Cork Institute of Technology

MSc in Data Science and Analytics

Assignment 1 – Visualization of a Dataset.

Givaldo Silva Jr



supervised by

Aengus Daly

19th March 2019

Abstract

This project explores the data set “World Development Indicators”, using R language, ggplot2, spread and gather. Most part of the graph are animated, what gives us a better view of the increasing or decreasing of some variable.

The focus of this exploration was Europe or some comparison between Europe and the world.

Acknowledgements

I would like to thank my supervisor, Aengus Daly, for all his support and always sharing his professional and academic experience.

Content

[1. Dataset 5](#_Toc35816976)

[2. Top ten countries GPD 7](#_Toc35816977)

[3. Life expectancy vs GDP - Europe 8](#_Toc35816978)

[4. Life expectancy vs GDP - continents 9](#_Toc35816979)

[5. GDP Europe – line graph interactive 10](#_Toc35816980)

[6. Evolution of GDP – Ireland and UK 11](#_Toc35816981)

[7. Evolution of corona virus 12](#_Toc35816982)

[8. GIS – Life expectancy comparing Europe and world 13](#_Toc35816983)

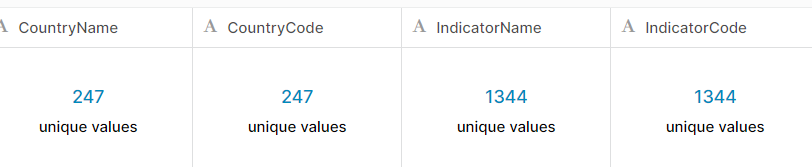
[9. Reference 14](#_Toc35816984)

# Dataset

This dataset (indicators) has more than 1.6 Million lines [6].

Two datasets were merged, Indicators and Country, it allows me to group countries by their continents.

This visualization (figure 1) was done using PowerBI, we can easily do that with PowerBI just loading the dataset, if both datasets have the same ID it will be done automatically, if not we can manually do that.



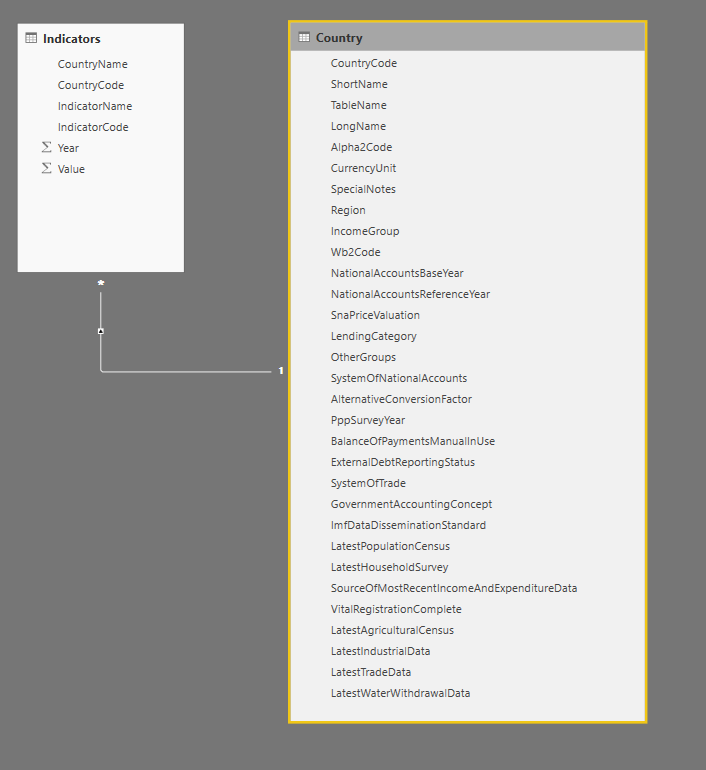


Figure 1

Indicators2 <- merge

(Indicators,Country\_Region, by=c("CountryCode","CountryCode"))

I have used the function **merge** to concatenate both datasets, by “**CountryCode**” column.

