**ALY6040 Data Mining Application**

**Final Project: Suicide Analysis**

Siyang Feng

Northeastern University

Instructor: Marcus Ellis

**Contents**

[**Introduction** 4](#_Toc12808562)

[**Dataset Introduction** 4](#_Toc12808563)

[**Data Processing and Analysis** 5](#_Toc12808564)

[**Data Wangling and Cleaning** 5](#_Toc12808565)

[**Exploratory Data Analysis** 8](#_Toc12808566)

[**Panel Data Regression** 10](#_Toc12808567)

[**Modeling** 11](#_Toc12808568)

[**Testing** 15](#_Toc12808569)

[**Conclusion** 17](#_Toc12808570)

[**Reference** 18](#_Toc12808571)

[**Appendix – R code** 18](#_Toc12808572)

[Figure 1 Dataset of Suicide 4](#_Toc12808737)

[Figure 2 Structure of Dataset 5](#_Toc12808738)

[Figure 3 Loaded Column name 6](#_Toc12808739)

[Figure 4 NA value detecting of Suicide dataset 6](#_Toc12808740)

[Figure 5 Outliers of country number per year 7](#_Toc12808741)

[Figure 6 Outliers of country number 7](#_Toc12808742)

[Figure 7 GDP per person and gender vs Suicide rate 8](#_Toc12808743)

[Figure 8 Density of suicide rate in gender 8](#_Toc12808744)

[Figure 9 Suicide rate vs Age 9](#_Toc12808745)

[Figure 10 Correlation graph 10](#_Toc12808746)

[Figure 11 Heterogeneity of suicide rate across years 11](#_Toc12808747)

[Figure 12 OLS Panel regression result 12](#_Toc12808748)

[Figure 13 Result of Between Estimation Model 12](#_Toc12808749)

[Figure 14 FD model result 13](#_Toc12808750)

[Figure 15 Result of Fixed effects model 14](#_Toc12808751)

[Figure 16 Result of random effects 14](#_Toc12808752)

[Figure 17 Hausman testing result 15](#_Toc12808753)

[Figure 18 pFtest result 15](#_Toc12808754)

[Figure 19 pFtest of two fixed effects model 15](#_Toc12808755)

# **Introduction**

This project is talking about the research of suicide. The aim of this project is to find out what factors will influence the suicide rate in the world and how these factors effect. By analyzing dataset of Suicide Rates Overview 1985 to 2016 (Kaggel Inc, 2019), I try to find out what kind of people is high risk of suicide which can be the support of social analysis on suicide portion.

Meanwhile, suicide problem has been one of the serious problems in modern society. One the statistic data indicates that suicide is already the 10th leading cause of death in USA and there are 47173 Americans died because of suicide in 2017 ("Suicide Statistics", 2019).

This project will use R to analysis the suicide dataset which comes from Kaggle. The R version in used is 3.6.0 (2019-04-26). The involved packages include dplyr, ggplot2, corrplot, foreign, plm and gplots.

# **Dataset Introduction**

Dataset in this project is Suicide Rates Overview 1985 to 2016 (Kaggel Inc, 2019) which includes 27820 observations and 12 columns in total loaded into R with function read.csv().

