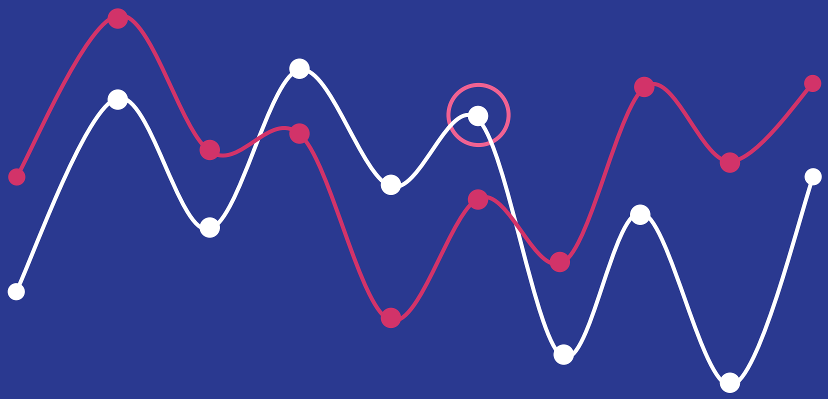




Exploratory Findings:

1. Suicides are generally decreasing over the last 20 years in Europe.
2. Per capita, elderly people (75+) are the most at risk population.
3. Men are three times more likely to die from suicide when compared to women.
4. Latitude and divorce rates appear to have a correlation to suicide rates.
5. Mediterranean countries as a group have the lowest divorce and suicide rates in Europe.
6. Eastern European countries as a group have higher suicide rates.
7. Mobile cellular subscriptions and year have a strong negative correlation to suicide rates.

Statistical Analysis



Hypothesis Testing: Divorce Rates

Null Hypothesis: Divorce rates and suicides are not correlated to each other. Therefore the slope of a linear regression would be 0.

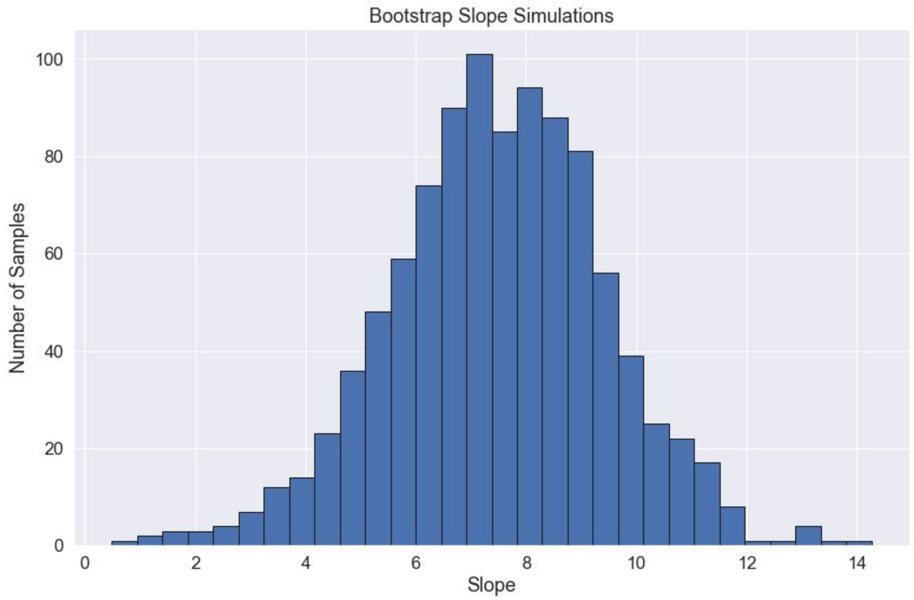
Alternative Hypothesis: Divorce rates and suicides are positively correlated.

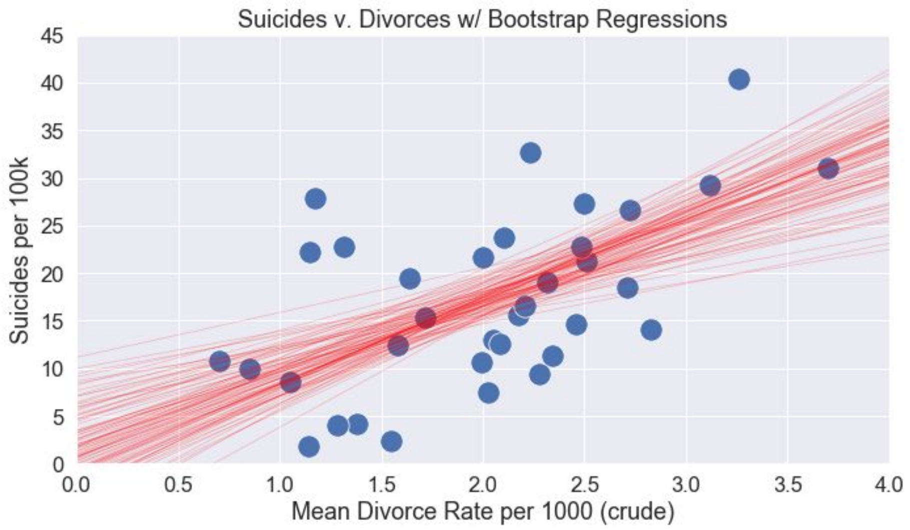
Methodology: 1. Run bootstrap simulations to resample the data with random selections of the data points (countries) with replacement.

