Contents

EXECUTIVE SUMMARY

This Project Was Taken To Build Deep Understanding of the Topic 'Suicide Rates' By-

- Figuring out The Variables From The Components Of Problem Statement
- Perform Panel Data Regression
- Find if the country's GDP per capita has any Influence in Suicide Death Rates

BACKGROUND

- This dataset contains information on suicide rates for years 1991, 2001 and 2011 for 40 countries in the world.
- Over the years studies have been conducted for exploring the possible causes that might increase the risk of suicide in societies.
- Therefore, deciding to working on this data set is our way of understanding this problem and finding out the relationship between socio economic factors and suicide rates.

PROBLEM STATEMENT

The objective of the project is to study the dataset, extract information about suicide rates by studying the data of suicide rate for different countries. Analyzing data, performing panel regression on the data, so as to find out the relationship between country's GDP per capita and suicide rates.

METHODOLOGY

In order to understand the relationship between socio economic factors and suicide rates, with the help of the dataset, we will have to take the following steps:

1. Import data from dataset and perform initial high-level analysis: look at the number of rows, look at the missing values, look at dataset columns and their values respective to the campaign outcome.
2. Clean the data: remove irrelevant columns, deal with missing and incorrect values, and turn categorical columns into dummy variables.
3. Performing panel data regression.

DATA SOURCE

United Nations Development Program. (2018). Human development index (HDI).
Retrieved from http://hdr.undp.org/en/indicators/137506
World Bank. (2018). World development indicators: GDP (current US$) by country: 1985 to 2016.
Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

	Country	Year	gdppercapita	suicidesrate	Female	Male	SEX Ratio
0	Argentina	1991	6404	9.940000	15133000	14490000	0.957510
1	Argentina	2001	7900	11.960000	17438298	16572648	0.950359
2	Argentina	2011	13946	8.820000	19491572	18524167	0.950368
3	Austria	1991	23808	29.620000	3812047	3487681	0.914910
4	Austria	2001	25848	21.830000	3949662	3684898	0.932965
...	...	...	...	...	...	...	...
112	Sweden	2001	28429	13.957500	4273200	4166117	0.974941
113	Sweden	2011	63380	11.814167	4466379	4418265	0.989228
114	Turkmenistan	1991	1014	10.578333	1610400	1552100	0.963798
115	Turkmenistan	2001	863	11.121667	2086627	2009354	0.962968
116	Turkmenistan	2011	6319	2.635833	2358391	2267704	0.961547

Variables

Dependent variable (Y):

- **Suicide Rate**- It shows the number of deaths per 100k people in a country for a given year, Numerical

Independent Variables (X) :

- **Country -** Name of the country, Categorical
- **Year -** Data of which year, Categorical
- **gdppercapita -** Shows gdp per capita for a country in a particular year, Numerical
- **Female -** Female population for a country in a particular year, Numerical
- **Male -** Male population for a country in a particular year, Numerical

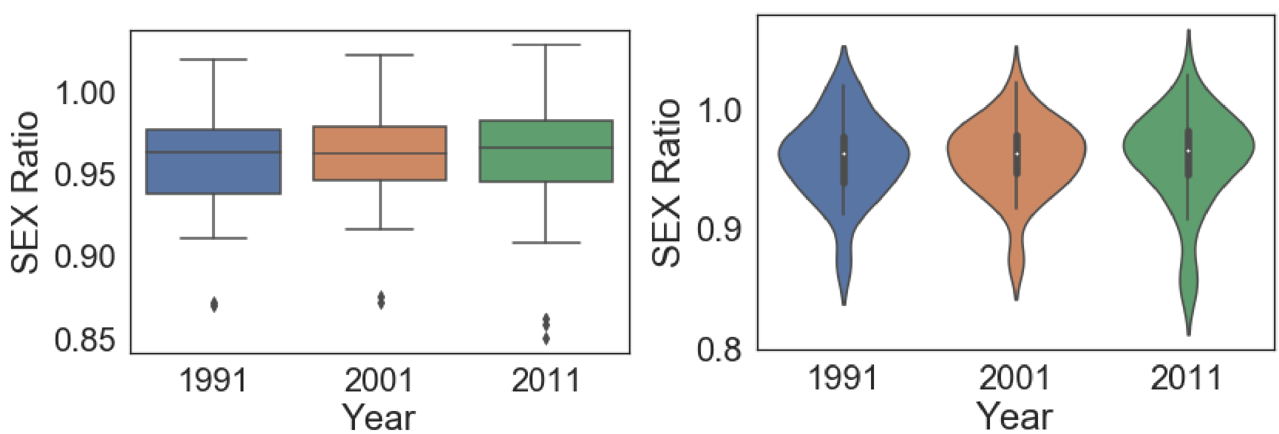
Descriptive Statistics

Descriptive statistics provide simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries form the basis of the initial description of the data.
Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables.