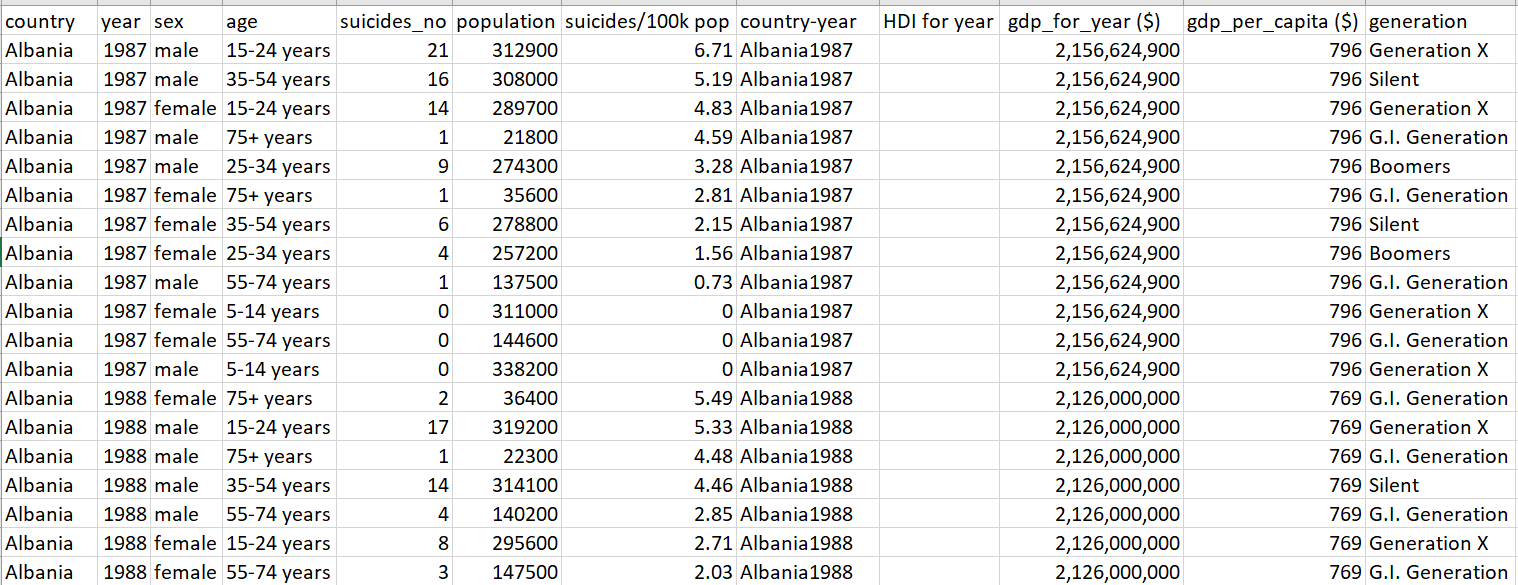


Figure 1 Dataset of Suicide

Figure 1 shows the screenshot of the suicide dataset which includes country, year, sex / gender, age range, suicides number, population, suicides number per 100k population, country-year, HDI, GDP for whole year, GDP per capita and generation. In the dataset, it’s important to understand every column. For instance, the populations column indicates the population of the corresponding country, year and age range. It’s not the whole county population in the corresponding year. Also, the age column contains four age ranges which are 5-14, 15-24, 25-34, 35-54, 55-74 and higher than 75 years. The column of suicide per 100k population indicates can be understand as the suicide rate in the corresponding situation. Generation column records the organization who records the corresponding observation.

This is the panel dataset which is consisted of cross-section data and time series data (year). The start of the project is to go through the dataset after loaded into R.

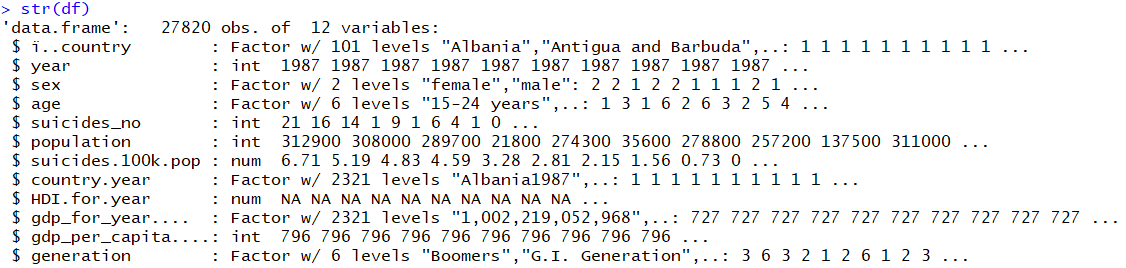


Figure Structure of Dataset

Table 1 display the structure of the dataset. We can easily find out that there are 5 factors, 4 integers and 3 numeric. In the further processing, it’s important to consider if the data type is appropriate in processing.

# **Data Processing and Analysis**

## **Data Wangling and Cleaning**

Data wangling is the critical step before analysis which also impacts the result of the later analysis which transform raw dataset into more appropriate structure. There are four issues we are facing.

* Columns names

When go through the column’s names, we find that some of the names are not standard enough.

A close up of a logo

Description automatically generated

Figure Loaded Column name

Figure 3 displays the original loaded column names and we find that there are a lot of dots following the name. And it makes name unstandardized. Thus, the first step of processing is to rename the column name.

