World Suicide Data

Satya Josyula

June 9, 2019

About

World Suicide Data - Analysis
This document is prepared as part of the project submission for Data Science Professional Certification

---

Preface

This exercise is an attempt to analyze the suicides data from across the world provided by the World Bank. The ultimate act of someone taking their own life depends on several Social, Economic and Political factors. With the wide disparities in the economic and social environment, political stability across the countries and regions of the world, this exercise tries to bring in few more factors that could be effecting the suicide rates across the world. More features from other data sources are added to the suicide data to get better insights into factors leading to the varied intensity of suicides. Various models are explored to predict the suicide rates across the countries for different population groups. Data is available from 1985 to 2016, with data for 2016 being very sparse.
Note: World Geo-spatial data is used to present disparities on the world map. Required software would need to be installed in the machine running this program.

Data Source used

The data source used is from Kaggle, sourced from World Bank under the Terms of Use as listed below.
https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets
World-wide geometric and economic indicators from Natural Earth Data libraries.

Load data from the CSV to Data Frame.

Load data into a data frame and remove any redundant data columns and alter column names to make them more inline with rest of the columns.

Get world geo data to get more insights into region wise intensity of suicides

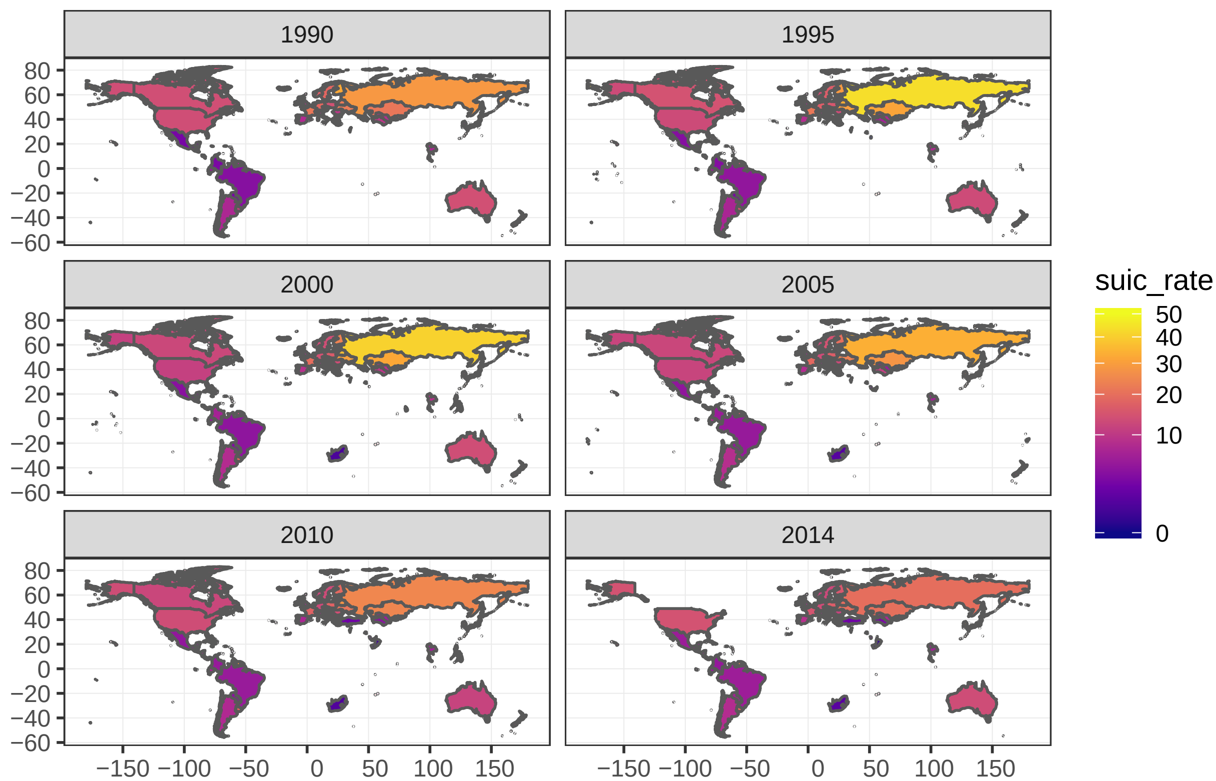
We combine some economic indicators of the countries in the world with the suicide rates data available and find any correlations between these indicators and suicide rates.

	countryid	agegroupid	generationid	suicides_no	year	sexid	population	gdp_per_capita
countryid	1.000	0.000	0.006	0.106	0.031	0.000	0.122	-0.042
agegroupid	0.000	1.000	-0.390	0.080	0.003	0.000	-0.061	0.001
generationid	0.006	-0.390	1.000	-0.043	0.236	0.000	0.014	0.084
suicides_no	0.106	0.080	-0.043	1.000	-0.004	-0.146	0.616	0.060
year	0.031	0.003	0.236	-0.004	1.000	0.000	0.009	0.342
sexid	0.000	0.000	0.000	-0.146	0.000	1.000	0.011	0.000
population	0.122	-0.061	0.014	0.616	0.009	0.011	1.000	0.078
gdp_per_capita	-0.042	0.001	0.084	0.060	0.342	0.000	0.078	1.000

Correlation matrix of the suicides data combined with world data shows not much correlation between the variables.

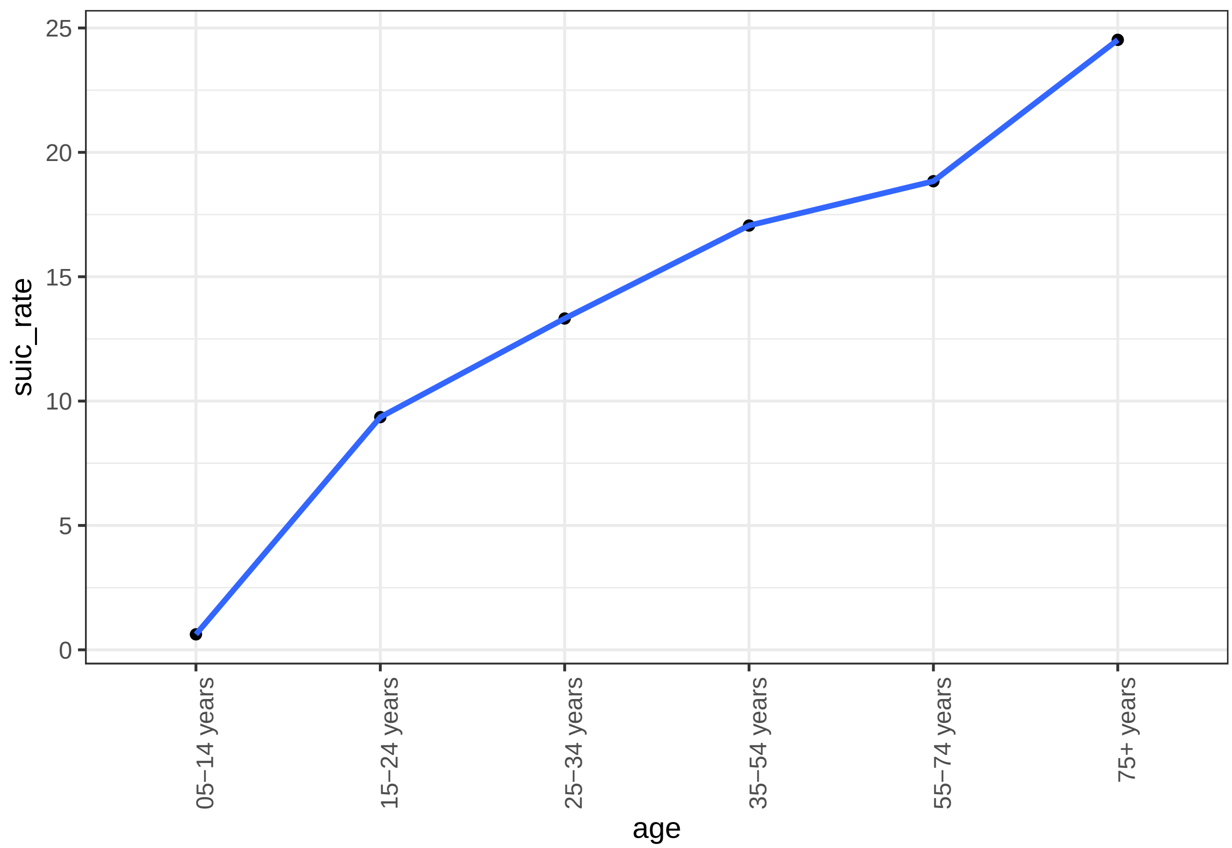
Variations in suicide rates across the countries in the world.

The data we merged has some macro indicators about the economies, regions they belong among others. These indicators help us get better insights into the suicide trends across several other factors than the ones currently available. We can see the heat mapped presentation of suicide rates across the world at file year intervals since 1990.



Variations in suicide rates across different age groups

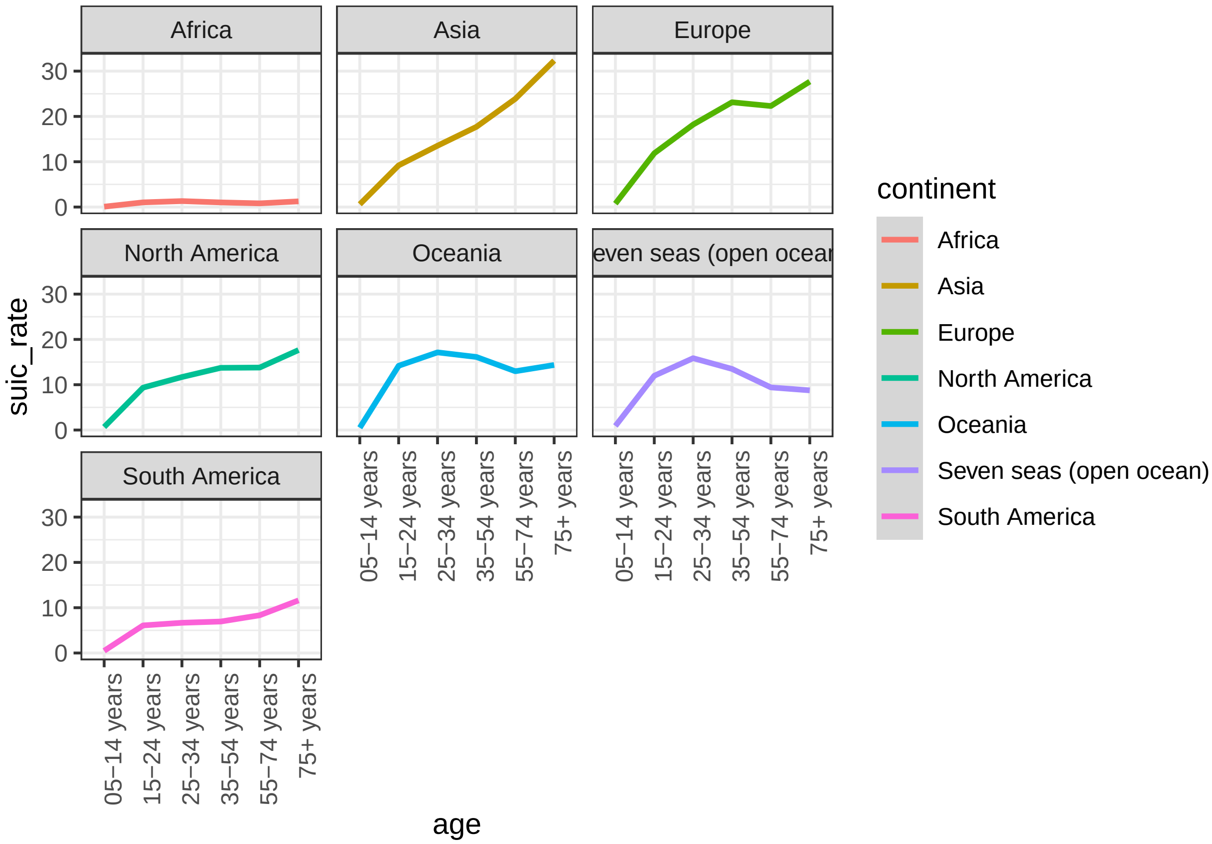
Causes for suicides vary across different age groups. Following image shows how suicide rates are influenced by age.