Indicators\_europe <- Indicators2 %>%

filter(Region=="Europe & Central Asia")

My focus was Europe, then I have used the function **filter** to get just Europe and Central Asia information.

wide\_format <- Indicators\_europe[,-c(3)] %>%

spread(key = "IndicatorCode", # Name of new key column

value = "Value") # Name of column for values

The function **spread** was used to turn the dataset in a wide format, it makes easier to plot the graphs.

# Top ten countries GPD

This animated bar graph (fugire2) is very useful to check the evolution of some trend, in that case, I am exploring the GDP per year between European countries [2].

The main propose of this graph is show for the audience how some specific country is changing over the time among others, we could as it for different things like for example, the profit of some companies or maybe how a specific product is making more profit compared with other products of that company. In the presentation, we can point a reason for the time where this product started to increase its profit.

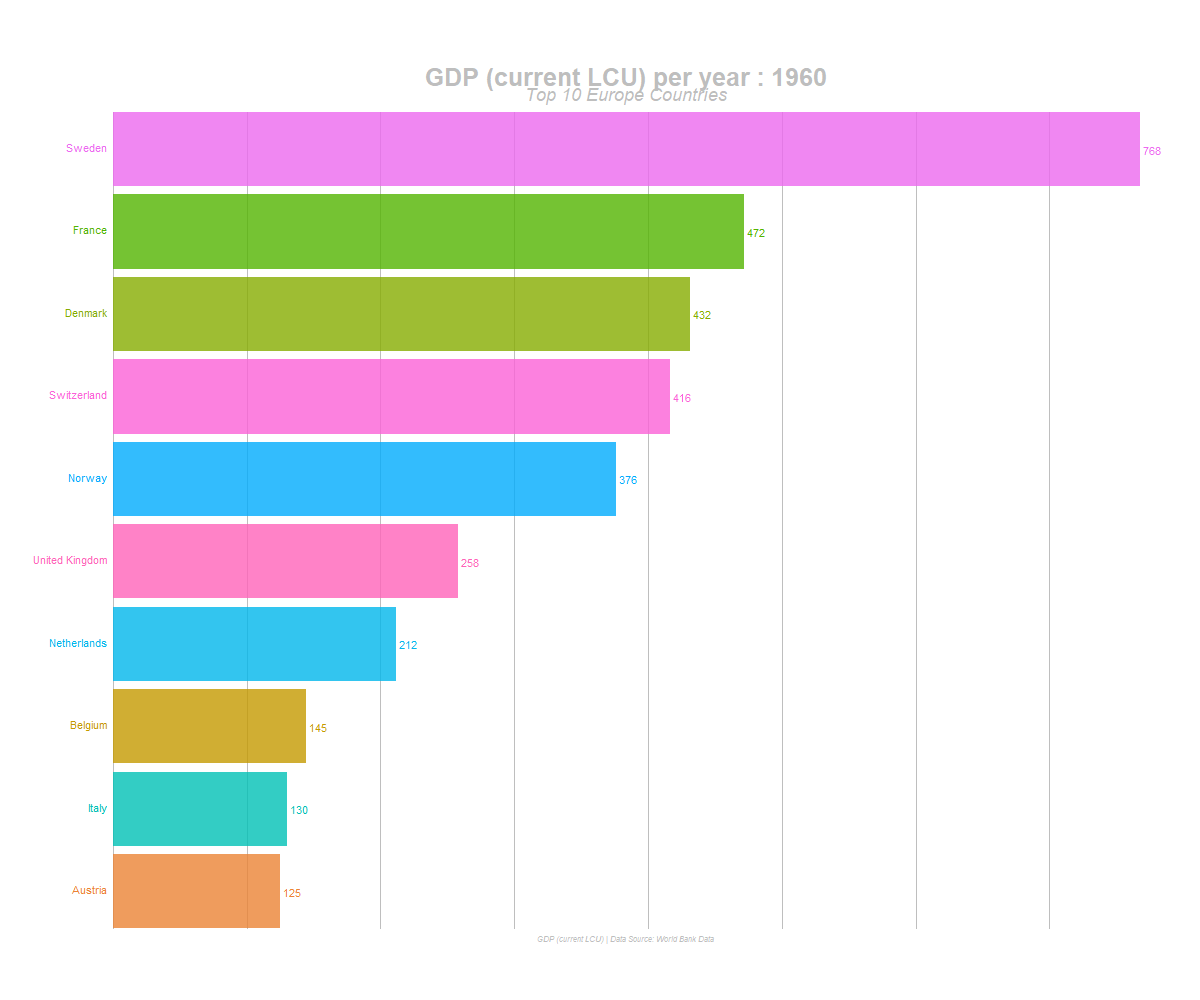


Figure 2

# Life expectancy vs GDP - Europe

Bubble chart (figure3) is very good to compare the evolution of three variables, in this animated graph we can observe how countries change during the years. It is easy to verify the tendencies, where they are going and when they reach some specific point.

In this example we can see the evolution of life expectancy and GDP over years, the size of population does not have so big change, then it is not possible to visualize it in the graph, but it still being a good information because we can make some comparisons between largest populations. In the second war for example there was a great decrease of life expectancy, and this graph show exactly that biggest countries suffered more [7].