2. Repeat this process 1000 times and extract an array of all of the slopes and intercepts of the regressions in each randomized array.

3. Score and plot each regression.

Divorces vs. Suicides





Out of the 1000 simulations, none of the simulations produced a slope 0 or less, therefore we should reject our null hypothesis due to an extremely small p-value.

Hypothesis Testing: Latitude

Null Hypothesis: Latitude and suicides are not correlated to each other. Therefore the slope of a linear regression would be 0.

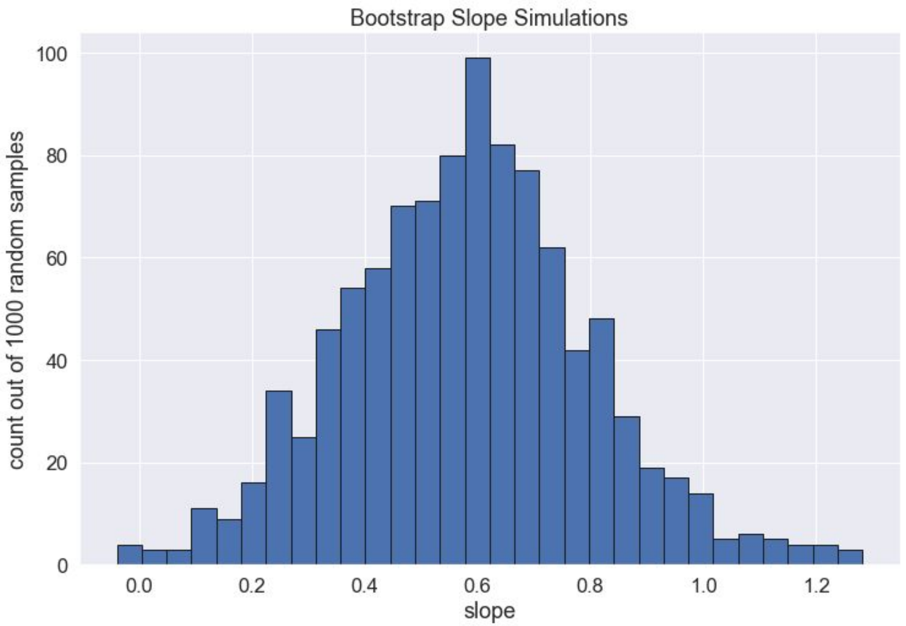
Alternative Hypothesis: Country Latitude and suicides are positively correlated.

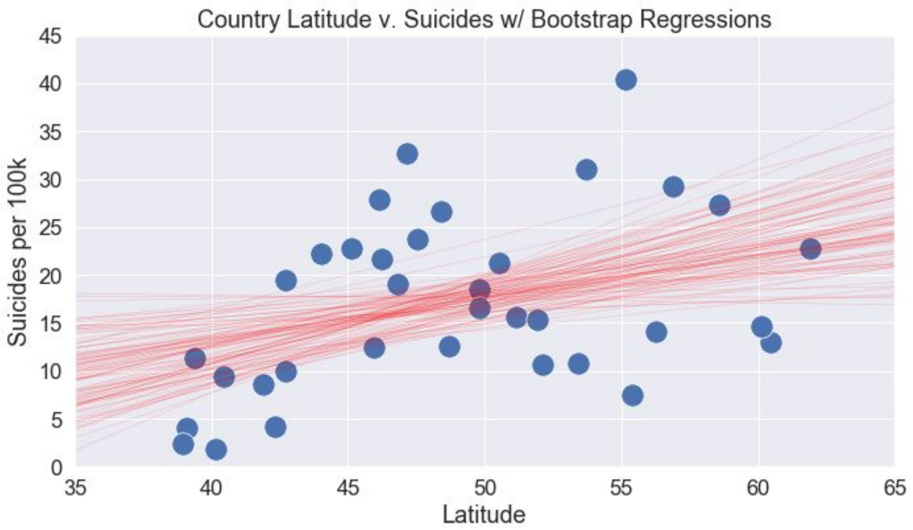
Methodology: 1. Run bootstrap simulations to resample the data with random selections of the data points (countries) with replacement.

2. Repeat this process 1000 times and extract an array of all of the slopes and intercepts of the regressions in each randomized array.