Suicide rate across different age groups shows uptrend in rate as population ages.Age and mix of the population in different age groups influences the overall suicide rate for a country. Let us see if the same trend is there in all continents of the earth.

Variations across different age group in various continents.

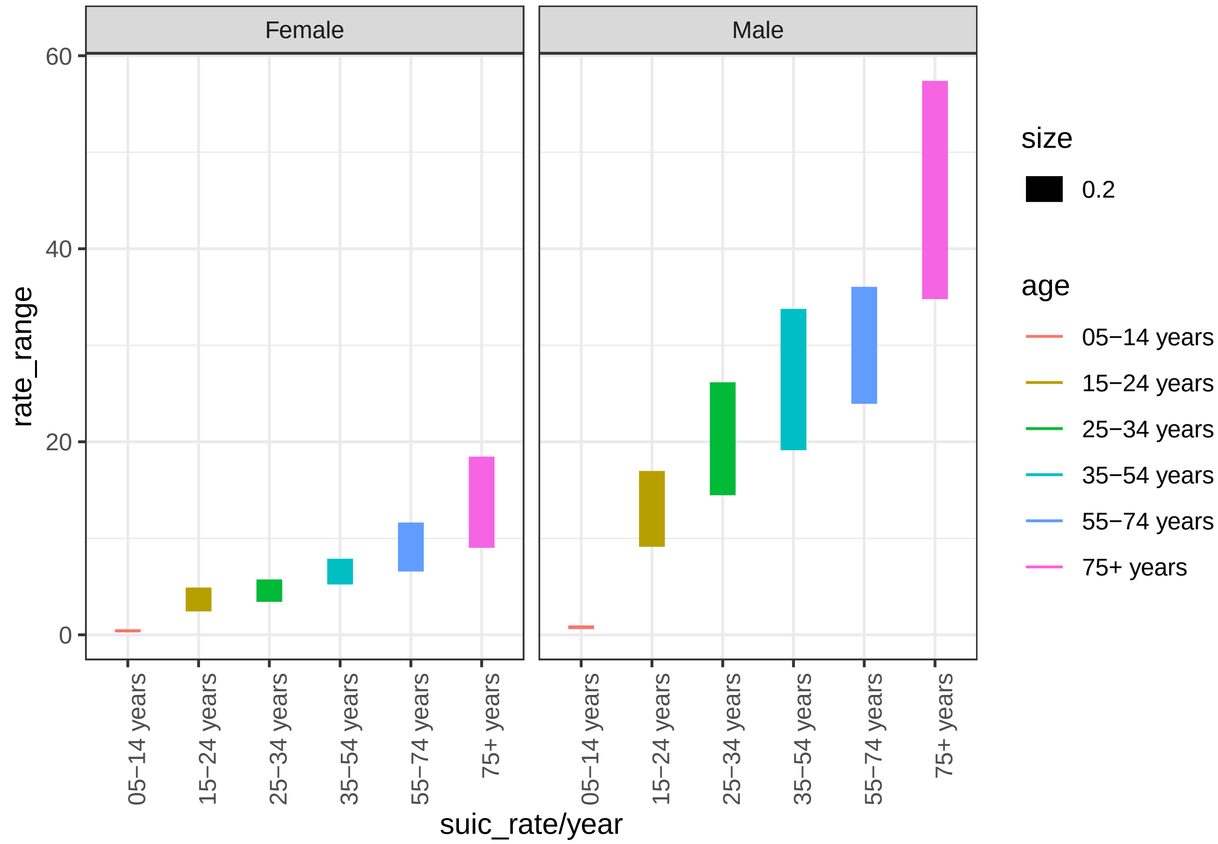
Following image shows the trends in suicide rates across ages in different contents of the world.



If we look at the trends of suicides as the population ages, all major continents show almost similar trend of rising suicide rates as the population ages. But Ocenia and Open seas see a peak in the age group of 25-34 years and comes down and flattens with age. Rise in suicide rates as population ages is more steep in Asia and Europe than South America and North America.

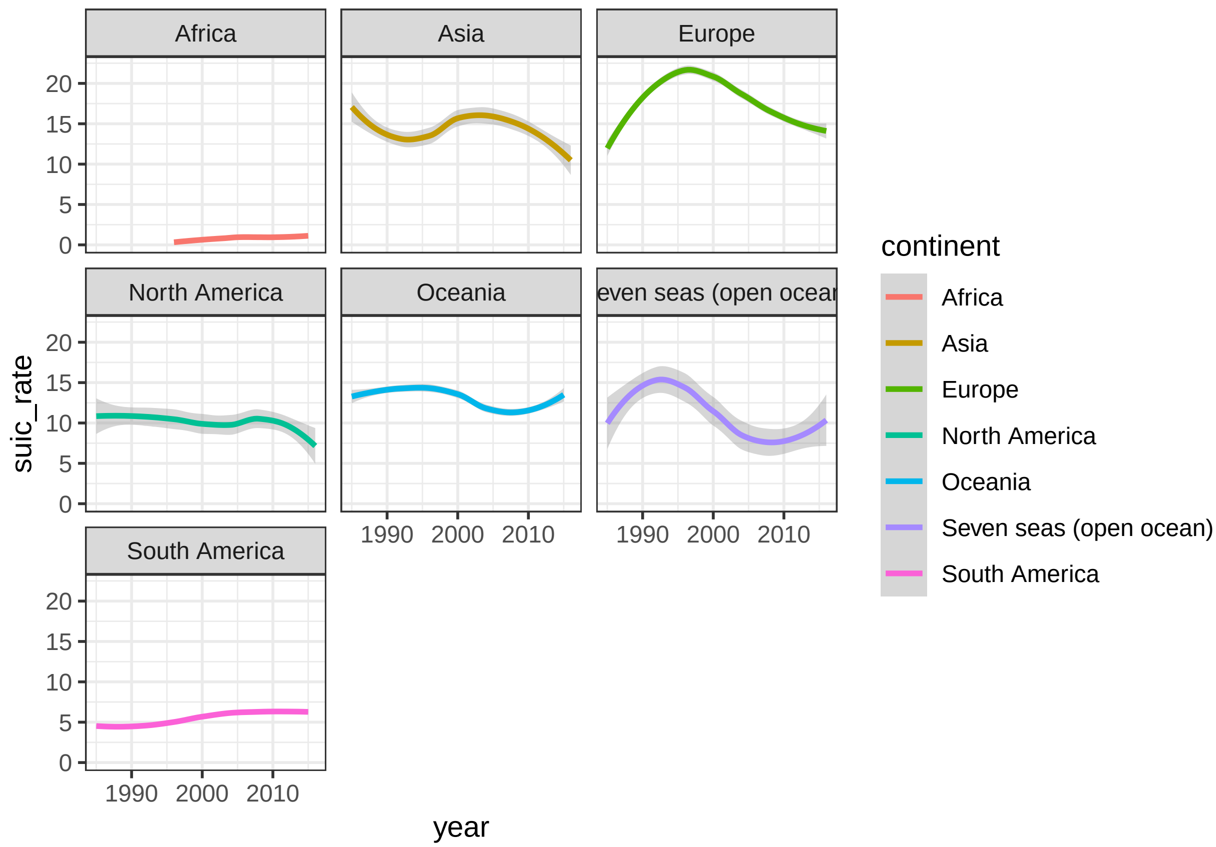
Variations in suicide rates across different age groups for male and female

Male and Female population react differently to different situations. That would reflect in the extreme steps they take in different situations. Following graph shows the trends for both sexes.



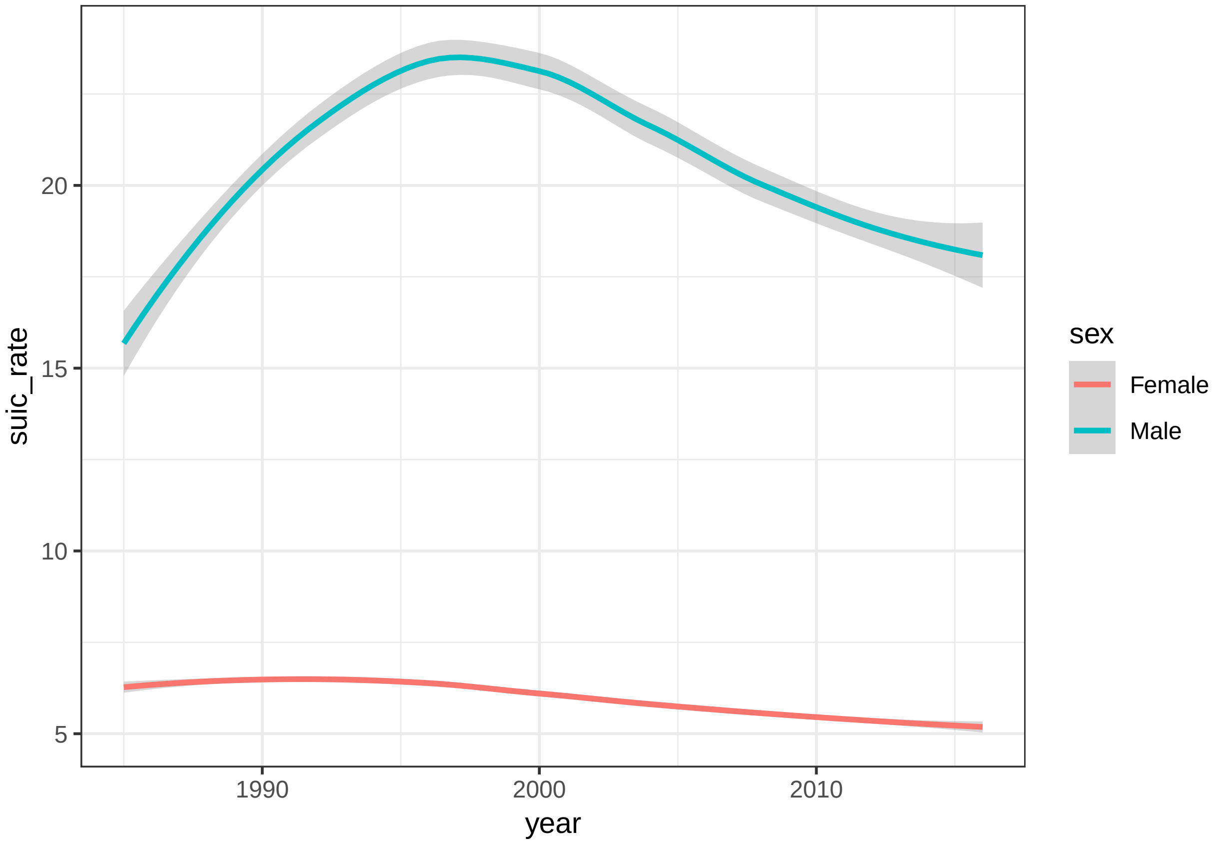
Suicide rates among men are significantly higher than women across all the age groups, except 5-14 years. The rise in suicide rates among women is not as intense as men as they age. Sex could be a dominant factor in determining suicide rates.