* Data type of some columns

From the figure 2, we find that there are 5 factors including GDP for year. However, the GDP for year should be numeric rather than factor. Because for the further modeling processing, it requires variables to be numeric. Moreover, the numbers in this column are represents with number and comma. To convert factor into numeric, the comma should be omitted and then convert it into numeric.

* Processing of NA value

The next step is checking NA value in the dataset. There are several methods of NA values. Some of the NA values can be replaced by the mean value of the specific conditions. Some of them should be omitted. In this project, we find that there are NA values in the dataset.

A picture containing bottle, indoor

Description automatically generated

Figure NA value detecting of Suicide dataset

Figure 4 displays that if the column contains NA value in the dataset and how many of it. From the figure, we find that only HDI column contains NA value. But this column contains 19456 NA values and only 8364 values are not NA. It means there are about 2/3 of the values are NA. Thus, it’s meaningless to fill the NA value with mean of the HDI. Finally, the HDI column is omitted.

* Outliers

The final step of wrangling is to find out the outliers. The outliers may affect the result of analysis and reduce the model accuracy. This part we will find out the outlier of the dataset.

A screenshot of a cell phone

Description automatically generated

Figure Outliers of country number per year

Figure 5 is the boxplot of recorded country numbers per year. From this figure, we find that in most of the year, it has more than 40 country records. There are one contains very less country records. In the extra analysis, we find that this outlier makes suicide rate extremely low. Thus, in this part, we omit the outlier which is the records in 2016.

A screenshot of a cell phone

Description automatically generated

Figure Outliers of country number

In the figure 6, we find that most of the country is recorded more than 10 year. However, this boxplot displays the outlier of the country which has less than 10 years. Thus, the country whose record number less than 10 will be omitted.

## **Exploratory Data Analysis**

EDA is one of the critical parts in the analysis processing. In this part, it provides the overview of the data structure and helps people to understand the whole data.

A close up of a map

Description automatically generated

Figure GDP per person and gender vs Suicide rate

Figure 7 indicates the relations between suicide rate and GDP per person comparing with the gender. From the graph we find out that people with lower GDP will have more risks on suicide. In general, male is easier to suicide compared with female. When the GDP per capita is very high, the suicide rate of male and female are closed.

A picture containing text

Description automatically generated

Figure Density of suicide rate in gender

Figure 8 displays the density graph of suicide rate comparing with the gender. The red line is female, and the blue line is male. From the graph, we find out that most of the suicide rate about female is closing to 0. However, male is different. In the figure, the density graph of male suicide rate has two peaks and one is far from 0. It indicates that the male suicide rate is significant high in most of the case compared with female.

A screenshot of a cell phone

Description automatically generated

Figure Suicide rate vs Age

Figure 9 shows the relationship between suicide rate and age range. In this figure, we can find out that the suicide rate is extremely low when the age is between 5-14. Also, this range is different because the suicide rate of female is as high as male. The gender is not significant in this age. The highest suicide rate comes from those people who is greater than 75. And the next highest suicide risk happens when the people is between 35 to 54.

Figure 10 indicates the correlation graph of most of the columns. The graph shows that the suicide rate has relations between gender, suicide number, and the age. But no relations between year and other columns. However, this dataset is panel data which related with time series and cross-section data. The correlation graph only tests the relations between the two columns which cannot avoid bias problem. The strength of panel data regression is considering the inter-relations between different variables and avoid bias problem in modeling.

A screenshot of a cell phone

Description automatically generated

Figure Correlation graph

## **Panel Data Regression**

The drawback of normal regression model is that the assumption of this kind model is that all the variables are independent which means the **interrelations** between the variables are ignored. It will cause high bias in most of the regression models. A book, Introduction to econometrics with R (Christoph Hanck, 2019), mentions that the one strength of panel data regression is mitigating the omitted variable bias of dataset whose time dimension changes constantly. This dataset follows the feature of the panel data by the time dimension. Another reason I choose panel data regression to solve this problem is that it’s a significant model in health and medical aspect, and finance and risk analysis.

Figure 11 indicates the heterogeneity of suicide rate across years. The blue line about each year shows the 95% confidence interval of the suicide rate in the year. This graph provide the main trend of the suicide rate about every year.