3. Score and plot each regression.

Country Latitude vs. Suicides

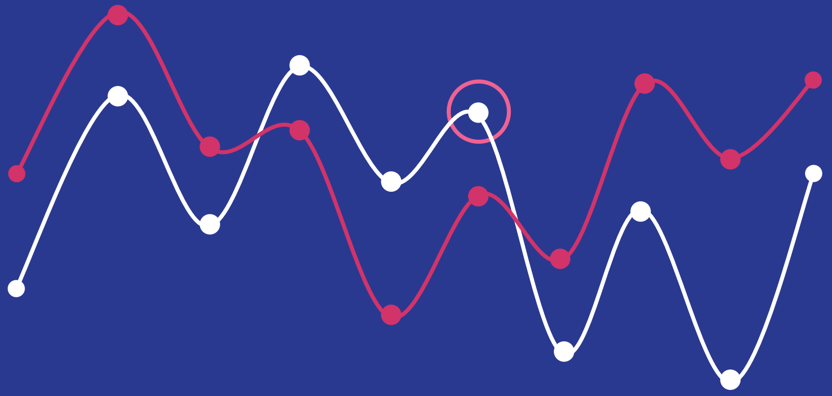




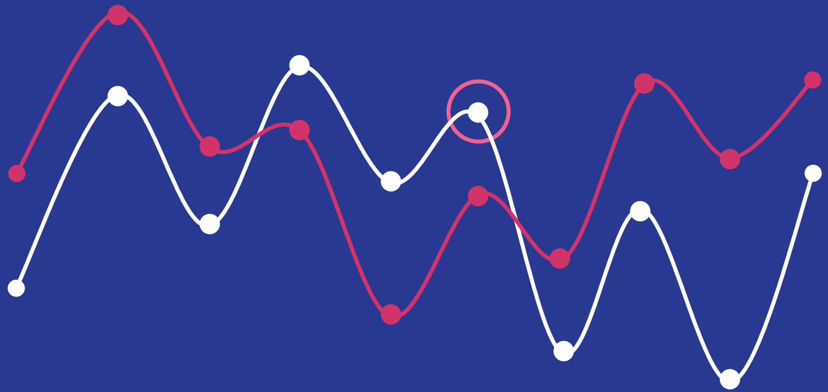
Out of the 1000 simulations, 0.4% of the simulations produced a slope 0 or less, therefore we should reject our null hypothesis due to a small p-value.

Statistical Findings:

1. Divorce rates and country latitude are highly likely to be correlated to suicide rates.
2. We can now develop a predictive model with these variables, along with mobile cellular subscriptions and year.



Modeling the Data



StatsModels: OLS results

Single Variable Model:

Dep. Variable:	suicides_100k	**R-squared:**	0.185
Model:	OLS	**Adj. R-squared:**	0.182
Method:	Least Squares	**F-statistic:**	71.92
Date:	Tue, 03 Mar 2020	**Prob (F-statistic):**	8.64e-16
Time:	18:46:06	**Log-Likelihood:**	-1049.8
No. Observations:	319	**AIC:**	2104.
Df Residuals:	317	**BIC:**	2111.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.7514	1.055	6.402	0.000	4.676	8.826
divorce_rate	4.0421	0.477	8.481	0.000	3.104	4.980

Using only divorce rates, **R2 is 0.185**, accounting for .185 of the variance in suicides.

Our **F-Stat is 71.92**, which is significant.

AIC is 2104

StatsModels: OLS results

6 Variable Model:

Dep. Variable:	suicides_100k	R-squared:	0.455
Model:	OLS	Adj. R-squared:	0.444
Method:	Least Squares	F-statistic:	43.38
Date:	Tue, 03 Mar 2020	Prob (F-statistic):	2.08e-38
Time:	18:46:10	Log-Likelihood:	-985.68
No. Observations:	319	AIC:	1985.
Df Residuals:	312	BIC:	2012.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	439.9245	214.666	2.049	0.041	17.549	862.300
Mobile_subs_100	0.0550	0.016	3.461	0.001	0.024	0.086
year	-0.2206	0.107	-2.062	0.040	-0.431	-0.010
divorce_rate	2.9087	0.451	6.448	0.000	2.021	3.796
latitude	0.2407	0.059	4.090	0.000	0.125	0.357
gdp_pc	-0.0001	1.2e-05	-9.448	0.000	-0.000	-8.95e-05
population	-1.139e-06	1.71e-07	-6.660	0.000	-1.48e-06	-8.03e-07

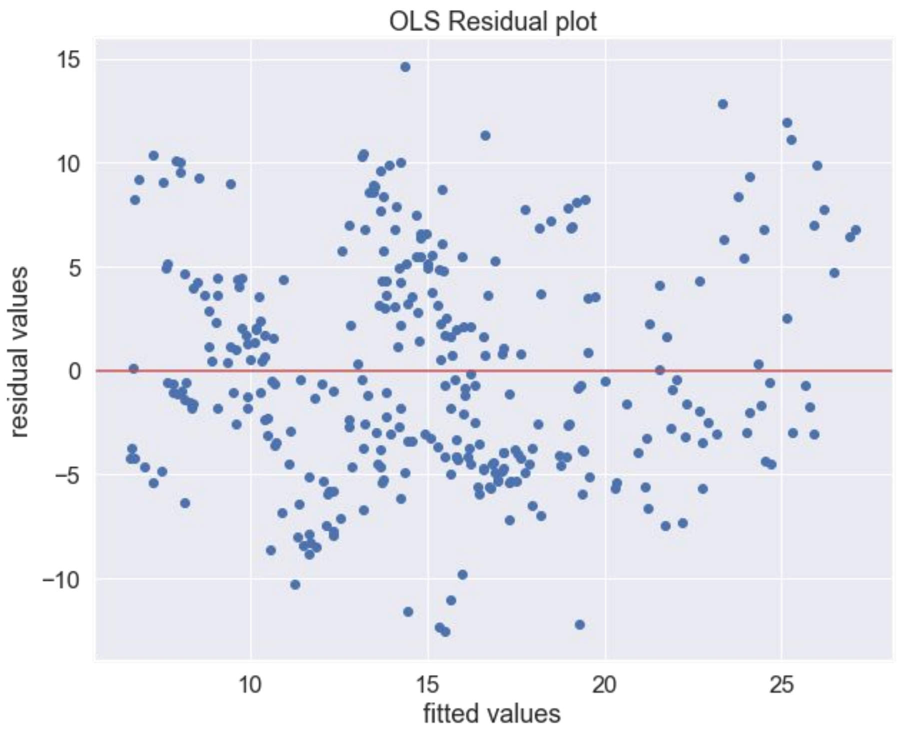
Using mobile subs per 100, year, divorce rate, latitude, gdp per capita, and population, **R2 is 0.455**, accounting for .455 of the variance in suicides.