Overall suicide rates across continents.



Suicide rates in continents Asia, North Americ and Europe have seen declining trend for the past two decades. South America and Africa (South Africa) see a little uptrend. Other Oceania and Open Oceans see a cyclic trend.

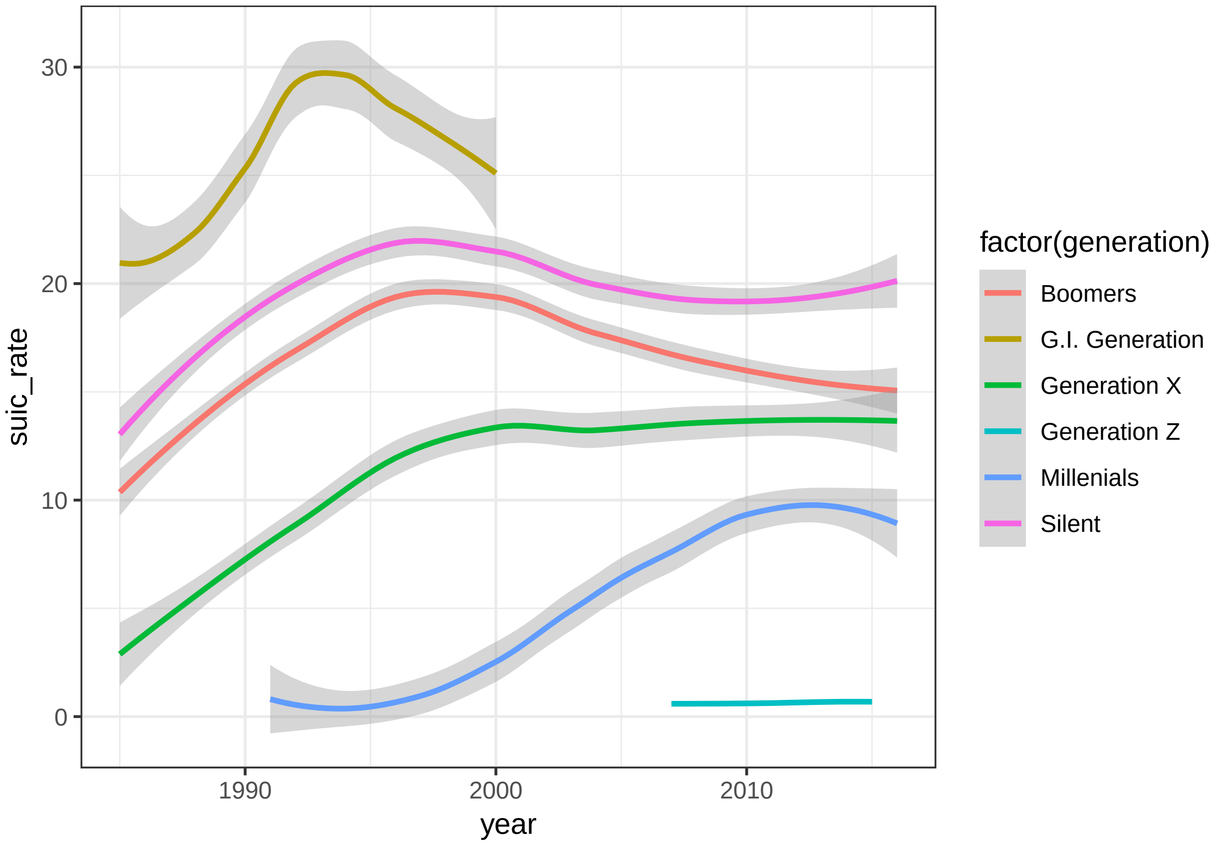
Suicide trends across years for both sexes.



Suicide rates among men are almost three times more than women. Rates among both men and women are declining in the same proportion.

Suicide rate trends across different generations.

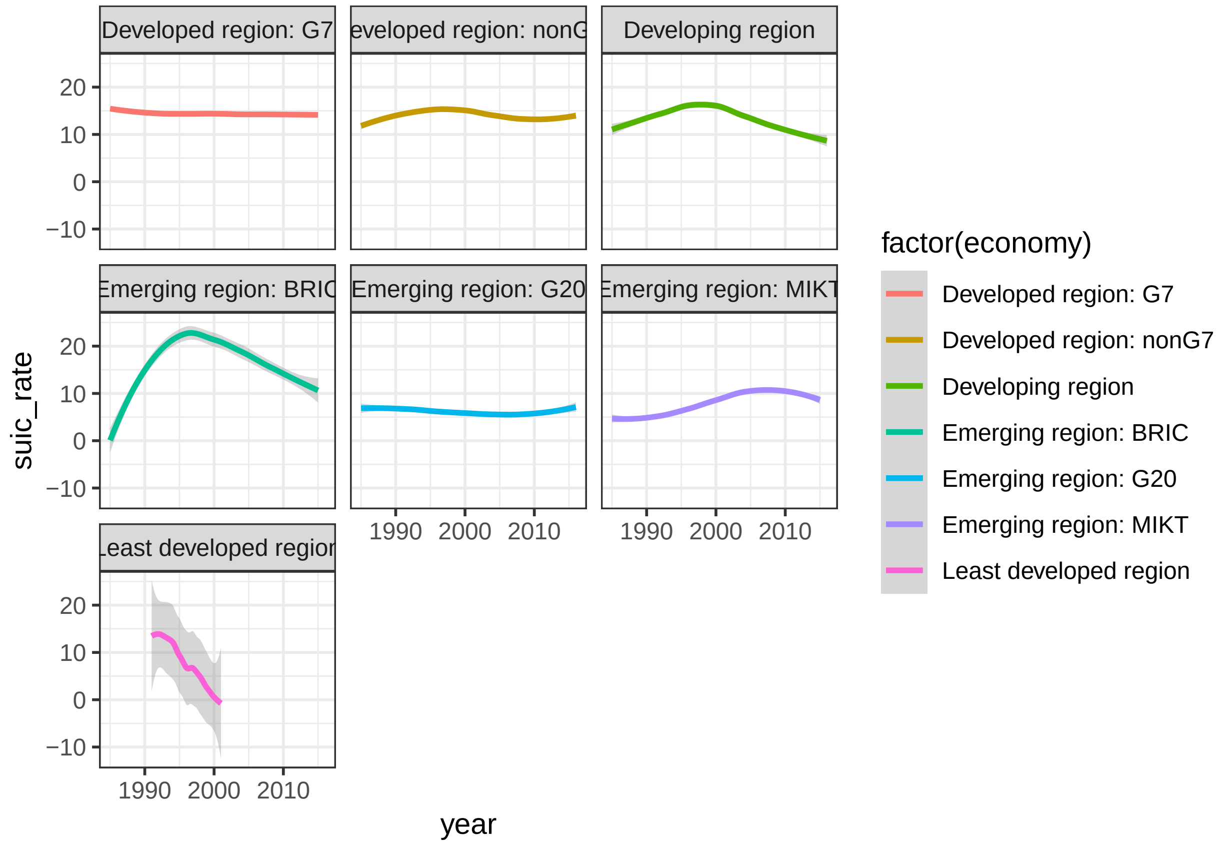
Causes for suicides vary for each generation of population. They could be factors ranging from economic, health, age etc.



The Generation wise suicide trends show that Generation-X have a raising suicide rate in general. Millennials experienced a rising suicide rate and started declining. Silent generation have seen a decline in rates but started to see rising suicide rates again. Generation Z, the youngest ones have a flat suicide rate. Boomers see declining suicide rates.

Influence of Economy on Suicide rates.

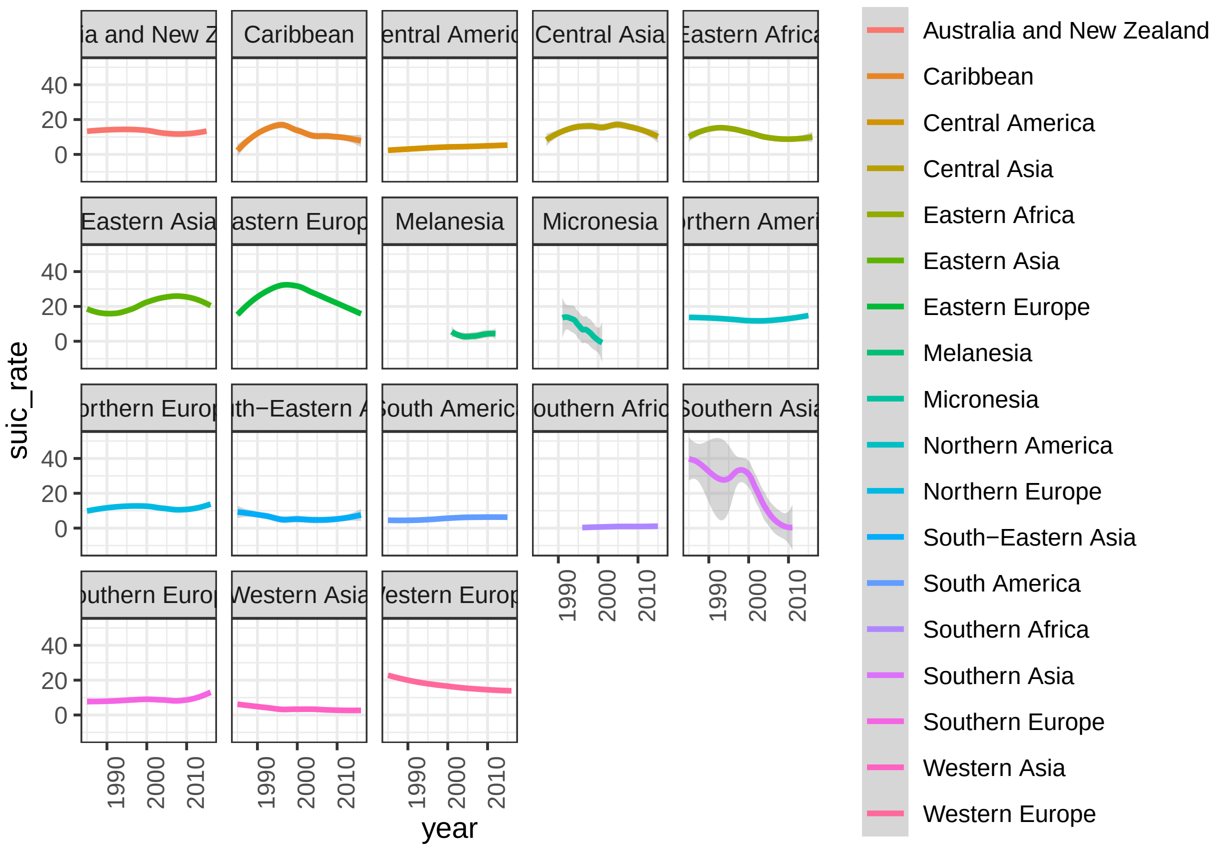
Macro economic factors could cause variations in suicide rates. Stress in Developed economies and poverty in less developed economies could be causes for suicides.