In my last job, I had the opportunity to use this kind of graph in a project to explain for the vice president of the company 7 different scenarios of the project, in the X-axis I put how long each project would take, in the Y-axis the size of the investment and the size of the bubble was the return of each project, how bigger was the bubble bigger was the return. That gave us a very clear visualization of something that could be difficult to identify just looking at the number, the best option was the biggest bubble most closely to zero.

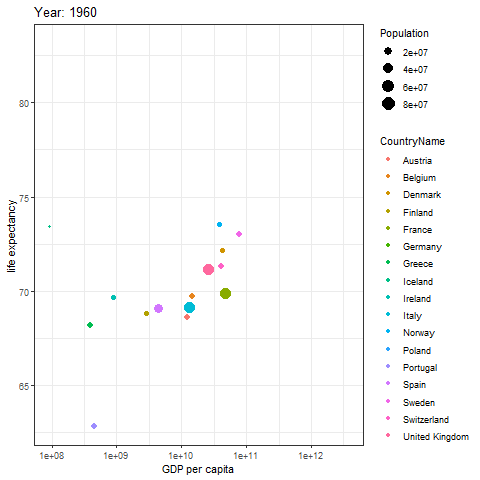


Figure 3

# Life expectancy vs GDP - continents

This bubble chart is almost the same as the previous one, but we can separate by continent or another category that we need, in that way it is possible to visualize the evolution by each category, and we could see some similarities or differences in the same category.

It also gives us some clear information, like in Europe at 1960 some countries had a high level of life expectancy and GDP while one country did not, Latin American had a very different level of life expectancy independently of their GDP, North America was very close and South Africa had a very low life expectancy in that year.

This graph also, allows us to see the speed that some countries can change its situation, to identify some abrupt decrease or increase, it is another very useful utility.

Mapa com linhas coloridas

Descrição gerada automaticamente

Figure 4

# GDP Europe – line graph interactive

This is an interactive line graph showing the evolution of the GDP of some European countries, it is easy to see the tendency between these countries, almost all of them follow the same variation over the years.

This type of graph is good because it enables us to explore different views like for example select just some specific country or few countries to do a better visualization, we can also give a zoom to check some details, visualize just one value or all values at the same time.

If we have too much information it is very useful because we can change it and choose what we want to do our analysis in real-time, instead of doing several different graphs.