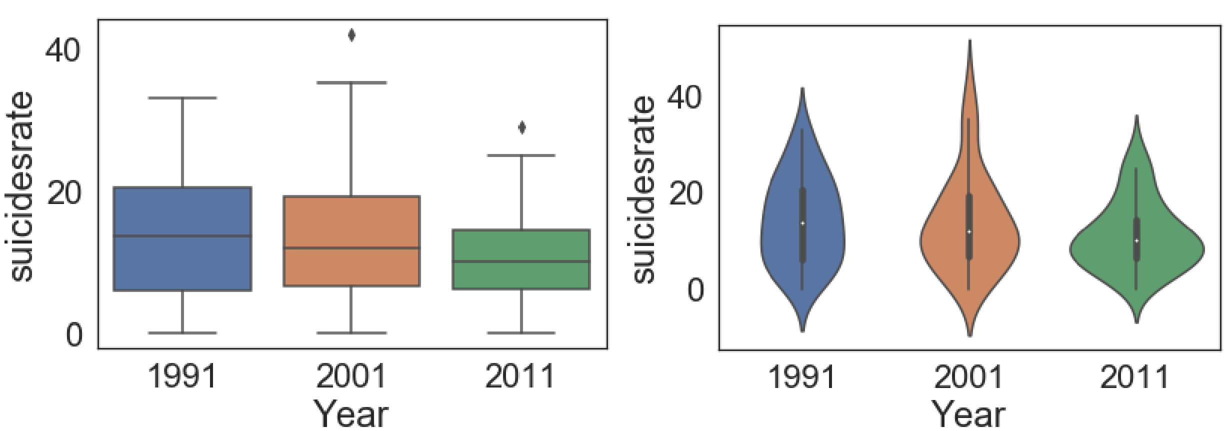
	Year	gdppercapita	suicidesrate	Female	Male	SEX Ratio
count	117.000000	117.000000	117.000000	1.170000e+02	1.170000e+02	117.000000
mean	2001.000000	19236.282051	12.973084	1.704315e+07	1.615862e+07	0.958968
std	8.200084	18798.422764	8.770762	2.802731e+07	2.660040e+07	0.034943
min	1991.000000	514.000000	0.000000	4.200000e+04	4.040000e+04	0.849769
25%	1991.000000	4204.000000	6.335833	1.801283e+06	1.772000e+06	0.942540
50%	2001.000000	13946.000000	11.450000	4.814895e+06	4.418265e+06	0.964046
75%	2011.000000	27214.000000	18.410000	1.978911e+07	1.892439e+07	0.980394
max	2011.000000	107430.000000	42.054167	1.472364e+08	1.430774e+08	1.028954

Data Visualization

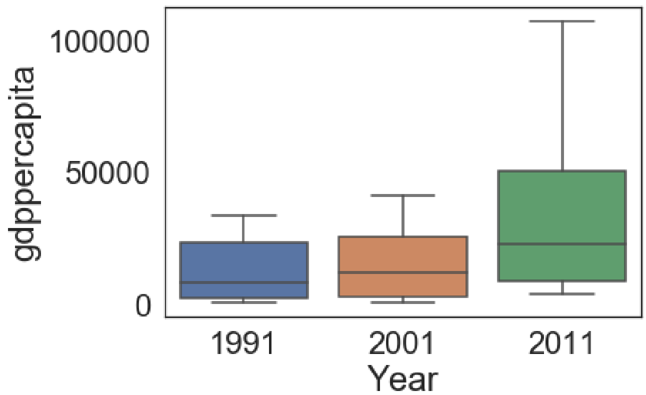
Sex ratio variation across 1991, 2001, 2011