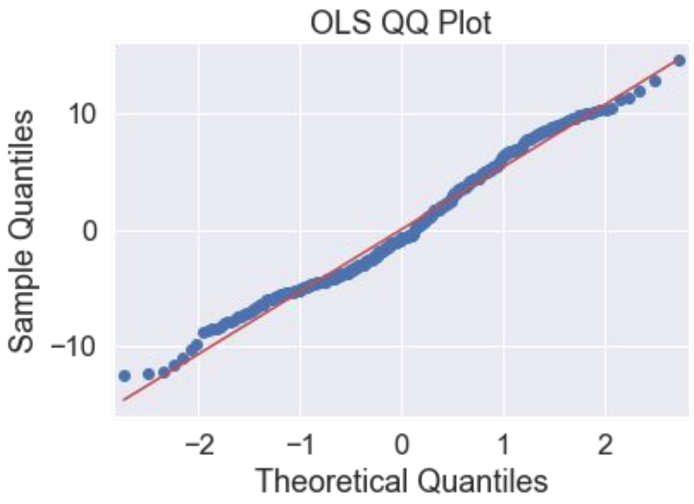
Our **F-Stat decreases to 43.38**, which is still significant. All individual variables have a p-value less than 0.05.

AIC decreases to 1985.

This model is our best thus far because it increases R2 and decreases AIC.

OLS results - 6 Variable Model





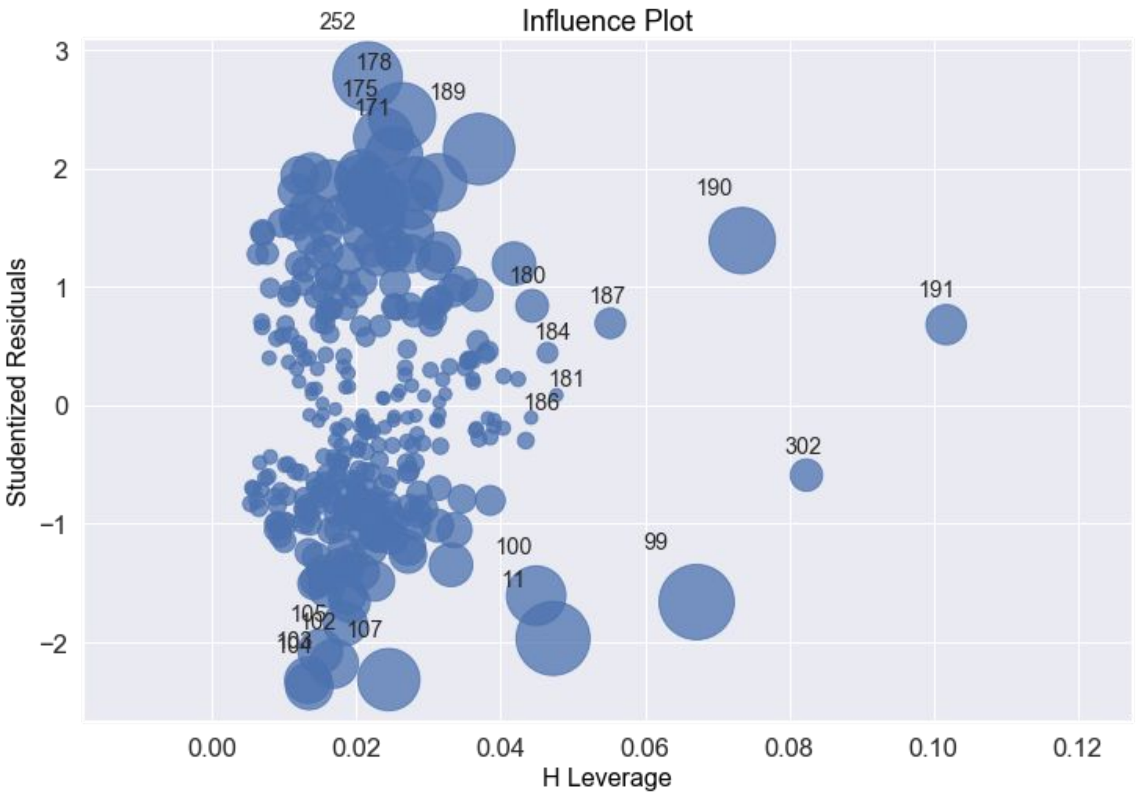
The residual plot and QQ plot suggest that a **linear model is the right choice** for our data. In a linear model, **residuals are uncorrelated and distributed normally.**