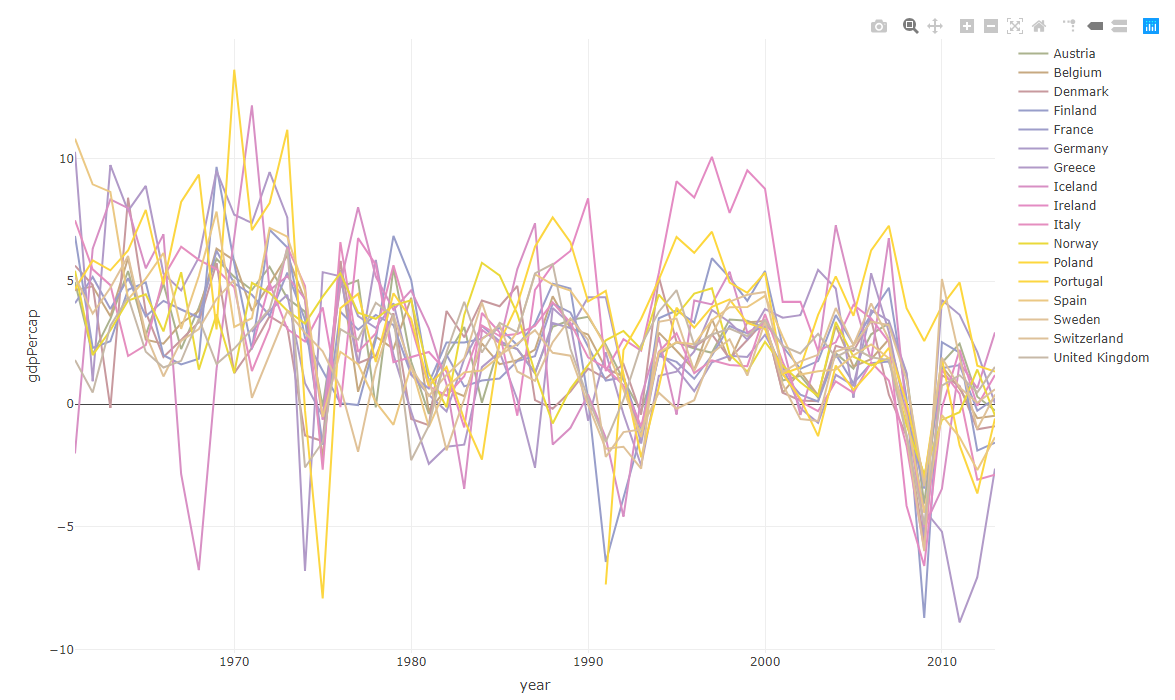


Figure 5

# Evolution of GDP – Ireland and UK

Here we can see a comparison between GDP from Ireland and UK, with this animated line graph it is possible to visualize how their economy is similar and maybe affect each other, the colors are different, which makes it easy to visualize both countries. The animation can emphasize how some country is increasing or decreasing faster than other one in a specific point over time, in that way the audience can discuss why that country had this increasing while the other country remained in a plateau.

We could use this kind of graph to compare different investments and the impact in the final profit, the result of sales, this would show us which investment gave the best result, the highest profit.