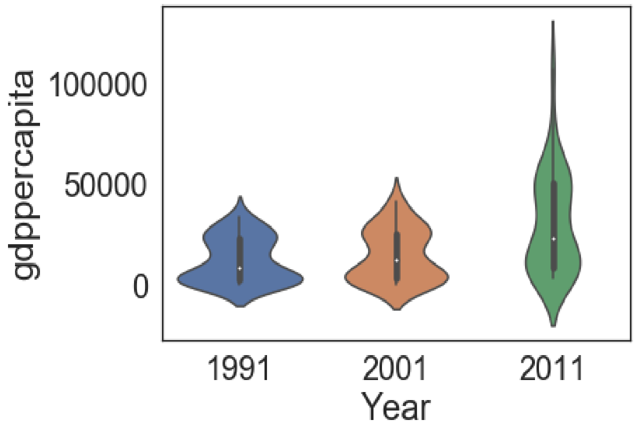


Suicide rate variation across 1991, 2001, 2011

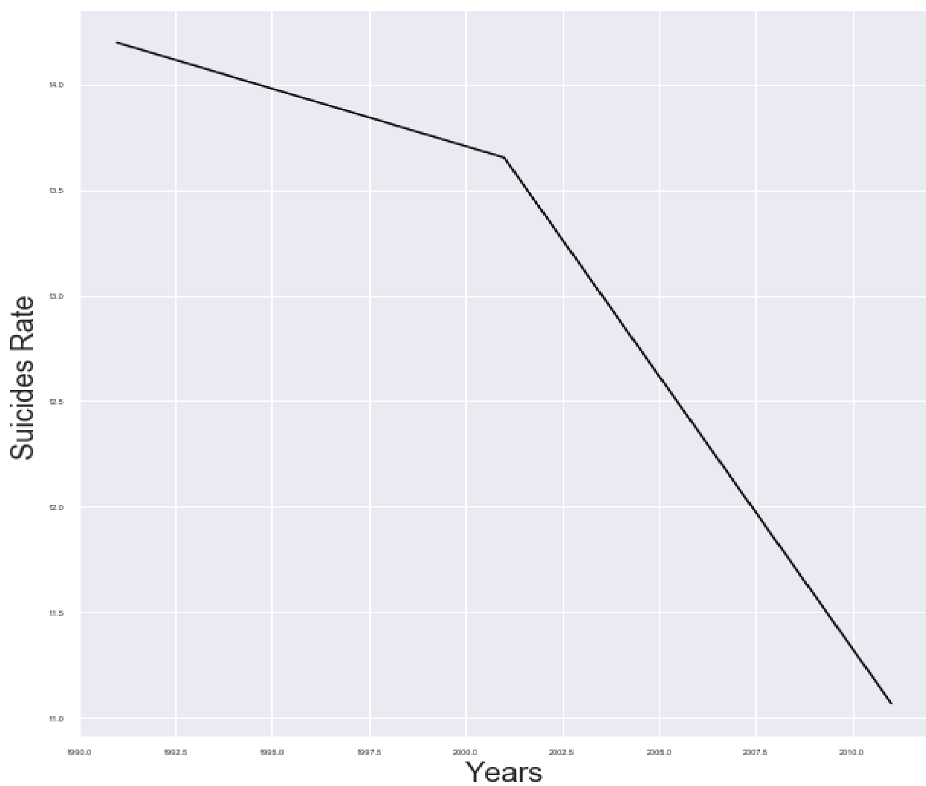


Suicide rate variation across 1991, 2001, 2011





Mean suicide rate



Additional Variables:

Variables representing suicides were having multiple bifurcations based on the age group and the Gender. So in order to attribute the useful data to our entity which is the country we derived two additional variables representing the entity.

1) Sex Ratio
2) Median Age

Sex Ratio:

The sex ratio is the ratio of males to females in a population. In most countries the ratio tends to be 1:1

Median Age:

Median age is the age that divides a population into two numerically equally sized groups - that is, half the people are younger than this age and half are older. It is a single index that summarizes the age distribution of a population

Summary of Models

We fitted regression model between a dependent variable (Suicide rate) and independent variable (GDP per capita). We have the data for 3 years 1991, 2001 and 2011 and we tried to understand the relationship using the below approaches

1) Pooling Model (Ordinary least squares)

2) Before and After (Change) Regression (2011 -1991)

3) Fixed Effects Regression

Pooling Model (Ordinary least squares):