If we look at countries as w whole in various economic development stages, Developed Regions have a constant suicide rate of around 15. Developing and Emerging nations (BRIC) have declining suicide rates. Emerging Regions (MIKT) had seen an uptrend and started declining. Least Developed countries are shown to have fastly decling suicide rates, but we dont have enough data.

Variations in suicide rates across Subregions of world.

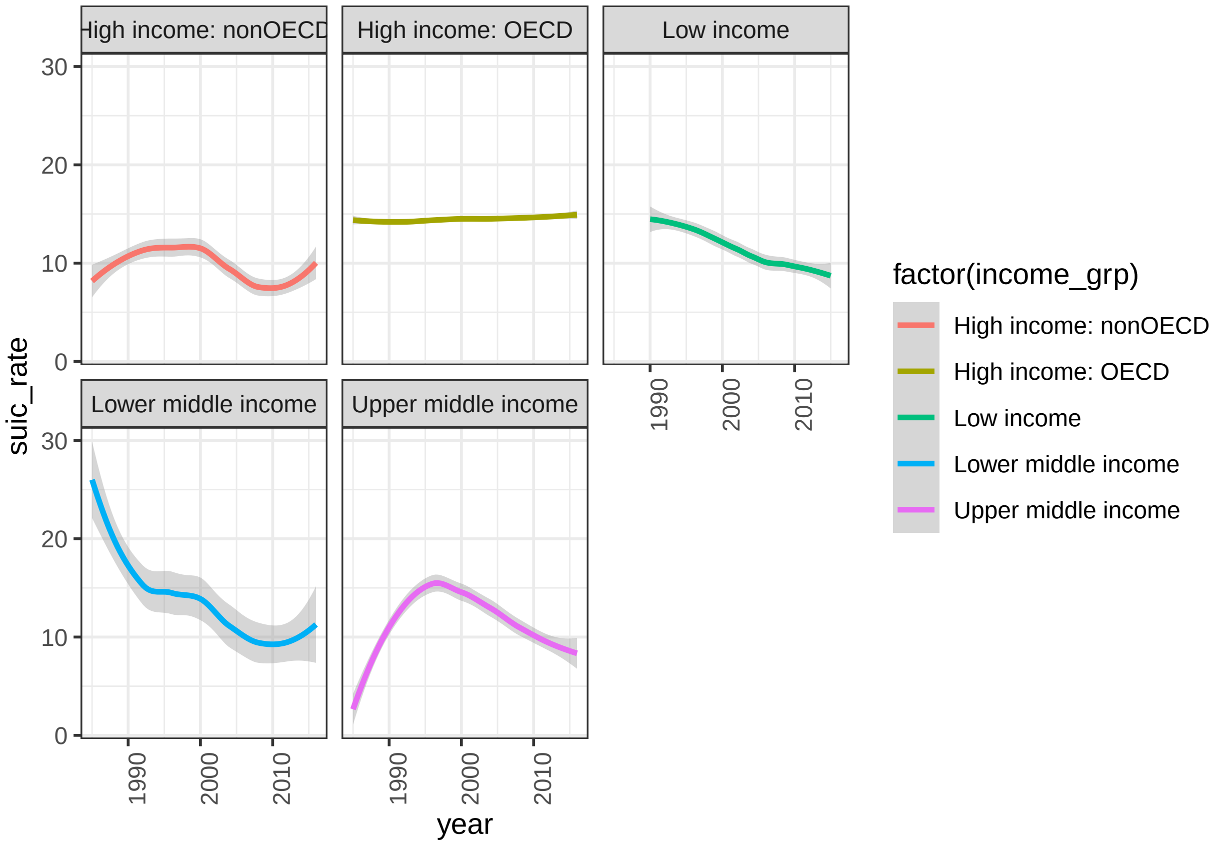
World is divided into several subregions, ecah could have their own culteral and economic factors influencing the suicide rates.



Most of the regions have a suicide rate of less than 20 per 100K of population. Southern Asia registered a decline from peak rate of about 40. Eastern Europe has peaked in Mid-90s and in a declining phase. Micronesia and Western Europe have declining suicide rates. We dont have complete data for Melanesia, Micronesia, Southern Africa, Southern Asia regions.

Income group driven variations in Suicide rates.

This factor could be similar to the economic well being of countries. Income levels of population drive their health and well-being and could have influence of rates of suicides in the population.



It is seen that countries belonging to High Income: OECD group almost constant Suicide rate and High Income: NonOECD have a suicide rate cycle that was in an uptrend. Low Income and Lower Middle Income countries have a declining suicide rate. Upper Middle Income countries have spiked in the rates and declining suicide rates.

Classify Suicide Rates

We first attempt to create a Classification model to get accurate predictions of suicide rates and then create a Regression model to get the best RMSE. As accurate estimates of the suicide rates is not necessary, We will try to categorize the suicide rates into different levels based on the range of suicide rates. We find that most of the rates fall in under 5 per 100K. We will call the new column as Suicides_Intensity. It ranges from 1 to 9. As the dataset is unbalanced and there is lot of data in the lower suicide rates, the data is classified as below. It can be reclassified to make it more balanced, if needed. 0-2 : 1, 2-6 : 2, 6-12 : 3, 12-20 : 4, 20-30 : 5, 30-40 : 6, 40-60 : 7, 60-100 : 8, 100+ : 9

Feature Selection

We begin with the features that originally came with the suicides data. Then we add more indicators from world data and see if any of those features help improve the model. Our target is to select a model that uses lowest number of features and provides us better predictions of suicide rates in a reasonable amount of time given the resources available on the machine running the program. Suicide data is not Categorical data. We try to find an algorithm that finds the minimum RMSE. We also attempt to use Classification algorithms to find Accuracy of the model.

Find High correlation among data

It is important to remove any highly correlated data from the dataset as it will not add much value to the model. We use a cut-off of 0.5. If we find any data that shows more than 0.5 we will remove those columns.

Feature Highly Correlated

We dont find any highly correlated factors. So, we will continue with the variables available to us to predict suicide rates.

Recursive Feature Elimination

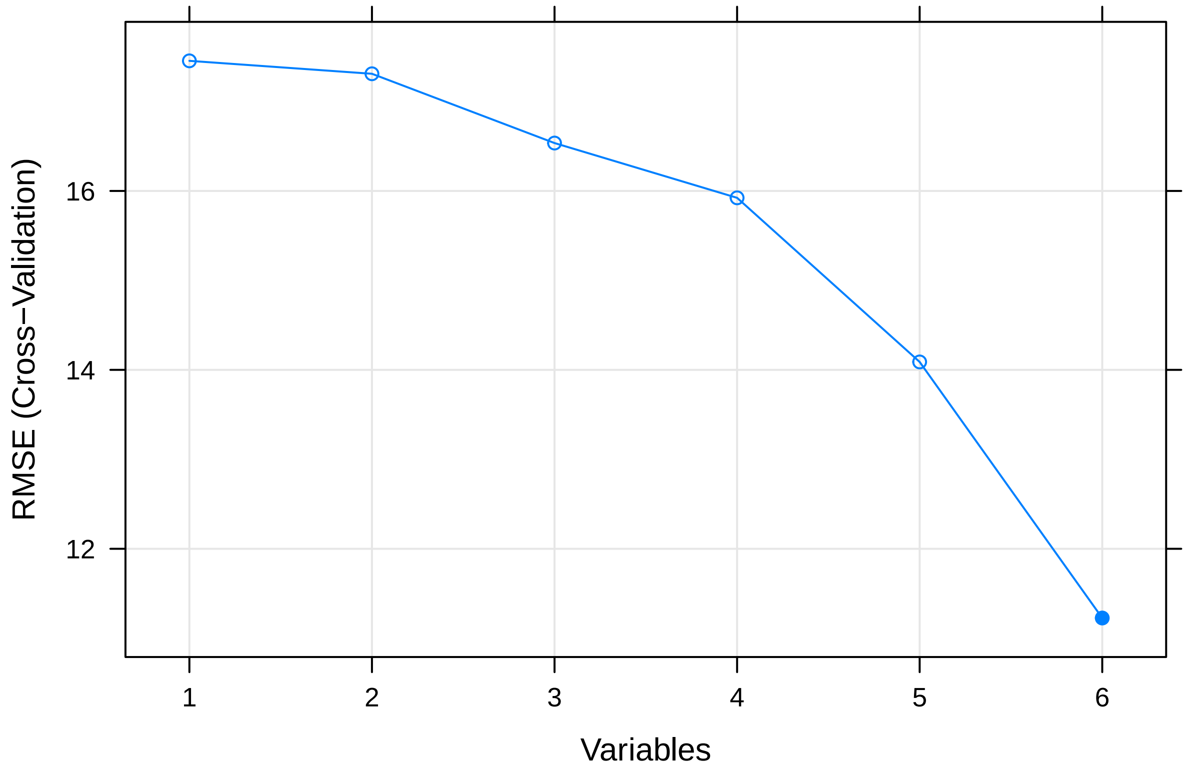
We try Recursive Feature Elimination to select only features that help us predict the suicide rates more accurately. We will use Random Forest (rf), Linear Discriminant Analysis (lda), Linear Model (lm), Bagged CART (treebag) to find the top features that provide us the best predictions.We use simple Cross Validation at 10 for all trainings.

RF Functions

We first use the Random Forest method to find contribution of features to the prediction of the Suicide rates. We also will save the statistics for the parameters that provide us the second best estimates.

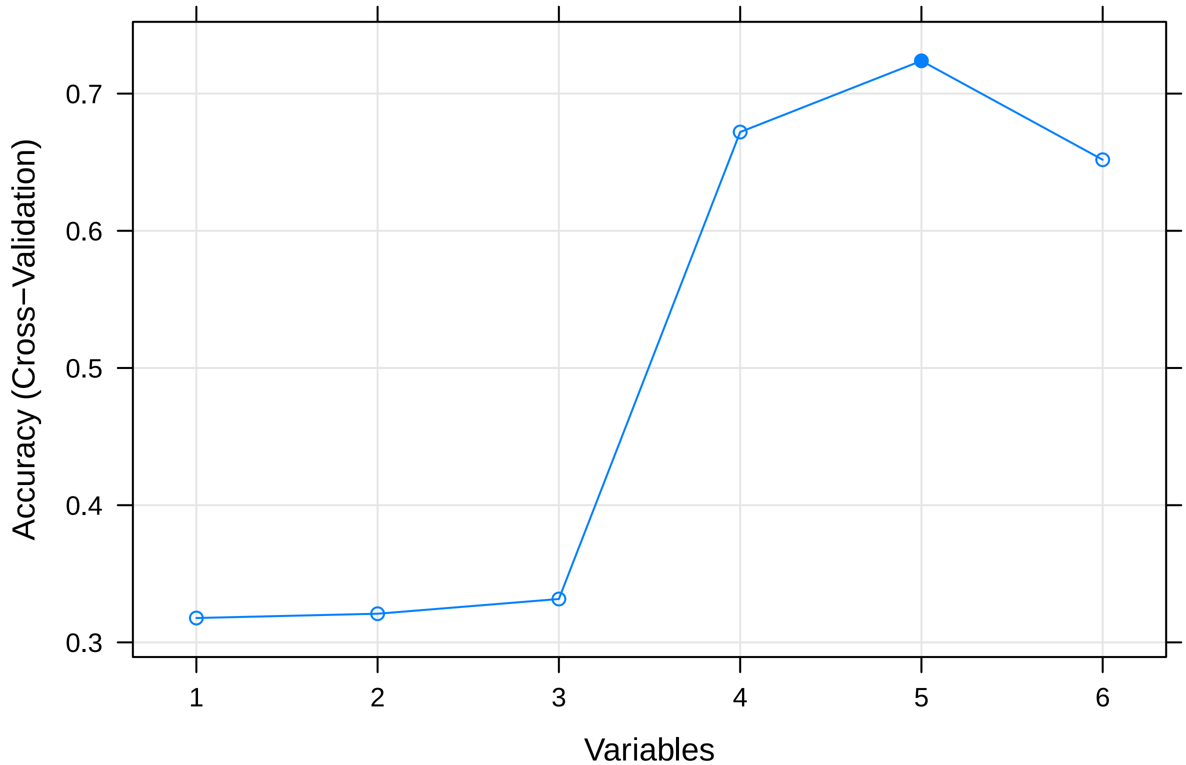
Variables selected
sexid
gdp_per_capita
countryid
year
agegroupid
generationid