Handling Outliers

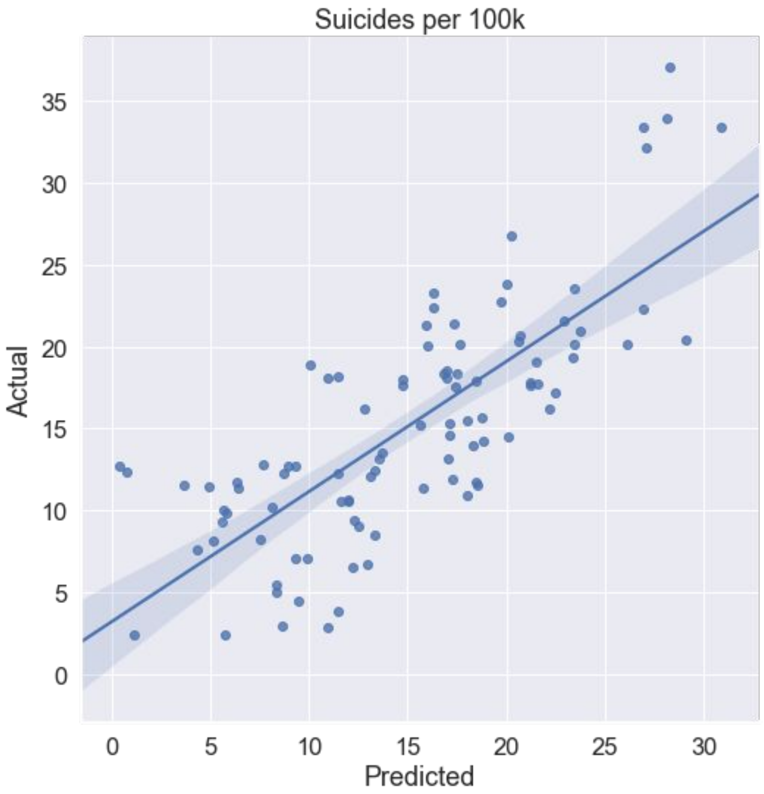
A high influence point has a **relatively large bubble** and is likely to influence our regression.

While there are some high influence points, none of them are invalid/misinterpreted data points.

Therefore, **no data points are removed.**



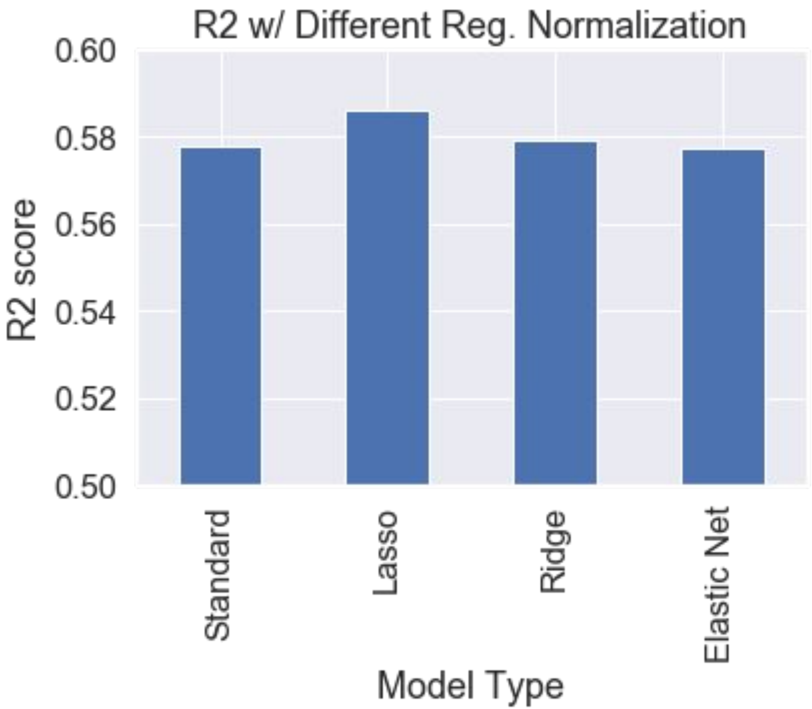
Further Tuning w/ SciKit-Learn



We use all 20 predictor variables, with the idea of normalization.

Our initial model accounts for an R2 score of 0.5577.