A close up of a logo

Description automatically generated

Figure Heterogeneity of suicide rate across years

In this project, I only use the linear panel regression model. There are totally 5 different linear panel regression models are in considered. The general function of panel data regression model is:

Comparing with the normal OLS model, , panel regression model not only considers the impacts of variable, but also the impact of variable in exactly time. The five panel models are all coming from the linear model with time.

### **Modeling**

Package plm (panel linear model) is used in this part to create linear panel regression model. In all the following models, suicide rate is estimated by the variables country population, GDP per capita and GDP for year. The panel data index is country and year which means the regression analysis is based on the time of year and category of country.

1. Pooled Ordinary Least Squares (OLS) Model

This model happens through OLS model but with time index. The error term is mostly correlated over time for given residual. In this model, method of plm() function is pooling. And figure 12 display the Pooled OLS model result of this project. In the result, we find out that the country population and GDP per capita are two significant variables in this model. And the variable GDP for year is not significant and can be omitted. The R-square is only about 0.0098.

A screenshot of text

Description automatically generated

Figure OLS Panel regression result

1. Between Estimation Model

Between estimation calculates the average of the dependent and the independent variables over the time. And then, it does the OLS regression. In this project, this model calculates the average value of each variable every year.

A screenshot of a cell phone

Description automatically generated

Figure Result of Between Estimation Model

Figure 13 shows the result of between estimation model. In this model, we find only the intercept is significant and other variables can be omitted. Thus, this model is useless in this case.

1. First differences Estimation Model

The first differences estimation model can exploit the features of a panel data and find the association between the individual-specific changing. The function of this model is:

Get the first difference:

The first differences Estimation Model result is calculated by the plm() function with model ‘fd’. Figure 14 displays the modeling result of this model. It’s easy to find out all the variables are not significant. And the p-value of F-statistic is 0.60177 which is larger than 0.05. Thus, this model is not suitable.

A screenshot of a cell phone

Description automatically generated

Figure FD model result

1. Fixed Effects Model

The fixed effected model and the later random effects model are most common model in panel regression. The result of the fixed effects model is calculated by the function plm() with model function ‘within’. The result is displayed in figure 15.

In the figure 15, it displays that all those there variables are significant and country population and GDP per capita is more significant than GDP for year. The R square of this model is about 0.0675 which is also very low.

A screenshot of a cell phone

Description automatically generated

Figure Result of Fixed effects model

1. Random Effects Model

The result of random effects model is calculated by the function plm() with the model ‘random’. This model assumes that the individual-specific effects are independent of the variables. The result of random effects model is displayed in figure 16.

A screenshot of a cell phone

Description automatically generated

Figure Result of random effects

This model shows that country population and GDP per capita are significant and this model can omit the GDP for year variable. Meanwhile, the R-square of this model is 0.05813 which is smaller than fixed effects model but larger than Pooled OLS model.

### **Testing**

This part is used to test which model is best. Hausman testing is used to test the fixed and random effects model. The null hypothesis of this test is that fixed effects model doesn’t perform better. And the alternative hypothesis is that fixed effects perform better.

A screenshot of a cell phone

Description automatically generated

Figure Hausman testing result

Figure 17 indicates the result of Hausman testing and the p-value is smaller than 0.05. Thus, the null hypothesis is rejected, and we can support fixed effects model in this case.

pFtest is used on fixed effects and pooled OLS model comparation. The null hypothesis is that pooled OLS model performance as good as fixed effects model. The figure 18 displays the result of pFtest. In the result of the test, the p-value is smaller than 0.05. Thus, fixed effects model performs better than pooled OLS model.

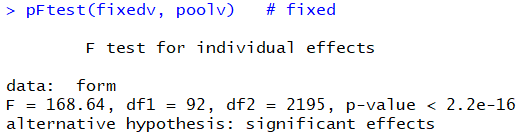


Figure pFtest result

After we defining the fixed model is best in linear panel regression. Then, we can create the time-fixed effects model and compare it with the previous fixed effects model. In the time-fixed model, a new factor is added in the training model which is factor(year). After creating the new model, pFtest is used to compare the new model and the previous fixed model.