Note:
Variables for Best RMSE



Variables selected
sexid
gdp_per_capita
countryid
agegroupid
year

Note:
Variables for Best Accuracy



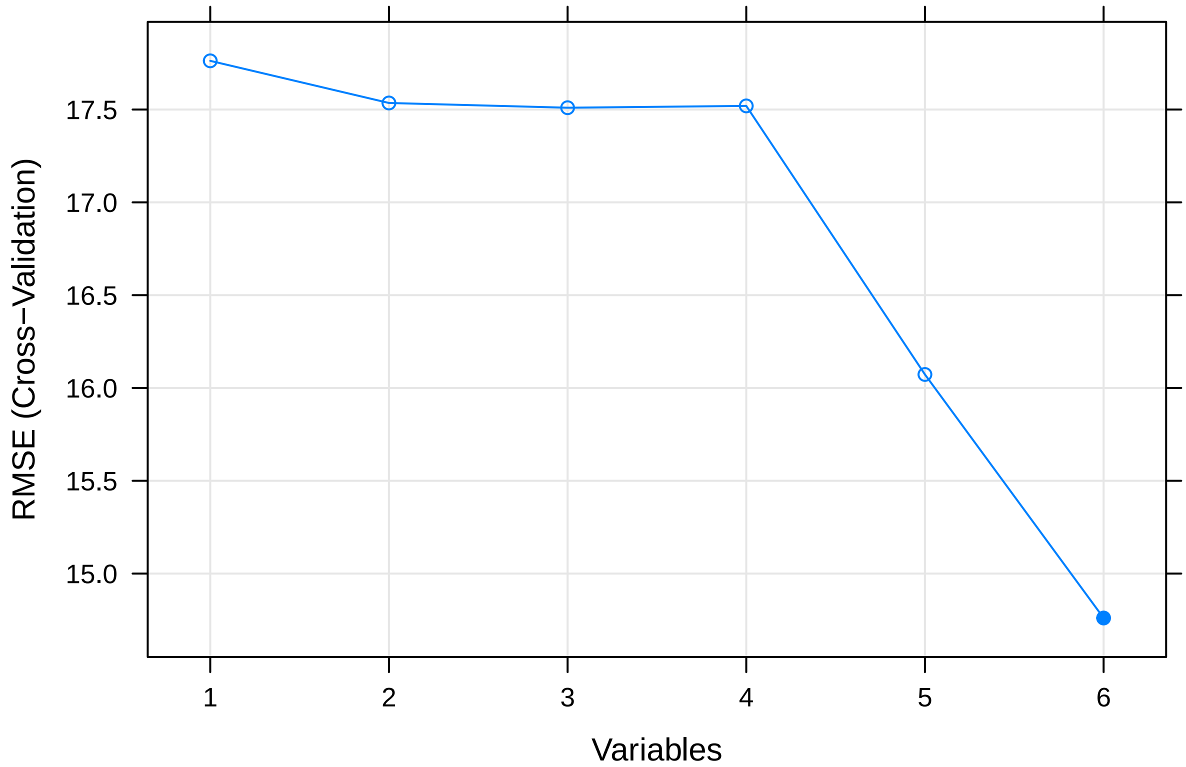
Lowest RMSE of about 11 is achieved with Suicide Rates prediction. Peak Accuracy of little more than 72% is acheived with predicting categorized Suicides_Intensity when we used 5 variables (sexid, gdp_per_capita, countryid, agegroupid, year - excluding Generationid).

Bagged CART

Here we use Bagging Model to check for features prediction accuracies.

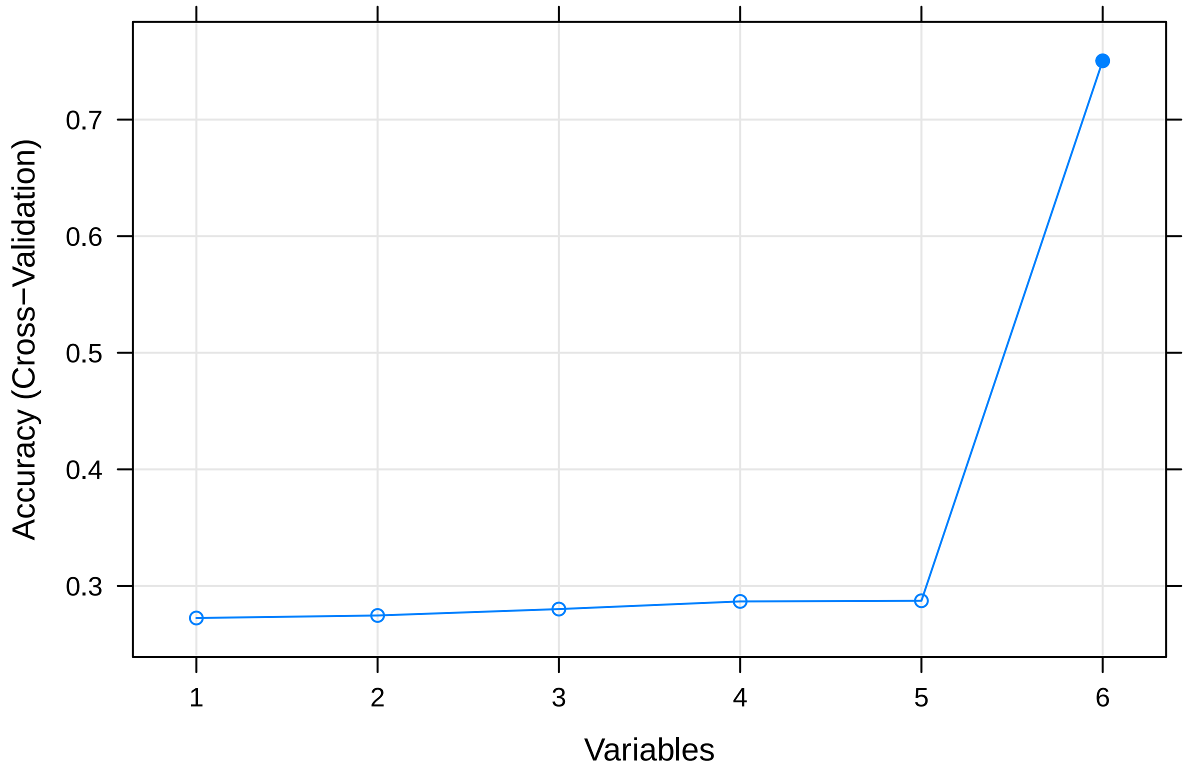
Variables selected
agegroupid
countryid
gdp_per_capita
generationid
year
sexid

Note:
Variables for Best RMSE



Variables selected
gdp_per_capita
year
countryid
agegroupid
generationid
sexid

Note:
Variables for Best Accuracy

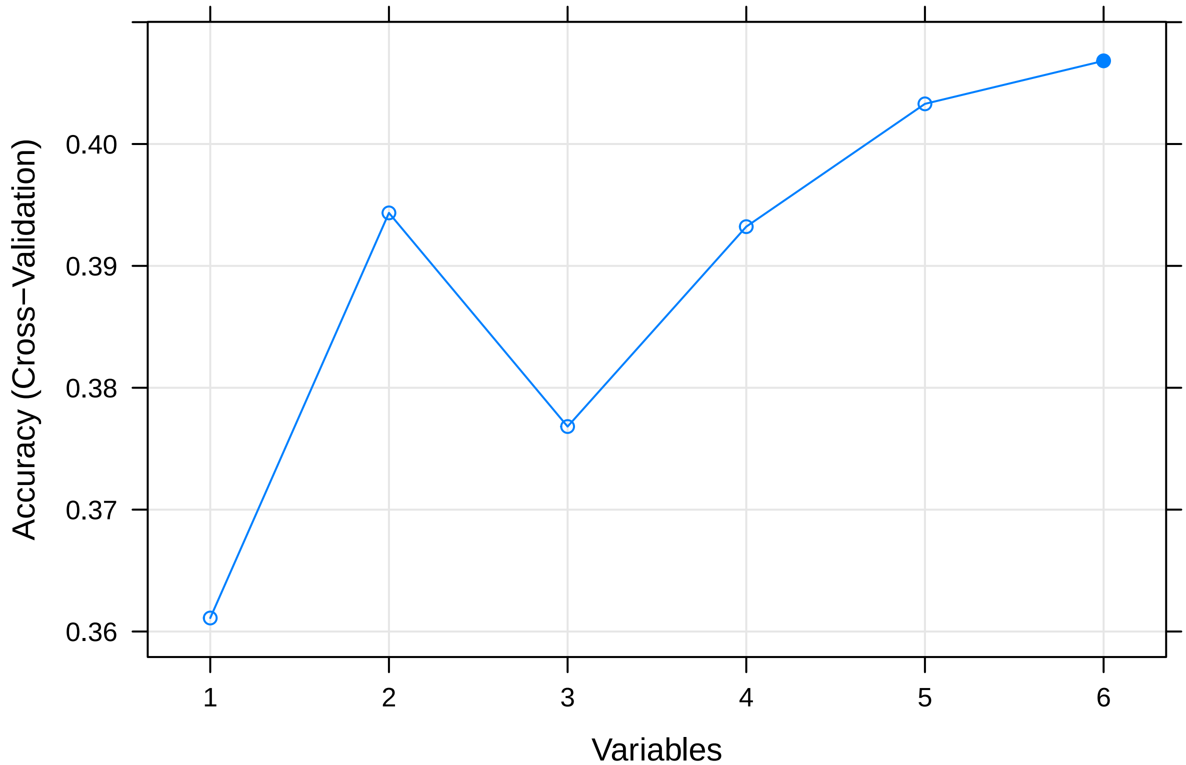


A minimum RMSE of little less than 15 with all 6 variables.An accuracy of about 75% with all 6 variables. RMSE did not fare as well as Bagged CART, but we got better accuracy.

Naive Bayes

We now attempt Naive bayes algorith to estimate the accuracy measures using RFE.

Variables selected
agegroupid
sexid
generationid
gdp_per_capita
countryid
year



A Peak Accuracy of more than 40% is found with all variables.