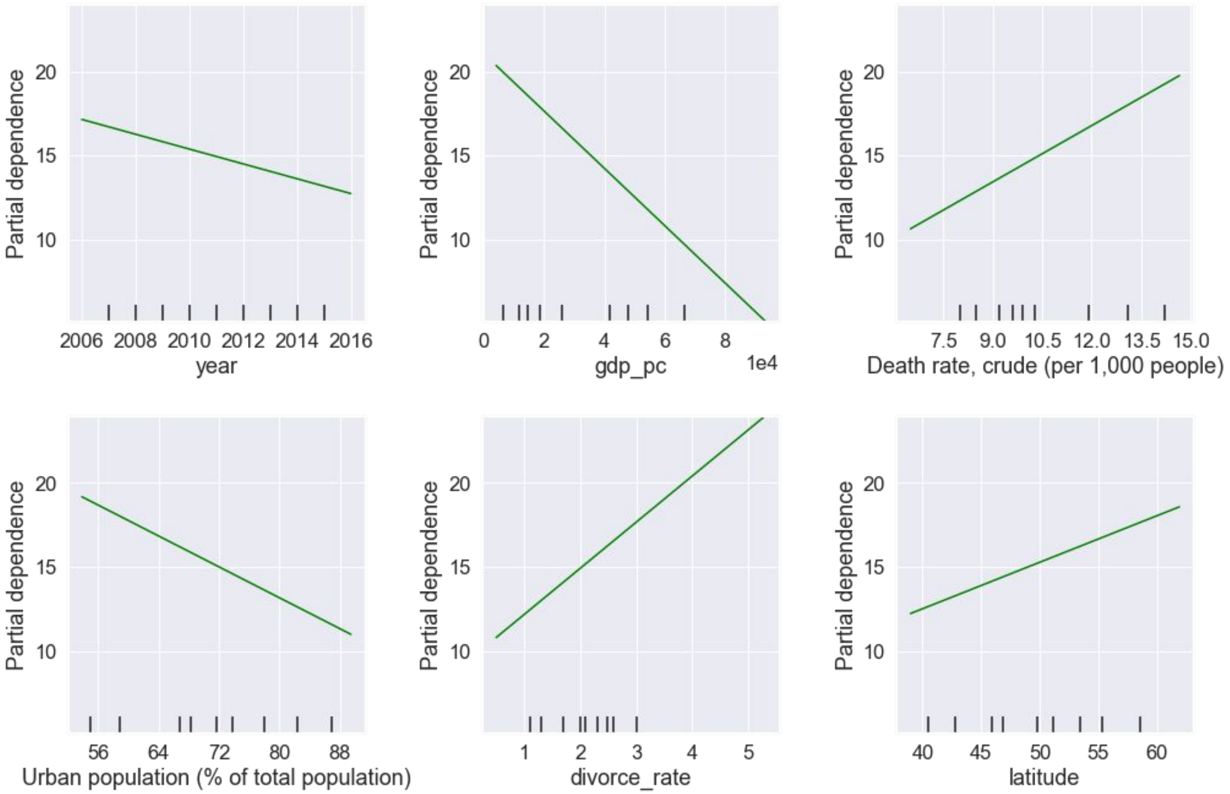
Further Tuning w/ SciKit-Learn



	R2 val
Standard	0.578
Lasso	0.586
Ridge	0.579
Elastic Net	0.577

Lasso Regularization, by reducing poor predictors towards 0, **yields the highest R2 score at 0.586**

Partial Dependence: Most Powerful Predictors

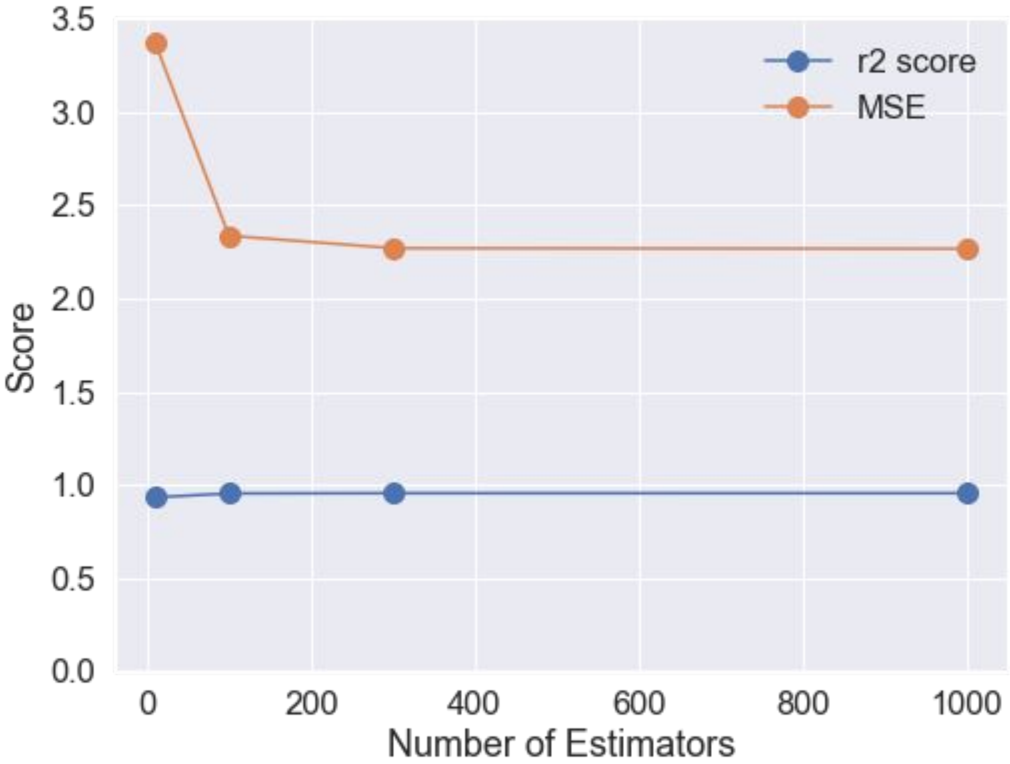


These 6 variables have the highest influence over suicide rates.

Higher divorce rates, crude death rates, and country latitudes **predict higher suicide rates**.

Higher urban population%, gdp per capita, and more years in the future **predict lower suicide rates**.