Pooling data refers to two or more independent data sets of the same type. This is nothing but OLS regression applied to the panel data neglecting panel effects.

```
Pooling Model

Call:
plm(formula = suicidesrate ~ gdppercapita, data = df, model = "pooling",
    index = c("Country", "Year"))

Balanced Panel: n = 39, T = 3, N = 117

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-12.9289  -6.4083  -1.8664   5.0010  29.5742

Coefficients:
               Estimate Std. Error t-value Pr(>|t|)
(Intercept)  1.2415e+01 1.1655e+00  10.652   <2e-16 ***
gdppercapita 2.9006e-05 4.3424e-05   0.668   0.5055
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    8923.6
Residual Sum of Squares: 8889.2
R-Squared:      0.0038647
Adj. R-Squared: -0.0047973
F-statistic: 0.446169 on 1 and 115 DF, p-value: 0.5055
>
```

The really low R squared value implies that there are significant amount of panel effects which we did not consider. Hence Pooling model is inappropriate for our given panel data.

Change Regression:

When data for each state are obtained for T = 2 time periods, it is possible to compare values of the dependent variable in the second period to values in the first period. By focusing on changes in the dependent variable, this “before and after” or “differences” comparison, in effect, holds constant the unobserved factors that differ from one state to the next but do not change over time within the countries

[8]:

	Country	gdp_per_capita ($)_1991	suicides/100k pop_1991	gdp_per_capita ($)_2011	suicides/100k pop_2011	Change in gdp_per_capita(2011 -1991)	Change in Suicide rate(2011 -1991)
0	Argentina	6404	9.936667	13946	8.822500	7542	-1.114167
1	Austria	23808	29.622500	53923	17.624167	30115	-11.998333
2	Barbados	8469	6.410833	17708	0.000000	9239	-6.410833
3	Belgium	22523	22.001667	50893	18.892500	28370	-3.109167
4	Brazil	4490	5.695000	14245	6.034167	9755	0.339167

Two additional variables showing the changes from 1991 and 2011 are added to the list of independent variables and those variables are considered in this analysis

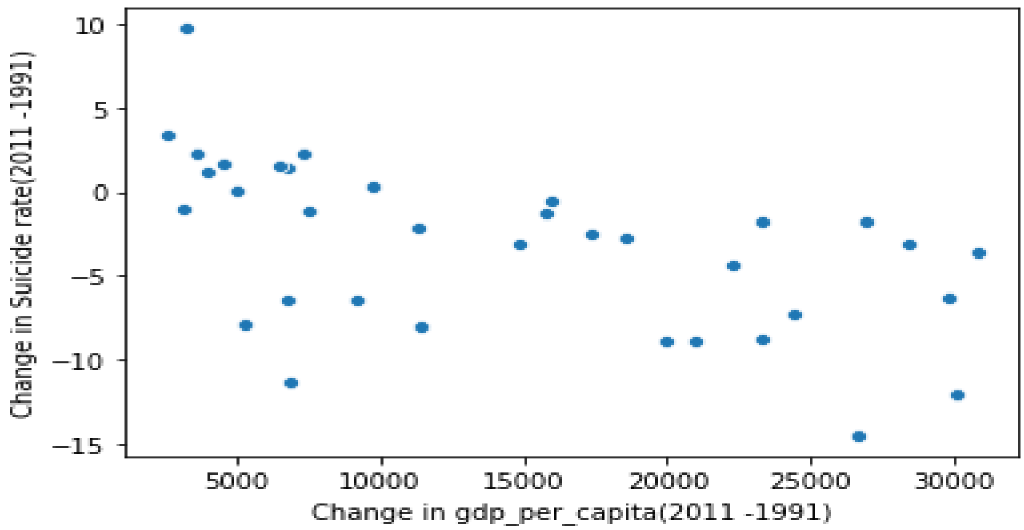
Suicide Rate 1991 = b0 + b1 (GDP percapita 1991) + b2 Z + error 1991