Add more features

Add more features from world data to suicide data and see if any of those new features help improve our model. We add the following features from world data:
Economy
Continent
Sub-Region
Income Level

Find High correlation instances with added world data

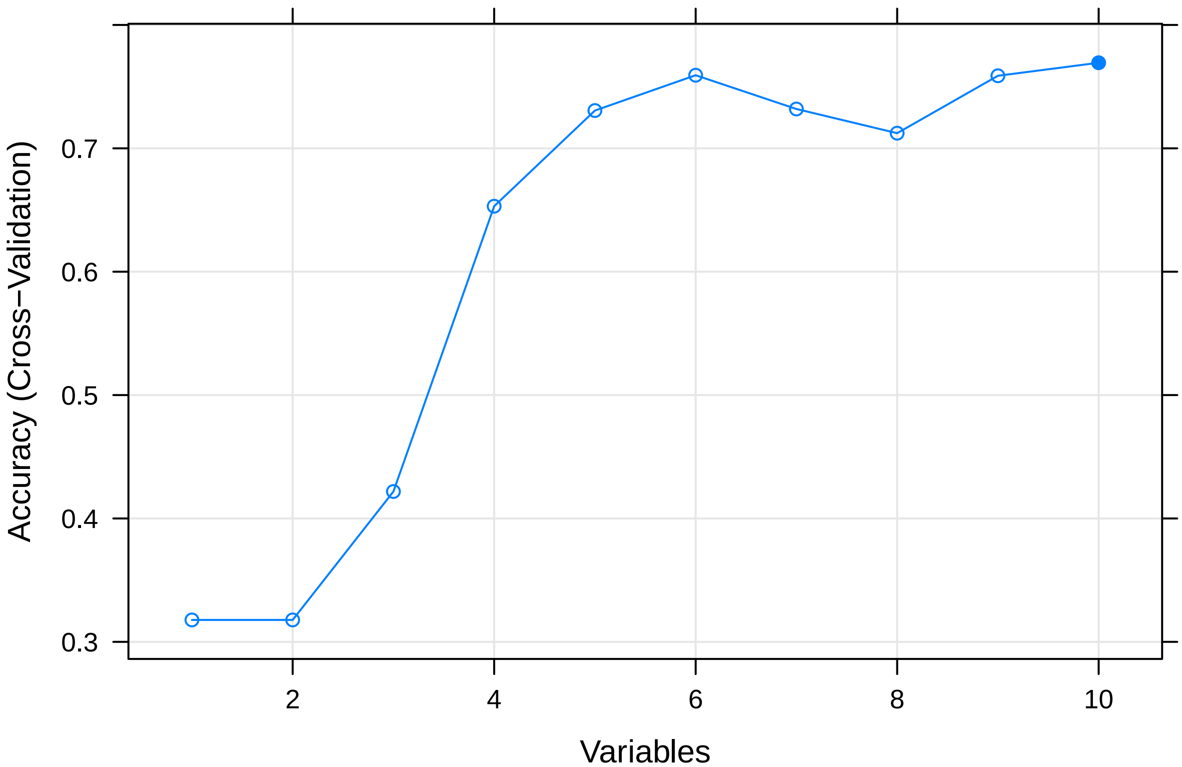
Features Highly Correlated

We dont find any highly correlated data even with added features.

Random Forest - More Variables

We achieved an accuracy of about 72% before. We now perform RFE again with more variables and see if this method can take advantage of additional features available.

Variables selected
sexid
countryid
agegroupid
subregion_id
gdp_per_capita
year
continent_id
ig_id
eco_id
generationid



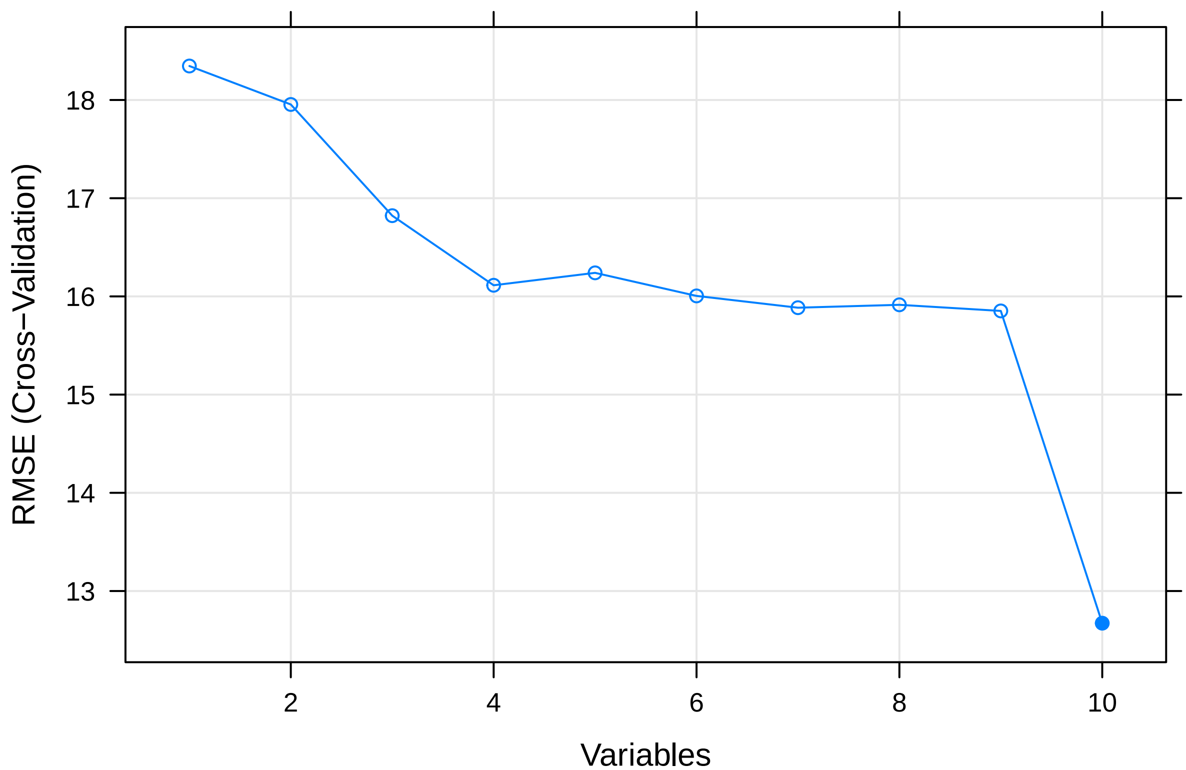
Peak Accuracy of little less than 77% with 10 predictors and about 76% with 6 (sexid, countryid, agegroupid, subregion_id, gdp_per_capita,year). This is the best accuracy we could achieve with about four full percentage points above the earlier estimate. Even with 6 variables we see a better accuracy than before.

Bagged CART - More Variables

We got the best accuracy earlier with 6 variable. But we did not get a good RMSE. We will check if addition of more features would give us a better Accuracy and lower RMSE.

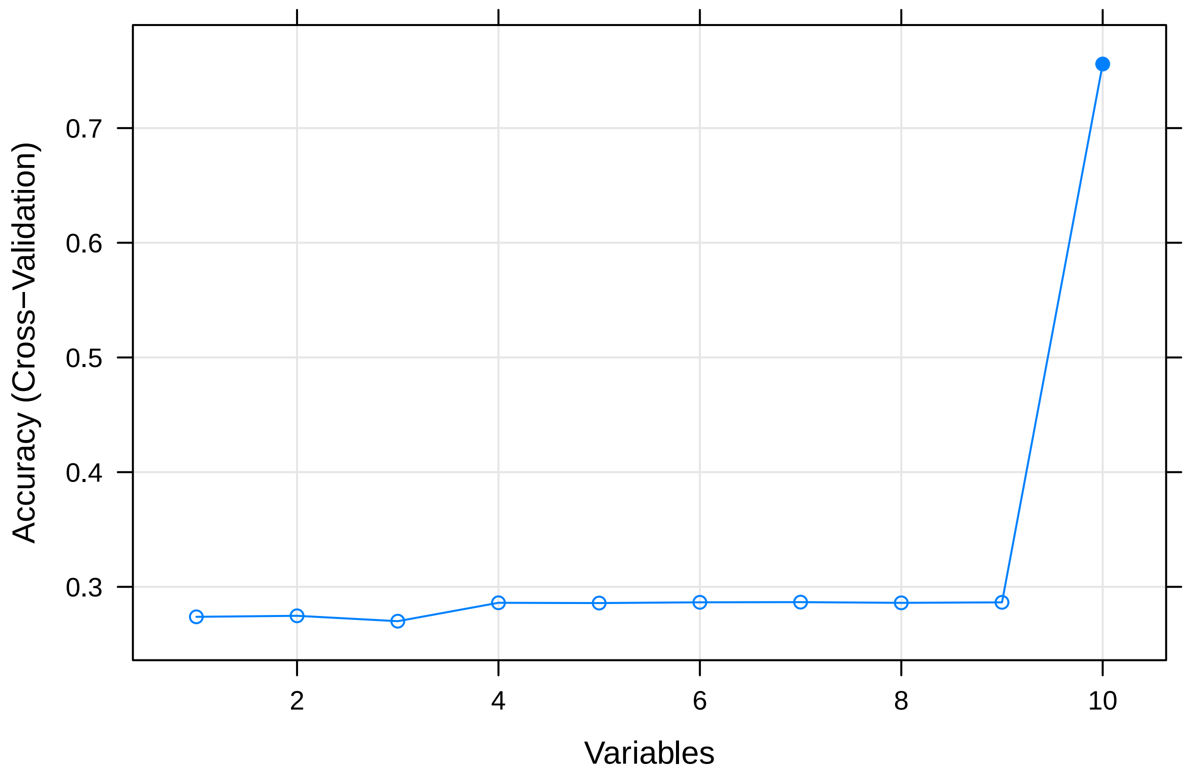
Variables selected
countryid
subregion_id
agegroupid
gdp_per_capita
ig_id
continent_id
eco_id
generationid
year
sexid

Note:
Variables for Best RMSE



Variables selected
gdp_per_capita
year
agegroupid
countryid
subregion_id
generationid
continent_id
eco_id
ig_id
sexid

Note:
Variables for Best Accuracy

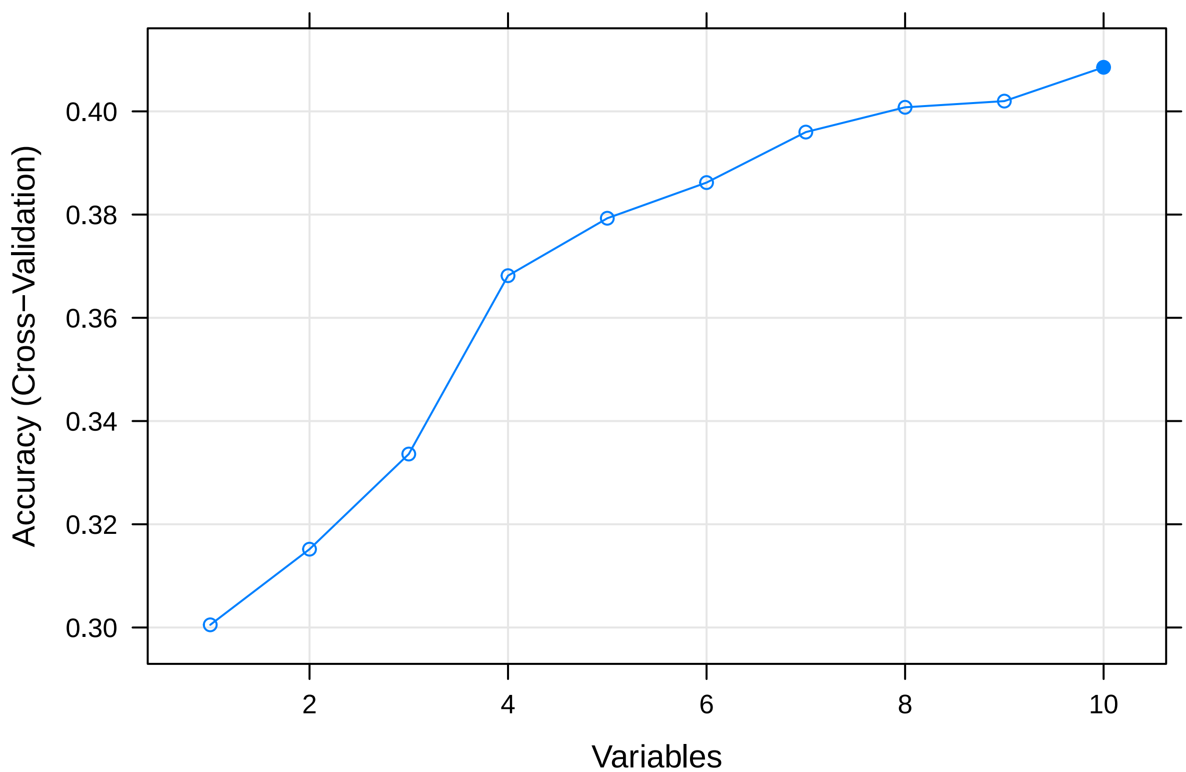


A Peak accuracy of little less than 76% with all 10 variables.This is half point better accuracy than we found with the original 6 variables. Lowest RMSE of little less than 13 is achieved with all 10 variables with Bagged Tree. This is a 2 point improvement from the RMSE we achieved Treegabg before with 6 original variables.