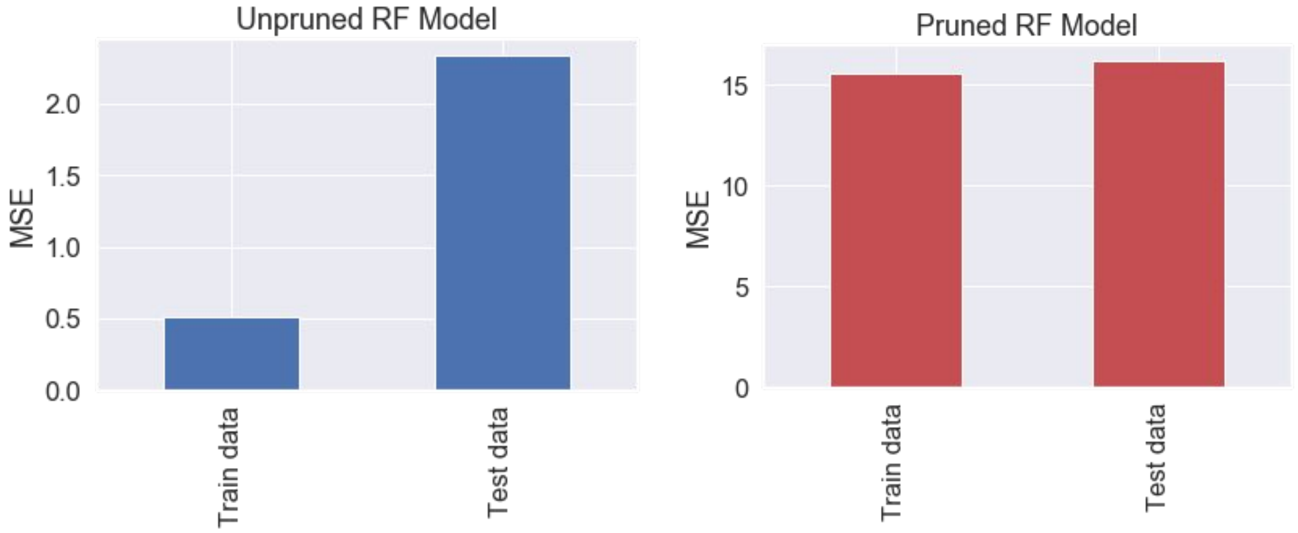
Random Forest Regression



A random forest fits decision trees on various sub-samples of our training data and aggregates averages to improve the predictive accuracy and control over-fitting.

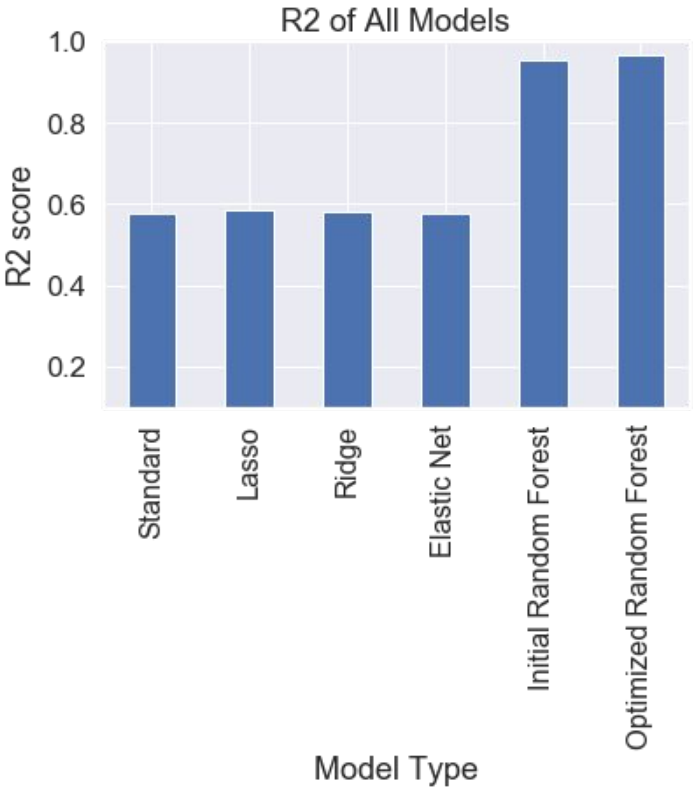
MSE drops and R2 increases with more estimators, but it reaches a limit.

Pruning: Random Forest Regression



To reduce overfitting, I compared an unpruned model with a pruned model. Although the pruned model has a very high MSE for both train and test data, **the pruned random forest** has less variability between the 2 scores, which means **less overfitting**.

High Performance RF with GridSearchCV



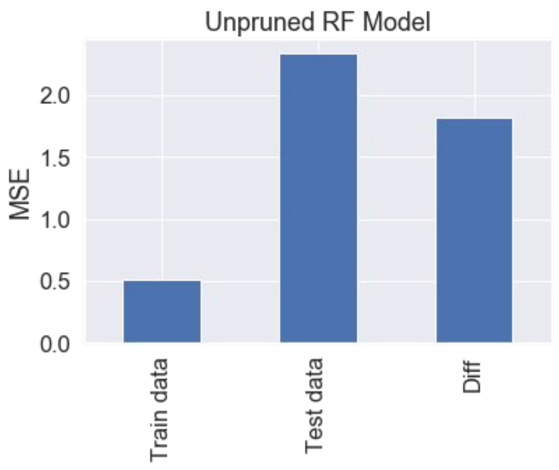
To help optimize the model, I used a grid search to filter many combinations of tuning variables for the optimal model. **Our high performance model was able to account for 0.964 of the variance**