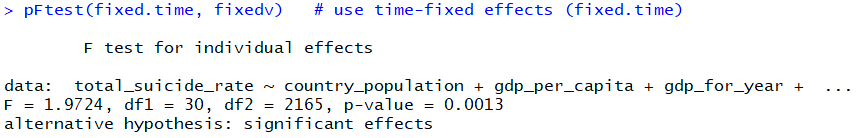


Figure pFtest of two fixed effects model

Figure 19 indicates the testing result and the p-value is smaller than 0.05 which means the time-fixed effects model has the significant effects compared with fixed effects model. The result of time-fixed effects model shows r-square is 0.092346.

|  |
| --- |
| Oneway (individual) effect Within Model  Call:  plm(formula = total\_suicide\_rate ~ country\_population + gdp\_per\_capita +  gdp\_for\_year + factor(year), data = suicide\_overview\_norm,  model = "within", index = c("country", "year"))  Unbalanced Panel: n = 93, T = 6-31, N = 2291  Residuals:  Min. 1st Qu. Median 3rd Qu. Max.  -0.3103247 -0.0251882 -0.0018856 0.0243684 0.3212586  Coefficients:  Estimate Std. Error t-value Pr(>|t|)  country\_population 0.76607955 0.13130770 5.8342 6.217e-09 \*\*\*  gdp\_per\_capita -0.15529369 0.02456788 -6.3210 3.146e-10 \*\*\*  gdp\_for\_year -0.17837891 0.06580279 -2.7108 0.006765 \*\*  factor(year)1986 -0.00937340 0.01277676 -0.7336 0.463254  factor(year)1987 0.00233423 0.01244148 0.1876 0.851194  factor(year)1988 0.00579739 0.01275512 0.4545 0.649504  factor(year)1989 0.00229643 0.01259400 0.1823 0.855330  factor(year)1990 0.00102852 0.01212397 0.0848 0.932401  factor(year)1991 0.00818363 0.01214582 0.6738 0.500522  factor(year)1992 0.01034650 0.01212997 0.8530 0.393770  factor(year)1993 0.01727848 0.01207979 1.4304 0.152757  factor(year)1994 0.01925003 0.01202224 1.6012 0.109478  factor(year)1995 0.03195045 0.01172104 2.7259 0.006464 \*\*  factor(year)1996 0.02395429 0.01177150 2.0349 0.041979 \*  factor(year)1997 0.02034272 0.01176631 1.7289 0.083970 .  factor(year)1998 0.02212561 0.01172086 1.8877 0.059198 .  factor(year)1999 0.02276089 0.01166253 1.9516 0.051112 .  factor(year)2000 0.02002433 0.01159799 1.7265 0.084394 .  factor(year)2001 0.01550828 0.01153159 1.3449 0.178814  factor(year)2002 0.01579850 0.01160837 1.3610 0.173669  factor(year)2003 0.01078916 0.01168778 0.9231 0.356050  factor(year)2004 0.01003780 0.01182634 0.8488 0.396105  factor(year)2005 0.01012022 0.01193405 0.8480 0.396525  factor(year)2006 -0.00082199 0.01196406 -0.0687 0.945231  factor(year)2007 0.00748838 0.01211883 0.6179 0.536698  factor(year)2008 0.01082023 0.01228536 0.8807 0.378555  factor(year)2009 0.00737402 0.01203073 0.6129 0.539986  factor(year)2010 0.00221211 0.01213911 0.1822 0.855420  factor(year)2011 0.00229622 0.01239027 0.1853 0.852992  factor(year)2012 0.00431367 0.01244361 0.3467 0.728883  factor(year)2013 -0.00069654 0.01254630 -0.0555 0.955731  factor(year)2014 -0.00829920 0.01267409 -0.6548 0.512656  factor(year)2015 -0.01967586 0.01302250 -1.5109 0.130957  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Total Sum of Squares: 9.2188  Residual Sum of Squares: 8.3674  R-Squared: 0.092346  Adj. R-Squared: 0.039941  F-statistic: 6.67488 on 33 and 2165 DF, p-value: < 2.22e-16 |

# **Conclusion**

In conclusion of this project, I create the time-fixed effects linear panel regression model to find out what both country population and GDP per capita and GDP for year are significant in this model. From the model, we find out that in the specific year, the country with higher GPD will has lower suicide rate. And in the specific country, when the country population increase, the suicide rate has the higher risk to increase. Then, comparing with the EDA part, we find that male is high risk group on suicide. Meanwhile, the people who is older than 75 has higher risk on suicide comparing with younger people. This model could be used to test if the country are in the high risk of suicide and solve the suicide problem in the high risk group.