Linear Discriminant Analysis - More Variables

This RFE will use the LDA functions to predict accuracy.

Variables selected
agegroupid
sexid
generationid
ig_id
eco_id
continent_id
subregion_id
gdp_per_capita
countryid
year

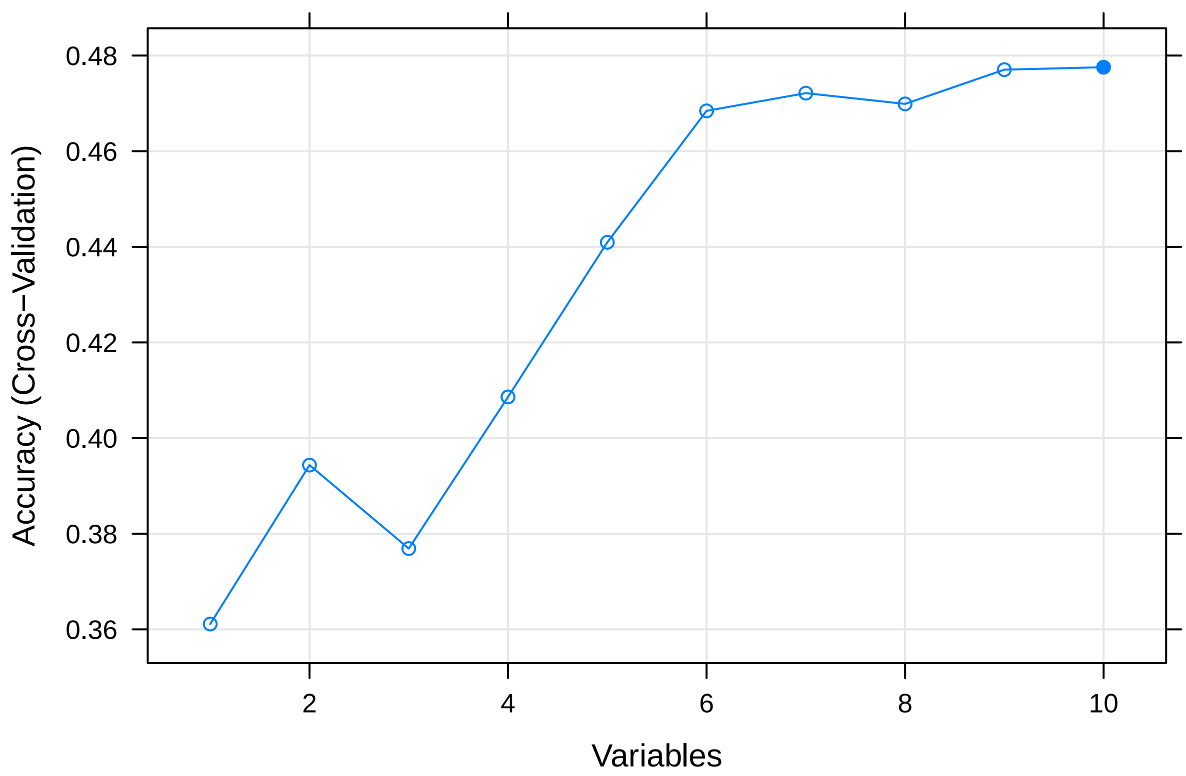


Peak Accuracy of more than 40% is achieved with all 10 variables

Naive Bayes - More Variables

This Classification model would be run with Naive Bayes to find the best accuracy that can be achieved with more features added to the Suicides data.

Variables selected
agegroupid
sexid
generationid
ig_id
eco_id
continent_id
subregion_id
gdp_per_capita
countryid
year

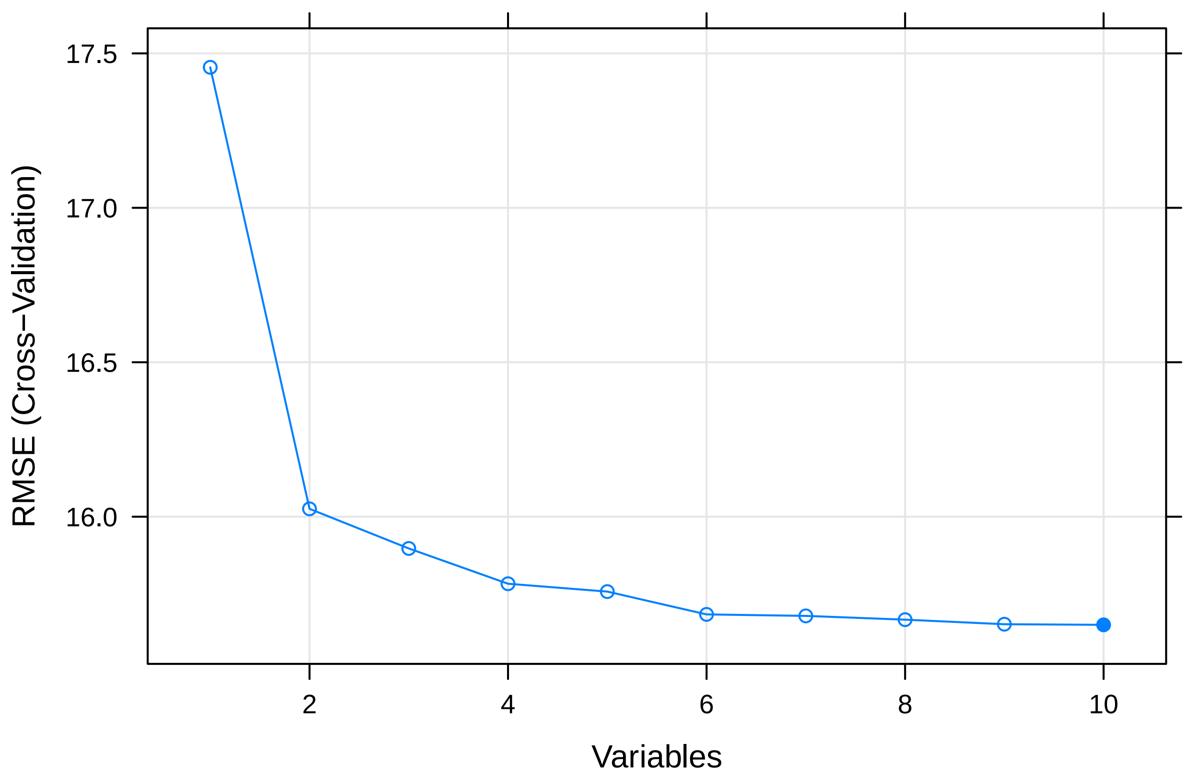


A Peak accuracy of little less than 48% is achieved with 9 variables.

Linear Model

We will use Linear Model for simple regression of the data and look for the best RMSE.

Variables selected
sexid
agegroupid
continent_id
ig_id
eco_id
subregion_id
generationid
year
countryid
gdp_per_capita



Lowest RMSE of more than 17 is found with 10 variables

RFE Results Summary

Table below summarizes the several RMSE and Accuracy estimates we calculated using RFE. We pick a reasonably suitable model for the purpose of training our data.

method	Accu	RMSE	Accu_vars	rmse_vars	Sec_accu	Sec_RMSE	Sec_accu_Vars
Random Forest	72.39	11.23	5	6	67.20	17.31	4
Bagged CART	75.04	14.76	6	6	28.72	17.54	5
Naive Bayes	40.68	NA	6	NA	40.33	NA	5
Random Forest - More Variables	76.93	NA	10	NA	75.92	NA	6
Bagged Tree - More Variables	75.59	12.67	10	10	28.67	17.95	7
LDA - More Variables	40.85	NA	10	NA	40.20	NA	9
Naive Bayes - More Variables	47.76	NA	10	NA	47.70	NA	9
LM - More Variables	NA	17.45	NA	10	NA	16.03	NA

Model Selection

Our model selection criterion would be to select a model that performs reasonably well with lowest possible number of variables. Random Forest models performed well for both kinds of measurements. We can use the Random Forest method with RMSE when we consider suicides rates data as continious data or "Random Forest - More Variables" (Third best Accuracy one with 6 variables) for finding Accuracy with categorized Suicide data.

Create a Classification Model with the 6 predictors from the RFE ressults and predict with test data

We select the features that gave us a relatively good performance with much less number of features. We model based on the data until 2014 and perform a prediction for 2015 data and compare the predicted suicide rates against the actual rates. We use Repeated Cross Validattion for training the model.
Following parameters are used:
metric: Accuracy (By default used for factorised data)
method: rf
number: 10
repeats: 3
ntree: 100
tunelength: 10
search: random

Classification Model With First Set of Tuning Parameters

mtry	Accuracy	Kappa	AccuracySD	KappaSD
1	0.5491272	0.4152249	0.0118556	0.0189855
2	0.7577315	0.6988002	0.0060590	0.0076568
3	0.7674157	0.7116633	0.0052431	0.0064487
4	0.7621907	0.7055025	0.0053031	0.0065289
5	0.7583191	0.7008320	0.0054074	0.0066544

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 100, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 23.3%
## Confusion matrix:
```