Tela de celular com publicação numa rede social

Descrição gerada automaticamente

Figure 6

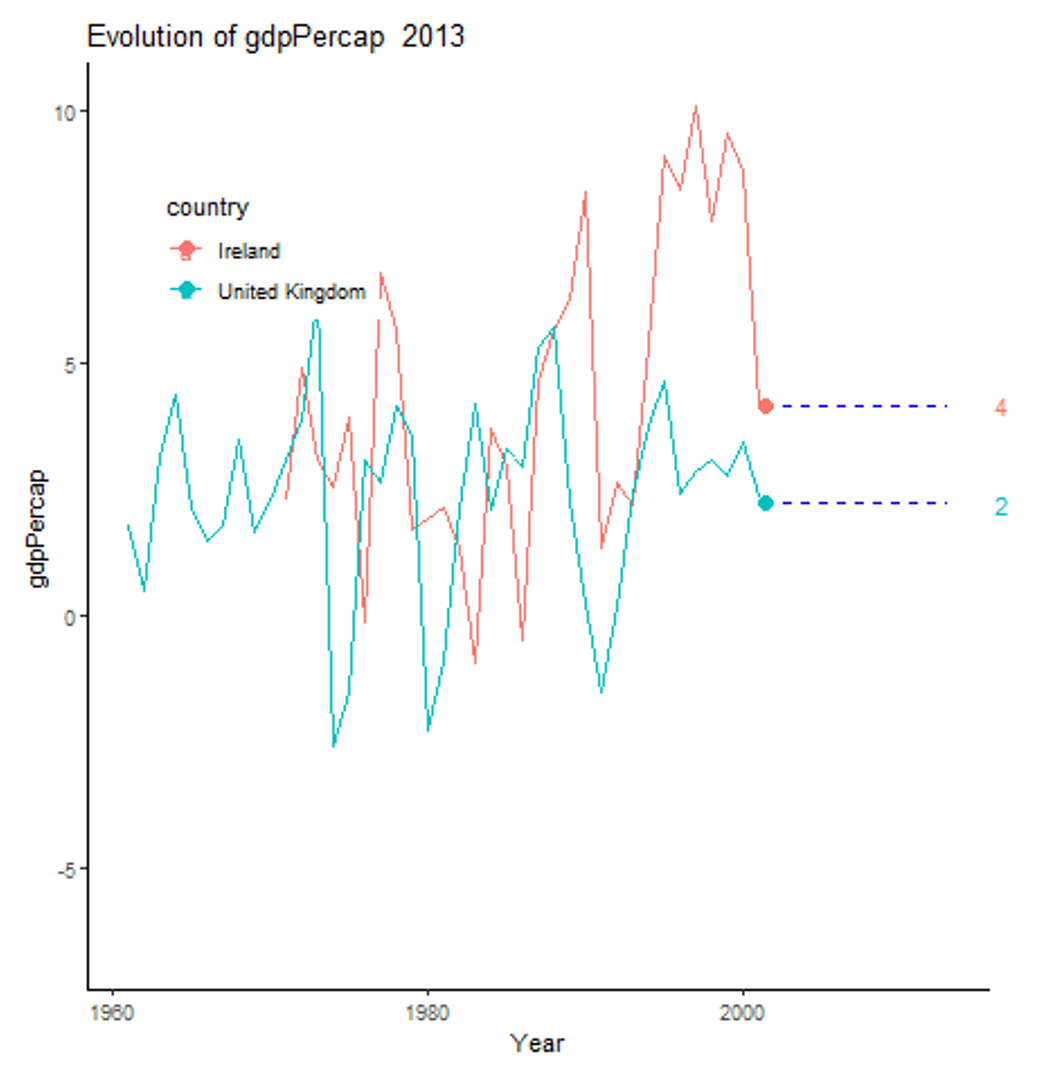


Figure 7

# Evolution of corona virus

This graph [8] shows the evolution of coronavirus over time and the difference between three-stage of cases, confirmed, recovered and deaths. I have copied it because it is a very good example of how we can use this graph, we can see clearly that the confirmed cases are much higher than deaths, so the probability of death is very low. And, it is very useful to see how it changes over time, how confirmed cases increase so fast, but deaths almost do not change at all.