However, we find out that the R-square of this model is only 0.092346 which is very low. The reason of the bad performance of this model could because the factors age and gender are omitted in this model. However, I cannot find any efficient way to add these two factors. Also, this project only uses the linear panel regression. In the further processing, the non-linear panel regression should be considered, and it might be more efficient. Moreover, we find that 75+ people have highest risk of suicide. It’s low possible that people who is older than 75 suicide because of the GDP. In this dataset, there are only population, GDP are included. The HDI data has a lot of missing data that we must omit it. One of the reasons of bad performance about this model could be the lack of the reference data.

# **Reference**

Kaggel Inc. (2019). Who Suicide Statistics. retrieved from: <https://www.kaggle.com/szamil/who-suicide-statistics>.

Suicide Statistics. (2019). Retrieved from <https://afsp.org/about-suicide/suicide-statistics/>

Christoph Hanck, A. (2019). 10 Regression with Panel Data | Introduction to Econometrics with R. Retrieved from <https://www.econometrics-with-r.org/10-rwpd.html>

# **Appendix – R code**

|  |
| --- |
| rm(list=ls()) # wipe environment  #install.packages("ggplot2")  require(dplyr)  require(tidyr)  require(ggplot2)  # read dataset  df <- read.csv("master.csv")  ################# data cleaning #################  # overview of dataset  dim(df)  str(df)  colnames(df)  # standard column name  colnames(df) <- c('country', 'year', 'sex', 'age', 'suicides\_no', 'population',  'suicide\_rate', 'country\_year', 'HDI', 'gdp\_for\_year',  'gdp\_per\_capita', 'generation')  attach(df)  colnames(df)  ## convert data structure  # define the function of converting factor into number  facToNum <- function(column) {  column <- as.character(column)  column <- gsub(',', '', column) %>%  as.numeric()  return(column)  }  # convert factor of gdp\_for\_year into numeric  df$gdp\_for\_year <- facToNum(df$gdp\_for\_year)  # check NA in each column  anyNA(df)  summary(is.na(df))  # omit HDI column because there are branch of NA value  newdf <- df[-c(9)]  newdf$gender <- sapply(newdf$sex, function(x) {ifelse(x=='male', 1, 2)})  # add column of age with value 1 to 6  ageLabel <- function(x) {  if (x == '5-14 years') return(1)  else if (x == '15-24 years') return(2)  else if (x == '25-34 years') return(3)  else if (x == '35-54 years') return(4)  else if (x == '55-74 years') return(5)  else return(6)  }  newdf$new\_age <- sapply(newdf$age, ageLabel)  str(newdf)  summary(is.na(newdf)) # check NA  ##################### Exploratory Data Analysis (EDA) and extra data wrangling ################  # total suicide number, per year per country  suicide\_overview <- newdf %>%  group\_by(year, country) %>%  summarise(suiNo\_year\_country = sum(suicides\_no),  country\_population = sum(population),  gdp\_for\_year = mean(gdp\_for\_year),  gdp\_per\_capita = mean(gdp\_per\_capita)) %>%  mutate(total\_suicide\_rate = suiNo\_year\_country / country\_population \* 100000) %>%  ungroup()  # check the recorded country number per year  country\_peryear <- suicide\_overview %>%  group\_by(year) %>%  summarise(country\_num = n()) %>%  ungroup()  #check record number with boxplot  boxplot(country\_peryear$country\_num,  main = "Recorded Country Number per Year",  xlab = "country number",  col = 'light blue',  notch = TRUE,  horizontal = TRUE)  # get the outlier based on the boxplot which is the min value  # outlier: 2016  filter(country\_peryear, country\_peryear$country\_num == min(country\_peryear$country\_num))  # count the number of recorded countries  country\_sum <- suicide\_overview %>%  group\_by(country) %>%  summarise(count\_freq = n())  boxplot(country\_sum$count\_freq,  main = "Recorded number for different country",  xlab = 'recorded number per country',  col = 'light green',  notch = TRUE,  horizontal = TRUE)  # check outliers based on the boxplot  outliers <- filter(country\_sum, country\_sum$count\_freq < 10)  outliers  # cleaning rare country observations  out\_country <- c(as.character(outliers$country))  cleaned\_df <- filter(newdf, !