Suicide Rate 2011 = b0 + b1 (GDP percapita 2011) + b2 Z + error 2011

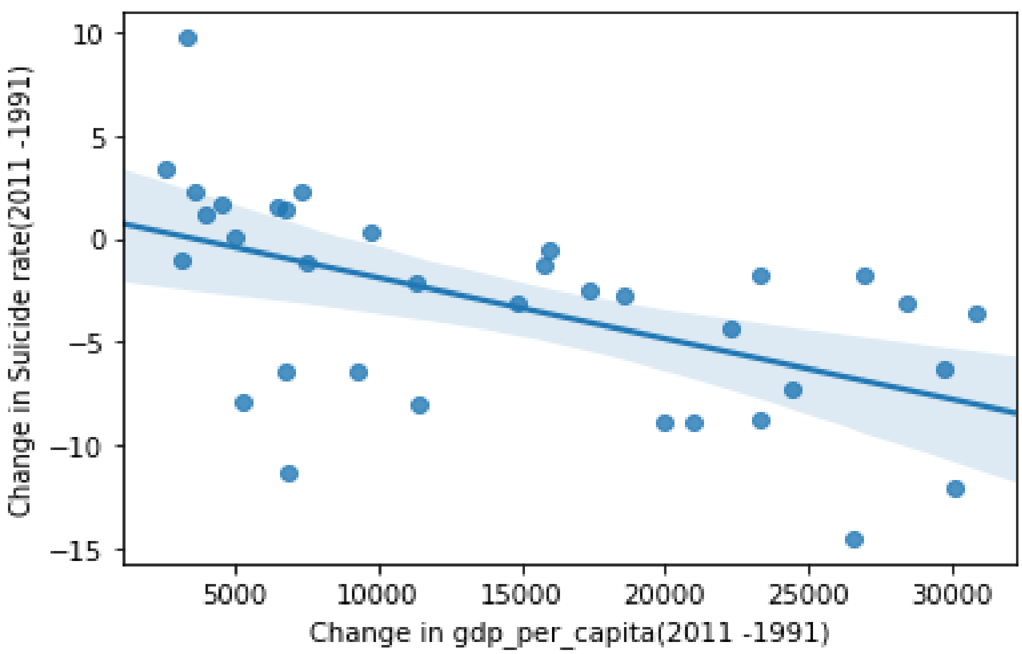
Subtracting first equation from second equation eliminates the effect of Z:

Suicide Rate 2011 - Suicide Rate 1991 = b1 (GDP percapita 2011 - GDP percapita 1991)) + error

The scatter plot for the above regression is given below:



The Regression plot for the above regression is given below:



This is a scatterplot of the change in the suicide rate and the change in the GDP per capita between 1991 and 2011 for 40 countries. There is a negative relationship between changes in the suicide rate and the change in the GDP per capita

Change in suicide Rate (1991 -2011) =1.0671 – 0.0003 (Change in GDP_per_capita (2011- 1991))

Dep. Variable:	Change in Suicide rate(2011 -1991)	R-squared:	0.295
Model:	OLS	Adj. R-squared:	0.274
Method:	Least Squares	F-statistic:	13.84
Date:	Wed, 10 Jun 2020	Prob (F-statistic):	0.000740
Time:	14:38:53	Log-Likelihood:	-99.754
No. Observations:	35	AIC:	203.5
Df Residuals:	33	BIC:	206.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0671	1.354	0.788	0.436	-1.687	3.822
Change in gdp_per_capita(2011 -1991)	-0.0003	7.92e-05	-3.720	0.001	-0.000	-0.000

Omnibus:	1.889	Durbin-Watson:	2.128
Prob(Omnibus):	0.389	Jarque-Bera (JB):	1.325
Skew:	-0.477	Prob(JB):	0.516
Kurtosis:	3.001	Cond. No.	3.18e+04

Hausman Test:

In panel data analysis (the analysis of data over time), the Hausman test can help you to choose between fixed effects model and a random effects model. The null hypothesis is that the preferred model is random effects; the alternate hypothesis is that the model is fixed effects. Essentially, the tests looks to see if there is a correlation between the unique errors and the regressors in the model. The null hypothesis is that there is no correlation between the two.

The result of the Hausman test shows that the Fixed Regression model would be appropriate for the given dataset.

```
> phtest(reg2,reg3)

        Hausman Test

data:  suicidesrate ~ gdppercapita
chisq = 12.211, df = 1, p-value = 0.0004751
alternative hypothesis: one model is inconsistent
```

Test for Heterskedasticity:

The null hypothesis for the Breusch-Pagan test is homoscedasticity.

```
        Breusch-Pagan test

data:  reg3
BP = 13.599, df = 1, p-value = 0.0002263
```

The BP test shows the presence of heteroskedasticity which implies we have to use some kind of robust standard errors to minimize this and the same has been followed.

