FT18_AnalyticsPracticum_II

Burcin_Sarac

30 05 2019

Data Preparation Part

This is an R Markdown document In this document, my aim is to understand position of suicide numbers of "Greece" among EU & some other OECD Countries by using visualizations with the helps of two different datasets. Both datasets include suicides throughout European Union; - The first dataset provided from Eurostat with absolute numbers for the period 2011-2015, which includes 33 countries including 29 EU countries with Norway,Iceland, Serbia, Turkey and EU Total as well as one time. This dataset manually copied into a xls file. - The second dataset provided from OECD for 2004-2014 period but only includes 23 of EU countries and it shows only suicide frequency of countries per 100 000 people one by one.This data also imported as txt document for further analyses.

```
library(shiny)
library(tidyverse)
library(ggrepel)
library(shinythemes)
```

After attaching packages, I read the datasets, however, since the datasets saved in a specific format, they needed to be cleaned or at least required a specific package to read.

Firstly I read excel file which included Eurostat data by using `read_xls()` function of `readxl` package and skipping first 2 rows. And then I used `melt()` function from `reshape2` package to collect all year columns into one column with name "Year".

```
suicide1 <- readxl::read_xls("suicide1.xls", skip=2)
colnames(suicide1)[1] <- "GEO"
suicide1 <- reshape2::melt(suicide1, id.vars=c("GEO"), variable.name = "Year",
              value.name = "Total")
```

Then I group data into countries by using `group_by()` function by defining a different dataset and in this dataset I also dropped EU Total rows.

```
suicide1_2 <- group_by(suicide1, GEO)
suicide1_2 <- suicide1_2[ ! suicide1_2$GEO %in% "European Union (current composition)", ]
```

After reading first dataset, I read the second .txt dataset with the helps of 'readLines' function, but dataset was not clean to read, so I combined this function with `gsub()` when reading dataset. With the helps of `gsub()` I dropped "" sign both from the beginning and the end of rows in data before reading it.

```
suicide2 <- read.csv(text=gsub("[\"\\]\\$", "",
                               readLines("suicide II OECD.csv.txt",
                                         skipNul = T)))
```

Since I read the data by using `readLines()` function, it read all data as one vector, so I required to clean data and define columns from scratch. For doing this I first chanced the separation of column names in the first row, which was `'.'` before and I changed it into `','`. And then I separated them from comma sign with the helps of `separate()` function from `dplyr` package under `tidyverse` package, which I imported at the beggining.

```
colnames(suicide2) <- "LOCATION,INDICATOR,SUBJECT,MEASURE,FREQUENCY,TIME,Value,Flag.Codes"

suicide2 <- separate(suicide2, colnames(suicide2),
                     into = c("LOCATION","INDICATOR","SUBJECT",
                              "MEASURE","FREQUENCY","TIME","Value",
                              "Flag.Codes"), sep=",")
```

After cleaning and separation, I dropped unnecessary columns and convert year and value columns into numeric with the codes below.

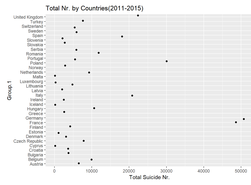
```
suicide2 <- suicide2[-c(2:5,8)]

suicide2$TIME <- as.numeric(gsub("[\"\\]$", "", suicide2$TIME))
suicide2$Value <- as.numeric(suicide2$Value)
```

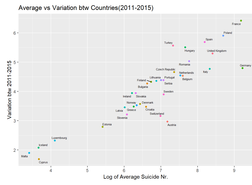
Visualization Part

Firstly, I sum all years into one value and showed by country only by using the first dataset, in other words from the range between 2011 and 2015 with absolute numbers. It seems from the graph below that, highest total number of suicides belongs to Germany and the lowest belongs to Cyprus together with Iceland and Malta.

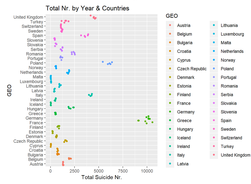
However since these are absolute numbers, with this dataset we cannot say anything about ration of suicide numbers to the population, so it might mislead us to draw conclusion. For example Germany is one of the highest populated country among these countries, so higher suicide numbers would be expected as it observed.



At the next graph below, I calculated mean and standard deviation of all years from 2011 to 2015 by countries, and for scaling data I took logarithm. As it observed from the previous graph, this graph also showed that Germany had highest average number of suicides, however it could be seen that highest variation observed in France, in other words suicide numbers in France differentiates more than other countries between these years.



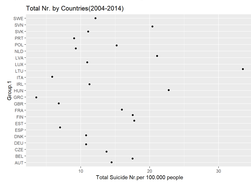
In the next graph, I showed suicide numbers by countries again, but this time I did not sum up years to see countries has close amount of suicide numbers in different years. And it can be said that this hypothesis would be true. However it seems better to draw a more detailed graph.



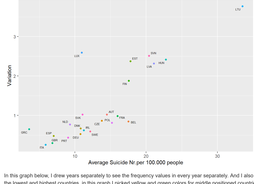
So I drawed next graph for more details. In this graph years were showed in different subgraphs and color is in the darkest blue form in the lowest point and become red with increase in numbers. SO this graph provided better visualization to understand that Greece has one of the lowest amount of suicide numbers among EU countries.

This time, I used second dataset for further visualizations. Since this dataset informs about frequency instead of absolute numbers, it allowed me to understand difference between countries.

In this first graph of second dataset, I just calculated average of all years and try to see average frequency numbers and from this graph I can easily say that Greece has the lowest frequency rate among countries.



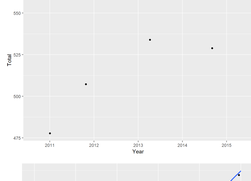
In this second graph I also took into account standard deviation together with means of frequency between 2004 and 2014. Since it could also be seen from this graph that Greece has lowest mean value, its variation changes over years compare to a couple of countries like Italy, GB, Spain. And also Lithuania is the highest by far beyond others in terms of both variation and frequency.

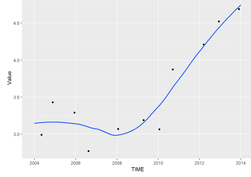


In this graph below, I drew years separately to see the frequency values in every year separately. And I also colorized frequency values to catch the lowest and highest countries. in this graph I picked yellow and green colors for middle positioned countries on purpose to highlight only the lowest and highest countries. And it showed us Greece as lowest in all years. And the highest country seen as Lithuania.

Lastly in these two graphs below, I aimed to show only the frequency numbers of Greece and also the difference between years. The first among these two was created by using the first dataset only includes absolute numbers and years between 2011 to 2015. And the second graph produced by using second dataset with frequencies between the years 2004 to 2014 and also the trend.

Both graphs showed that there is an increasing trend among years generally.





In conclusion, I can say that, Greece has the lowest suicide ratio among EU countries and although there is an increase trend, it keeps its position in long run.

Lastly, I added a dynamic plot below, to see country trends one by one only with selecting the desired country name from the selector for further information.