(newdf$country %in% out\_country))  # suicide rate per year  suicide\_year <- cleaned\_df %>%  group\_by(year) %>%  summarise(sui\_rate\_year = sum(suicides\_no) / sum(population) \* 100000) %>%  ungroup()  ggplot(suicide\_year, aes(suicide\_year$year, suicide\_year$sui\_rate\_year)) +  geom\_line(color = "blue", linetype = "dashed", size = 1.2) +  geom\_point(size = 3) +  ggtitle("Average suicide rate (per 100k population) Per Year")  # cleaning all the observation from 2016  cleaned\_df = filter(cleaned\_df, cleaned\_df$year != 2016)  ############################ log transformation and data normalization ########################  #### working for modeling  View(cleaned\_df[c(5, 6, 7, 9, 10)])  new\_suicide\_overview <- cleaned\_df %>%  group\_by(year, country) %>%  summarise(suiNo\_year\_country = sum(suicides\_no),  country\_population = sum(population),  gdp\_for\_year = mean(gdp\_for\_year),  gdp\_per\_capita = mean(gdp\_per\_capita)) %>%  mutate(total\_suicide\_rate = suiNo\_year\_country / country\_population \* 100000) %>%  ungroup()  model\_df <- cleaned\_df  # normalization (Min-Max normalization)  model\_df[c(5, 6, 7, 9, 10)] <- sapply(model\_df[c(5, 6, 7, 9, 10)],  function(x) {  return((x - min(x)) / (max(x) - min(x)))  })  suicide\_overview\_norm <- new\_suicide\_overview  suicide\_overview\_norm[-c(1, 2)] <- sapply(new\_suicide\_overview[-c(1, 2)],  function(x) {  return((x - min(x)) / (max(x) - min(x)))  })  ##############################################################################################  ###### EDA  ggplot(model\_df, aes(x = gdp\_per\_capita, y = suicide\_rate, shape = sex, color = sex)) +  geom\_point() +  geom\_smooth(method = 'lm', se = FALSE) +  ggtitle("GDP per person VS suicide rate (sex identify)")  ggplot(model\_df, aes(suicide\_rate, color = sex)) +  geom\_density() +  ggtitle("Suicide rate density compared with sex")  ggplot(model\_df, aes(x = age, y = suicide\_rate, fill = sex, color = sex)) +  geom\_point() +  ggtitle("Relations between age and suicide rate")  #install.packages("corrplot")  library(corrplot)  corrplot(cor(model\_df[-c(1, 3, 4, 8, 11)]), order = 'hclust', tl.col = 'black', tl.cex = .65)  corrplot(cor(suicide\_overview\_norm[-c(2)]), order = 'hclust', tl.col = 'black', tl.cex = .65)  ################################# panel data regression #####################################  #install.packages("foreign")  #install.packages("plm")  require(foreign)  require(plm)  require(gplots)  attach(suicide\_overview\_norm)  plotmeans(total\_suicide\_rate ~ year, main = "Heterogeineity across years", data = suicide\_overview\_norm)  #### model 1: OLS (Ordinary Least Squares) Model  form <- total\_suicide\_rate ~ country\_population + gdp\_per\_capita + gdp\_for\_year  poolv <- plm(form, data = suicide\_overview\_norm, model = 'pooling', index = c("country", "year"))  summary(poolv)  #### model 2: Between estimation  btv <- plm(form, data = suicide\_overview\_norm,  model = 'between',  index = c("country", "year"))  summary(btv)  #### model 3: First differences estimation  fdv <- plm(form, data = suicide\_overview\_norm,  model = 'fd',  index = c("country", "year"))  summary(fdv)  #### Model 4: fixed effects model  fixedv <- plm(form, data = suicide\_overview\_norm,  model = 'within',  index = c("country", "year"))  summary(fixedv)  #### Model 5: Random effects model  randomv <- plm(form, data = suicide\_overview\_norm,  model = 'random',  index = c("country", "year"))  summary(randomv)  #### model test  # fixed vs random (hausman test)  phtest(randomv, fixedv) # fixed  # fixed vs OLS (pFtest)  pFtest(fixedv, poolv) # fixed  fixed.time <- plm(total\_suicide\_rate ~ country\_population + gdp\_per\_capita + gdp\_for\_year + factor(year),  data = suicide\_overview\_norm,  model = 'within',  index = c("country", "year"))  summary(fixed.time)  pFtest(fixed.time, fixedv) # use time-fixed effects (fixed.time) |