Test for panel Effect:

The LM test helps you decide between a fixed effects regression and a simple OLS regression. The null hypothesis in the LM test is that variances across entities is zero. This is, no significant difference across units (i.e. no panel effect).

```
        Lagrange Multiplier Test - time effects (Breusch-Pagan) for balanced
        panels

data:  suicidesrate ~ gdppercapita
chisq = 0.14636, df = 1, p-value = 0.702
alternative hypothesis: significant effects
```

Fixed Effect Regression:

This “before and after” or “differences” analysis works when the data are observed in two different years. Our data set, however, contains observations for three different years but the “before and after” method does not apply directly when T is greater than 2. To analyze all the observations in our panel data set, we use the method of fixed effects regression.

Fixed effects regression is a method for controlling for omitted variables in panel data when the omitted variables vary across entities (countries) but do not change over time. Unlike the “before and after” comparisons, fixed effects regression can be used when there are two or more time observations for each entity.

A fixed effects regression is an estimation technique employed in a panel data setting that allows one to control for time-invariant unobserved individual characteristics that can be correlated with the observed independent variables. It involves of subtracting the time mean from each variable in the model and then estimating the resulting transformed model by Ordinary Least Squares. This procedure, known as “within” transformation

Entity Fixed Effect:

It involves of subtracting the time mean from each variable in the model and then estimating the resulting transformed model by Ordinary Least Squares. This procedure, known as “within” transformation. In this data we have data of 40 countries so the mean for three years across all entities and variables would be calculated and subtracted from all the data points. Then Ordinary least squares is applied. The resulting linear equation is given below.

Suicides Rate = - 0.0001 (Gdppercapita) + Entity Fixed Effects

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:          suicidesrate   R-squared:                        0.2154
Estimator:                  PanelOLS   R-squared (Between):             -0.3499
No. Observations:                117   R-squared (Within):               0.2154
Date:               Wed, Jun 10 2020   R-squared (Overall):             -0.3295
Time:                       14:32:23   Log-likelihood                   -279.30
Cov. Estimator:           Unadjusted
                                       F-statistic:                      21.135
Entities:                         39   P-value                           0.0000
Avg Obs:                      3.0000   Distribution:                    F(1,77)
Min Obs:                      3.0000
Max Obs:                      3.0000   F-statistic (robust):             21.135
                                       P-value                           0.0000
Time periods:                      3   Distribution:                    F(1,77)
Avg Obs:                      39.000
Min Obs:                      39.000
Max Obs:                      39.000

                              Parameter Estimates
================================================================================
              Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
gdppercapita    -0.0001  2.855e-05    -4.5973     0.0000     -0.0002   -7.44e-05
================================================================================

F-test for Poolability: 20.175
P-value: 0.0000
Distribution: F(38,77)