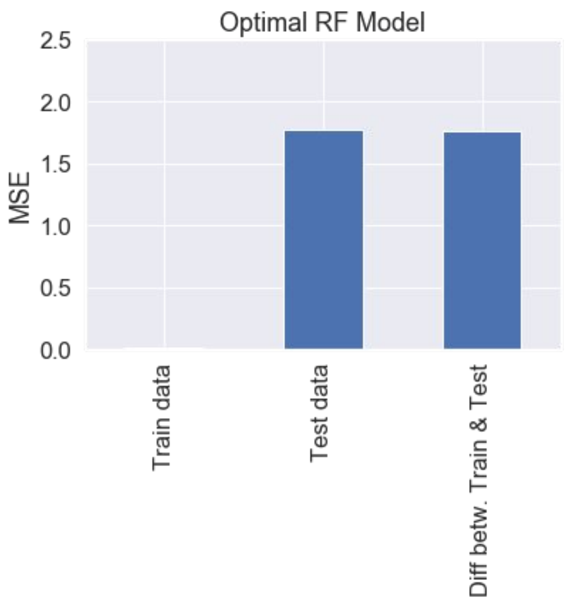
```
]: {'bootstrap': False,
    'max_depth': 100,
    'max_features': 5,
    'min_samples_leaf': 1,
    'min_samples_split': 2,
    'n_estimators': 300}
```

```
]: gspred= gs.predict(X_test)
   explained_variance_score(y_test, gspred)
```

```
]: 0.9648106850265727
```

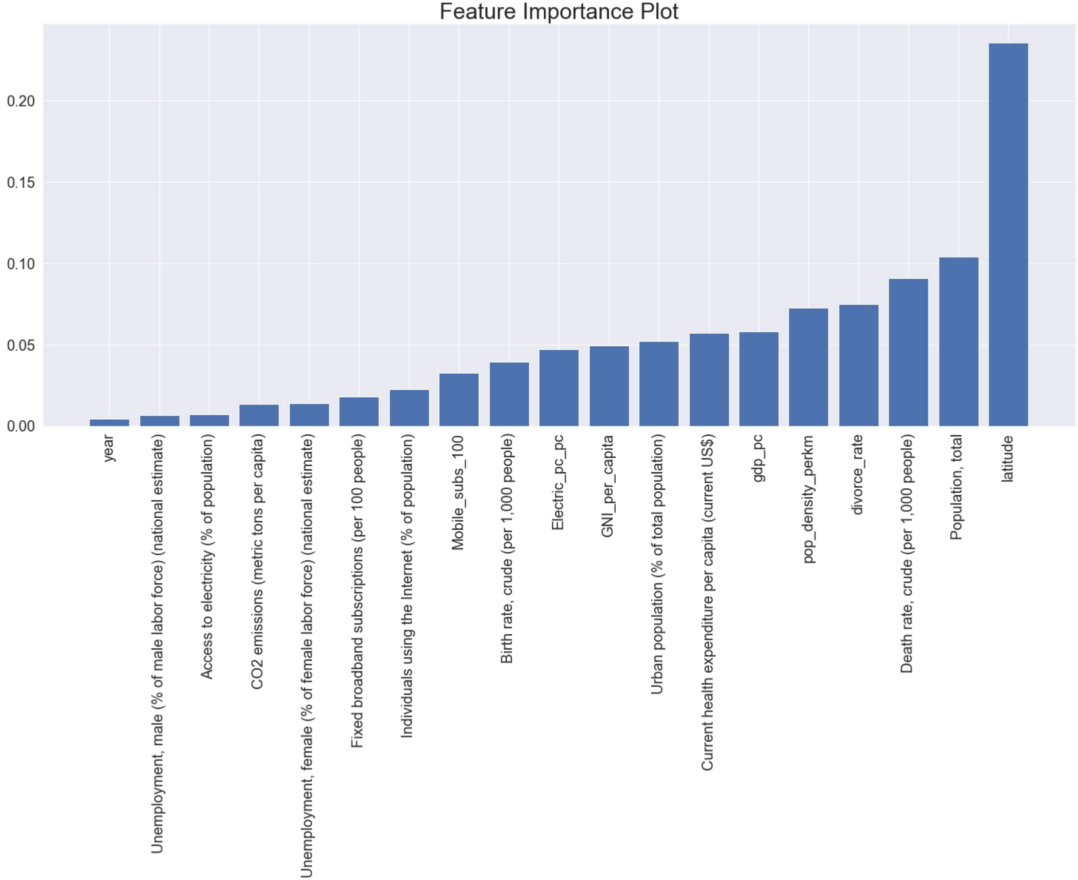
High Performance RF with GridSearchCV





Our optimized model is likely overfitted with a training MSE of 0.0127, but it achieved the lowest Test MSE and was able to reduce the difference between the training and test MSE.

Random Forest Feature Importance Plot



The optimized Random Forest determined that **latitude, population, crude death rate, divorce rate and population density** were the most important features for predicting suicide rates.

Overall Findings:

1. **The Random Forest Model outperformed our other models** by a wide margin, but it's **likely to be overfitted** to the training data.
2. **From partial dependence, our most predictive variables** are latitude, divorce rates, urban population %, gdp per capita, crude death rates, and the year.
3. **From our Random Forest model, our most important features** are latitude, population total, crude death rates, divorce rates and population density.
4. Latitude and divorce rates are in the top 5 most important features for the Lasso Model and the RF Model.
5. **Men are thrice as likely to commit suicide** compared to women and **seniors (75+) are the most susceptible age group** per capita.
6. **Suicides are trending lower over time in Europe**.
7. **Characteristics of modernization**, like high mobile subscriptions, GDP's, and urban populations, **are correlated with decreases in suicides** per capita.