Tela de celular com publicação numa rede social

Descrição gerada automaticamente

Figure 8

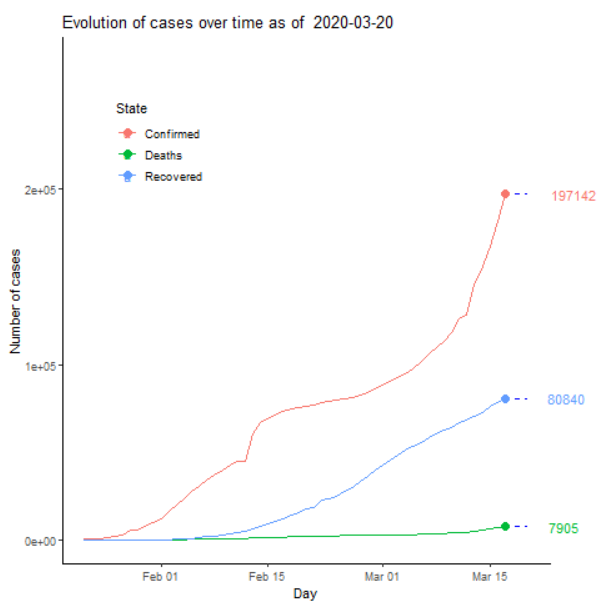


Figure 9

# GIS – Life expectancy comparing Europe and world

Geographic information system [10] helps us to analyses spatial location and layers of information’s using 3D maps for example. With different colors we can demonstrate changes between some periods among different places, in the example below (figure 12-15) we can see life expectancy between European countries compared with the rest of the world.

Figure 10 shows us the countries where coronavirus have started in February, this graph was made using PowerBI, this system like Tableau, Orange, Qlikview, etc are quite useful to plot some graphs very fast, with just a few clicks we can get a good visualization or a dashboard. But if we need something more complex maybe only this would not be a good solution. In the case of PowerBI, it is possible to insert some R code to improve the visualization, as we can see in figure 11.

GIS graphs are very powerful visualization, but we need to take care of some points like the colors, for example, otherwise, we can lose the accuracy of the information and our audience will not be able to get the main information.

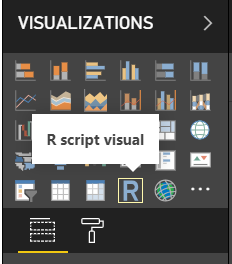
 

Figure 10 Figure 11

Uma imagem contendo mapa

Descrição gerada automaticamenteUma imagem contendo texto, mapa

Descrição gerada automaticamenteUma imagem contendo texto, mapa

Descrição gerada automaticamenteUma imagem contendo mapa

Descrição gerada automaticamente

Figure 13

Figure 14

Figure 12

Figure 15

Total of words: [ 1122]

# Reference

I am using IEEE stile reference.

[1]G. Grolemund and H. Wickham, 5 Data transformation | R for Data Science. .

[2]“amrrs/animated\_bar\_charts\_in\_R: Code to demo how to build Animated Bar Charts in R.” [Online]. Available: https://github.com/amrrs/animated\_bar\_charts\_in\_R. [Accessed: 15-Mar-2020].

[3]“Bar Chart Races With gganimate.” [Online]. Available: https://emilykuehler.github.io/bar-chart-race/. [Accessed: 15-Mar-2020].

[4]“Data Science,” edX. [Online]. Available: https://www.edx.org/micromasters/uc-san-diegox-data-science. [Accessed: 15-Mar-2020].

[5]“Emigration by age and sex - ecodp.common.ckan.site\_title.” [Online]. Available: https://data.europa.eu/euodp/en/data/dataset/MfkEgsB8RmSr3syAcyHCw. [Accessed: 15-Mar-2020].

[6]R. L. Muenchow Jakub Nowosad, Jannes, Geocomputation with R. .

[7]M. Roser, E. Ortiz-Ospina, and H. Ritchie, “Life Expectancy,” Our World in Data, May 2013.

[8]P. Rotzetter, protzetter/coronavirus. 2020. <https://github.com/protzetter/coronavirus>

[9]“World Development Indicators.” [Online]. Available: https://kaggle.com/worldbank/world-development-indicators. [Accessed: 15-Mar-2020].

[10]“What is GIS? | Geographic Information System Mapping Technology.” [Online]. Available: https://www.esri.com/en-us/what-is-gis/overview. [Accessed: 22-Mar-2020].