Included effects: Entity
```

<u>With N-1 Entity Regressors:</u>

```
factor(Country)Czech Republic        2.086e+01  1.900e+00  10.980  < 2e-16 ***
factor(Country)Ecuador               6.432e+00  1.876e+00   3.428 0.000981 ***
factor(Country)Finland               2.826e+01  2.135e+00  13.241  < 2e-16 ***
factor(Country)France                2.505e+01  2.088e+00  11.994  < 2e-16 ***
factor(Country)Germany               2.084e+01  2.091e+00   9.966 1.68e-15 ***
factor(Country)Greece                6.163e+00  1.935e+00   3.185 0.002091 **
factor(Country)Grenada               1.343e+00  1.882e+00   0.714 0.477583
factor(Country)Guyana                2.699e+01  1.875e+00  14.396  < 2e-16 ***
factor(Country)Iceland               1.745e+01  2.145e+00   8.138 5.52e-12 ***
factor(Country)Ireland               1.599e+01  2.113e+00   7.567 6.91e-11 ***
factor(Country)Israel                1.260e+01  2.007e+00   6.281 1.84e-08 ***
factor(Country)Italy                 1.197e+01  2.039e+00   5.868 1.05e-07 ***
factor(Country)Jamaica               1.370e+00  1.877e+00   0.730 0.467678
factor(Country)Japan                 2.608e+01  2.182e+00  11.954  < 2e-16 ***
factor(Country)Kazakhstan            2.946e+01  1.880e+00  15.665  < 2e-16 ***
factor(Country)Mauritius             1.293e+01  1.881e+00   6.875 1.43e-09 ***
factor(Country)Mexico                5.930e+00  1.887e+00   3.142 0.002381 **
factor(Country)Norway                2.161e+01  2.535e+00   8.526 9.83e-13 ***
factor(Country)Paraguay              4.559e+00  1.876e+00   2.430 0.017410 *
factor(Country)Portugal              1.338e+01  1.925e+00   6.948 1.04e-09 ***
factor(Country)Russian Federation    3.428e+01  1.885e+00  18.187  < 2e-16 ***
factor(Country)Singapore             1.967e+01  2.211e+00   8.899 1.88e-13 ***
factor(Country)Spain                 1.185e+01  1.976e+00   5.996 6.15e-08 ***
factor(Country)Sweden                2.012e+01  2.222e+00   9.056 9.37e-14 ***
factor(Country)Trinidad and Tobago   1.526e+01  1.900e+00   8.030 8.89e-12 ***
factor(Country)Turkmenistan          8.471e+00  1.876e+00   4.516 2.24e-05 ***
factor(Country)United Kingdom        1.143e+01  2.080e+00   5.497 4.83e-07 ***
factor(Country)United States         1.898e+01  2.194e+00   8.649 5.69e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.246 on 77 degrees of freedom
Multiple R-squared:  0.9716,    Adjusted R-squared:  0.9569
F-statistic: 65.97 on 40 and 77 DF,  p-value: < 2.2e-16
```

In this model instead of considering Entity effects we created N-1 binary regressors for 40 entities and those were added in the model to understand the significance in each countries. we notice that most of the variables are in significant in this approach but the R squared value is considerably high.

Time Fixed Effect:

Time fixed effects model assist in controlling for omitted variable bias due to unobserved heterogeneity when this heterogeneity is constant over time. This heterogeneity can be removed from the data through differencing, for example by subtracting the group-level average over time, or by taking a first difference which will remove any time invariant components of the model

Dummy variable

Dummy variables are required to represent a categorical variables the number of dummy variables depends on the number of values that particular categorical variable can take. If we have to represent a categorical variable that can take N different values, we need to define N - 1 dummy variables.

Here in this dataset we have the year in the data is collected, In order to include this attribute in the model we need to encode this to a numerical values or include time demeaned effects.since we have only 3 years data we include 2 binary regressors. While converting this categorical attribute to numerical attribute we need to take care of dummy variable trap issue.

When dummy variables are defined, we need to be careful or else we might end up defining too many variables. If a particular categorical variable takes on N values, it is highly possible that we may define N dummy variables. If we define N dummy variables then we will end up in this trap called Dummy variable trap. This could lead us to linear dependence between these variables so you only need N - 1 dummy variables. A Nth dummy variable is redundant as it carries no new information. And it creates a severe multicollinearity problem for the Regression analysis. In this dataset we have removed one dimension 1991 from the years overcome this problem.

In this data we have data of 40 countries so the mean for all the entities across all years would be calculated and subtracted from all the data points. Then panel regression is applied. The resulting linear equation is given below.

Suicides Rate = - 0.00007 (Gdp per capita) - 0.3485(Year_2001) – 1.74 (Year_2011)
+ Enitity Fixed Effects

```
                        PanelOLS Estimation Summary
================================================================================
Dep. Variable:            suicidesrate   R-squared:                        0.2430
Estimator:                    PanelOLS   R-squared (Between):             -0.2825
No. Observations:                  117   R-squared (Within):               0.2430
Date:                 Wed, Jun 10 2020   R-squared (Overall):             -0.2635
Time:                         14:35:18   Log-likelihood                   -277.20
Cov. Estimator:             Unadjusted
                                         F-statistic:                      8.0261
Entities:                           39   P-value                           0.0001
Avg Obs:                        3.0000   Distribution:                    F(3,75)
Min Obs:                        3.0000
Max Obs:                        3.0000   F-statistic (robust):             8.0261
                                         P-value                           0.0001
Time periods:                        3   Distribution:                    F(3,75)
Avg Obs:                        39.000
Min Obs:                        39.000
Max Obs:                        39.000

                             Parameter Estimates
================================================================================
              Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
gdppercapita -7.713e-05  4.341e-05    -1.7768     0.0796     -0.0002   9.344e-06
Year_2001       -0.3485     0.7399    -0.4711     0.6390     -1.8224      1.1254
Year_2011       -1.7442     1.0700    -1.6301     0.1073     -3.8758      0.3873
================================================================================