```
##       1    2    3    4    5   6   7   8   9 class.error
## 1 7719  598   75   48   28  10   8   4   1   0.0909198
## 2  760 3415  546   42    4   0   0   0   0   0.2836165
## 3  153  576 3251  450   41   4   6   0   0   0.2744923
## 4  109   44  482 2280  377  17   8   1   0   0.3128391
## 5   51    7   45  398 1660 211  34   5   0   0.3114890
## 6   30    0    6   36  279 646 129   9   0   0.4308370
## 7   32    0    6   16   46 146 666  93   3   0.3392857
## 8   18    0    0    4    9   2  92 653  31   0.1928307
## 9    6    0    0    0    0   2   1  61 114   0.3804348
```

Apply the above model to the test data for 2015 and predict the Suicide Intensity

1	2	3	4	5	6	7	8	9
189	24	2	5	0	0	1	0	0
18	108	21	1	0	0	0	0	0
3	16	97	14	2	0	0	0	0
1	0	21	61	6	0	0	0	0
0	1	2	13	40	4	0	0	0
0	0	0	2	13	16	3	0	0
1	0	1	0	2	10	15	1	0
0	0	0	0	0	0	5	12	0
0	0	0	0	0	0	0	0	1

	x
Accuracy	0.7363388
Kappa	0.6741985
AccuracyLower	0.7028249
AccuracyUpper	0.7679351
AccuracyNull	0.2896175
AccuracyPValue	0.0000000
McnemarPValue	NaN

An accuracy of about 73% is achieved, but the model seems to be little overtrained. The variation in accuracies could be because the 2015 data does not include few countries that were part of the trained model. Mtry value of about 3 or 4, which would be about the default value is resulting in better models. We increase the number of trees to 200 and perform training with mtry of 3.46. We expect better accuracy of the model as well as predictions.

Create a Classification Model with 200 number of trees

We select the features that gave us a relatively good performance with much less number of features. We model based on the data until 2014 and perform a prediction for 2015 data and compare the predicted suicide rates against the actual rates. We use Repeated Cross Validattion for training the model.
Following parameters are used:
metric: Accuracy (By default used for factorised data)
method: rf
number: 10
repeats: 3
ntree: 200

tunegrid: 3.46
search: random

Trees 200: Classification Model With 200 trees

mtry	Accuracy	Kappa	AccuracySD	KappaSD
3.464102	0.7686685	0.7131974	0.0079709	0.0099887

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 200, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 22.91%
## Confusion matrix:
##      1    2    3    4    5   6   7   8   9 class.error
## 1 7739  565   77   60   27   5  13   4   1  0.08856436
## 2  758 3421  541   44    2   1   0   0   0  0.28235788
## 3  149  559 3288  431   45   4   5   0   0  0.26623522
## 4  105   42  468 2303  382  13   4   1   0  0.30590717
## 5   52    9   41  388 1679 206  34   2   0  0.30360846
## 6   29    0    7   32  280 642 137   7   1  0.43436123
## 7   32    0    7   17   46 139 672  94   1  0.33333333
## 8   18    0    0    4    7   4  85 655  36  0.19035847
## 9    6    0    0    0    0   1   6  61 110  0.40217391
```

Trees 200: Apply the above model to the test data

We apply the model to the teset data for 2015 and predict the Suicide Intensity.

Trees 200: Confusion Matrix and Accuracy estimates for Predictions

Confusion Matrix for the predictions with test data with NTREE of 200

1	2	3	4	5	6	7	8	9
189	24	2	5	0	0	1	0	0
17	107	22	0	0	0	0	0	0
4	16	96	16	2	0	0	0	0
1	0	21	60	6	0	0	0	0
0	1	2	15	40	4	0	0	0
0	1	0	0	13	16	3	0	0
1	0	1	0	2	10	16	1	0
0	0	0	0	0	0	4	12	0
0	0	0	0	0	0	0	0	1

	x
Accuracy	0.7336066
Kappa	0.6708216
AccuracyLower	0.6999988
AccuracyUpper	0.7653191
AccuracyNull	0.2896175
AccuracyPValue	0.0000000
McnemarPValue	NaN

It is seen that prediction accuracy improved with the model with larger number of trees. We increase the number of trees to 500 and retrain.

Create a Classification Model with 500 number of trees

We select the features that gave us a relatively good performance with much less number of features. We model based on the data until 2014 and perform a prediction for 2015 data and compare the predicted suicide rates against the actual rates. We use Repeated Cross Validattion for training the model.
Following parameters are used:
metric: Accuracy (By default used for factorised data)
method: rf
number: 10
repeats: 3
ntree: 500
tunegrid: 3.46
search: random

Trees 500: Classification Model With 500 trees

mtry	Accuracy	Kappa	AccuracySD	KappaSD
3.464102	0.7701212	0.7149798	0.0076548	0.0095198

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 500, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 22.8%
## Confusion matrix:
##      1    2    3    4    5   6   7   8   9 class.error
## 1 7749  562   73   56   26   9  14   2   0  0.08738664
## 2  737 3442  543   43    2   0   0   0   0  0.27795259
## 3  152  560 3272  445   46   2   4   0   0  0.26980585
## 4  109   40  450 2312  387  14   5   1   0  0.30319470
## 5   55    7   44  382 1687 199  33   4   0  0.30029034
## 6   31    0    7   27  286 638 140   5   1  0.43788546
## 7   34    0    5   13   49 136 675  94   2  0.33035714
## 8   18    0    1    4    7   2  92 653  32  0.19283066
## 9    6    0    0    0    0   1   5  63 109  0.40760870
```

Trees 500: Apply the above model to the test data

We apply the model to the teset data for 2015 and predict the Suicide Intensity.

Trees 500: Confusion Matrix and Accuracy estimates for Predictions

Confusion Matrix for the predictions with test data with NTREE of 200

1	2	3	4	5	6	7	8	9
189	24	2	5	0	0	1	0	0
18	108	20	0	0	0	0	0	0
3	15	98	14	2	0	0	0	0
1	0	21	62	6	0	0	0	0
0	1	2	15	40	4	0	0	0
0	1	0	0	13	16	3	0	0
1	0	1	0	2	10	16	0	0
0	0	0	0	0	0	4	13	0
0	0	0	0	0	0	0	0	1

	x
Accuracy	0.7418033
Kappa	0.6810288
AccuracyLower	0.7084817
AccuracyUpper	0.7731626
AccuracyNull	0.2896175
AccuracyPValue	0.0000000
McnemarPValue	NaN

Prediction accuracy improved further with the model with 500 number of trees.

Final Classification Model

RANDOM FOREST Number of Classes: 9 Classes (Suicide Rate Range-Value): 0-2: 1, 2-6: 2, 6-12: 3, 12-20: 4, 20-30: 5, 30-40: 6, 40-60: 7, 60-100: 8, 100+: 9
number: 10
repeats: 3
ntree: 500
tunegrid: 3.46
search: random
Features Used: sexid, countryid, agegroupid, subregion_id,gdp_per_capita,year

Summary of Accuracies with predictions on Test data with Random Forest

We find the accuracies with the predictions on test data across different features. Following tables present a overall breakup of accuracies across different feature. These may provide some areas of focus for achieving better accuracies.

country	correct	incorrect	accuracy
Antigua and Barbuda	11	1	91.67
Argentina	11	1	91.67
Armenia	9	3	75.00
Australia	11	1	91.67
Austria	11	1	91.67
Belgium	11	1	91.67
Belize	5	7	41.67
Brazil	12	0	100.00
Chile	11	1	91.67
Colombia	12	0	100.00
Croatia	10	2	83.33
Cuba	10	2	83.33
Cyprus	7	5	58.33
Czech Republic	8	4	66.67
Denmark	7	5	58.33
Ecuador	4	8	33.33
Estonia	4	8	33.33
Finland	5	7	41.67
Georgia	7	5	58.33
Germany	10	2	83.33

sex	correct	incorrect	accuracy
Female	286	80	78.14
Male	257	109	70.22

age	correct	incorrect	accuracy
05-14 years	115	7	94.26
35-54 years	88	34	72.13
15-24 years	86	36	70.49
75+ years	86	36	70.49
25-34 years	85	37	69.67
55-74 years	83	39	68.03

economy	correct	incorrect	accuracy
Emerging region: G20	64	8	88.89
Emerging region: MIKT	32	4	88.89
Developed region: G7	52	8	86.67
Developed region: nonG7	204	84	70.83
Developing region	175	77	69.44
Emerging region: BRIC	16	8	66.67

income_grp	correct	incorrect	accuracy
Low income	10	2	83.33
Upper middle income	217	59	78.62
High income: OECD	222	78	74.00
High income: nonOECD	48	24	66.67
Lower middle income	46	26	63.89

continent	correct	incorrect	accuracy
Oceania	11	1	91.67
Africa	10	2	83.33
South America	60	12	83.33
North America	92	28	76.67
Asia	117	39	75.00
Europe	239	97	71.13
Seven seas (open ocean)	14	10	58.33

subregion	correct	incorrect	accuracy
Northern America	12	0	100.00
Australia and New Zealand	11	1	91.67
Caribbean	42	6	87.50
South America	60	12	83.33
Southern Africa	10	2	83.33
Western Europe	59	13	81.94
Eastern Asia	19	5	79.17
South-Eastern Asia	19	5	79.17
Southern Europe	64	20	76.19
Central Asia	27	9	75.00
Western Asia	52	20	72.22
Eastern Europe	50	22	69.44
Central America	38	22	63.33
Northern Europe	66	42	61.11
Eastern Africa	14	10	58.33

Regression Model for minimizing RMSE

We achieved an RMSE of about 11 before. We used simple Cross Validation at 10 during the RFE. Now we use Repeated Cross Validation with 3 repeats. As the training is slow as a results of the number of variables, we limit the number of trees to 100.

mtry	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
5	2.932627	0.9764452	1.507261	0.2533045	0.0035213	0.0272713
6	2.845286	0.9777833	1.408304	0.2494705	0.0034951	0.0267397
7	2.815827	0.9782582	1.360707	0.2389870	0.0032953	0.0242436
8	2.789747	0.9786438	1.336019	0.2457388	0.0033863	0.0243350
10	2.776270	0.9788365	1.318164	0.2507491	0.0034905	0.0216008
11	2.773446	0.9788775	1.314779	0.2505062	0.0034481	0.0207809
13	2.785618	0.9787234	1.312521	0.2286108	0.0031623	0.0187681
14	2.791271	0.9786142	1.313047	0.2408002	0.0033963	0.0201791

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 100, mtry = param$mtry)
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 11
##
##           Mean of squared residuals: 7.409882
##                     % Var explained: 97.97
```

Method
rf

Apply the above model to the test data for 2015 and predict the Suicide Rate

There is a dramatic improvement in the RMSE with mtry values of around 10 with 100 trees. We achieved the best RMSE of about 3.

FINAL REGRESSION MODEL

RANDOM FOREST
number: 5
repeats: 3
ntree: 100
mtry: 10
search: random
Features Used: sexid, gdp_per_capita, countryid, year, agegroupid, generationid

RMSE across different features in the test data.

Following tables summarize the overall RMSE on the test data when the model is applied and the RMSEs we achieved across different features.

country	rmse
Antigua and Barbuda	0.4830387
Argentina	0.8284049
Armenia	1.4090103
Australia	1.7227039
Austria	1.4151343
Belgium	2.0105683
Belize	2.7324786
Brazil	0.9003173
Chile	2.2253394
Colombia	0.5516278

sex	rmse
Female	1.029664
Male	2.598696

age	rmse
05-14 years	0.3563526
15-24 years	1.3761967
25-34 years	1.9095482
35-54 years	2.1154667
55-74 years	1.7706622
75+ years	3.1878788

generation	rmse
Boomers	1.7706622
Generation X	2.1154667
Generation Z	0.3563526
Millenials	1.6643755
Silent	3.1878788

Conclusion

World suicides data is used to prepare regression as well as classification models and estimate RMSE and Accuracies. Several methods are applied to the training data and the model best suited for the data is selected using Recursive Feature Elimination process. Attempt is made to enhance the data by adding more features in order to peform better analysis. RFE helps to minimise the number of features used in the prediction algorithm to achieve a reasonably better performing model with lower number of features. It is found that suicide rates among male are three times more than women and they rise with age. Regional variations in suicide rates across the world are shown. These could be attributed to different socia-economic factors in those regions. We could make major improvemnts in the RMSE and Accuracies using the Random Forest models using Repeated Cross Validation and selected the best models to perform predictions on the world's suicide data.