F-test for Poolability: 19.524
P-value: 0.0000
Distribution: F(38,75)
```

It turns out that all the N-1 variables included in the model to capture the time fixed effects are statistically insignificant which suggests that there are very less time fixed effects in our data.

	FINAL MODELS	
	Change Regression	**Fixed Effects = Entity**
Change in Gdp Per capita (2011 -1991)	**– 0.0003**	**---**
GDP Per capita	**---**	**- 0.0001**
Sex Ratio	**Insignificant**	**Insignificant**
Median Age of Population	**Insignificant**	**Insignificant**

Conclusion

Before and after (Change) Regression:

Change in suicide Rate (1991 -2011) = – 0.0003 (Change in GDP_per_capita (2011- 1991))

Fixed Effects Regression:

Suicides Rate = - 0.0001 (Gdppercapita) + Entity Fixed Effects

The most appropriate model is the Entity fixed regression as we have negligible time fixed effects, From the above summary of the regression models it is conclusive that there exists a relationship between the GDP per capita of a country and the suicides rate of the country and the relationship turns out to be negative between changes in the suicide rate and the change in the GDP per capita. As the socio economic status grows the country as a whole is more likely to have a good standard of living and there by having a good mental health.

Appendix

PYTHON CODE

IMPORTING LIBRARIES

```
importnumpyasnp
importpandasaspd
importmatplotlib.pyplotasplt
importseabornassns
```

```
df = pd.read_excel("Preprocesseddata.xlsx",sheet_name= "Sheet8")

Original = df.copy()

df.head()

sns.scatterplot(x=df["Change in gdp_per_capita(2011 -1991)"], y= df["Change in Suicide rate(2011 -1991)"] ,data = df)

sns.regplot(x=df["Change in gdp_per_capita(2011 -1991)"], y= df["Change in Suicide rate(2011 -1991)"] ,data = df)

importstatsmodels.api as sm

X = df["Change in gdp_per_capita(2011 -1991)"]

Y = df["Change in Suicide rate(2011 -1991)"]

X = sm.add_constant(X)
```

```
ols= sm.OLS(endog=Y, exog = X)
result = ols.fit()
result.summary()
mi_data = df.set_index(['Country', 'Year'])
print(mi_data.head())
fromlinearmodels import PanelOLS
importstatsmodels.api as sm
Y= mi_data["suicidesrate"]
X= mi_data["gdppercapita"]
#X = sm.add_constant(X)
mod = PanelOLS(Y,X, entity_effects=True)
print(mod.fit())

fromlinearmodels import PanelOLS
Y= mi_data["suicidesrate"]
X= mi_data[["gdppercapita","SEX Ratio"]]
#X = sm.add_constant(X)
mod = PanelOLS(Y,X,entity_effects=True)
print(mod.fit())
fromlinearmodels import PanelOLS
Y= mi_data["suicidesrate"]
X= mi_data[["gdppercapita","Year_2001","Year_2011"]]
#X = sm.add_constant(X)
mod = PanelOLS(Y,X,entity_effects=True)
print(mod.fit())
```

R CODE

```
library("readxl")
library("plm")
df <- read_excel("Finaldata.xlsx")
```

```
head(df)

reg1 = plm(suicidesrate ~ gdppercapita ,
           data = df, index = c("Country","Year"),model = "pooling")
summary(reg1)

reg2 = plm(suicidesrate ~ gdppercapita ,
           data = df, index = c("Country","Year"),model = "random")
summary(reg2)

reg3 = plm(suicidesrate ~ gdppercapita ,
           data = df, index = c("Country","Year"),
           model = "within" )
summary(reg3)
phtest(reg2,reg3)
fixed.dum <-lm(suicidesrate ~ gdppercapita + factor(Country) - 1, data=df)
summary(fixed.dum)
plmtest(reg3, c("time"), type=("bp"))
pbgtest(reg3)
library(lmtest)
bptest(reg3, data = df, studentize=F)
```

References:

https://stattrek.com/multiple-regression/dummy-variables.aspx

https://en.wikipedia.org/wiki/Descriptive_statistics

https://en.wikipedia.org/wiki/Multicollinearity

https://en.wikipedia.org/wiki/Fixed_effects_model#:~:text=In%20panel%20data%20analysis%20the,invariant%20intercept%20for%20each%20subject).

https://en.wikipedia.org/wiki/List_of_countries_by_median_age

https://en.wikipedia.org/wiki/Sex_ratio#:~:text=The%20sex%20ratio%20is%20the,ratio%2C%20either%20periodically%20or%20permanently.

https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/fixed-effects-regression

https://dss.princeton.edu/training/Panel101R.pdf