**Riley Fencl**

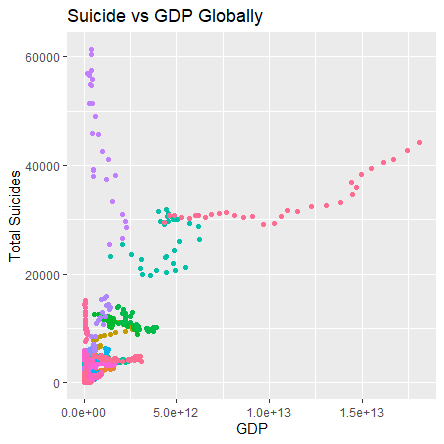
**Final Project – Part III**

**Introduction**

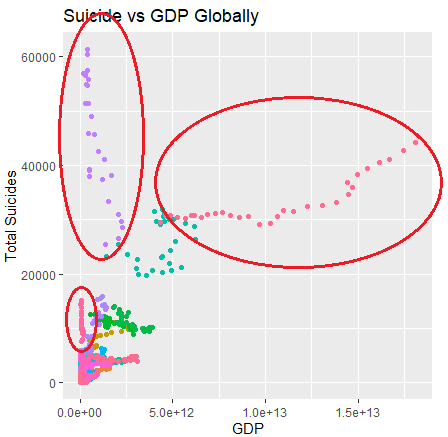
Initially, I set out on this project to gain a better understanding of how suicide permeates itself in the global environment. Admittedly, I believe I have finished it with far more questions than I had going into to it, however I feel like I have gained quite the understanding of how this problem looks globally. I tackled this endeavor because I had heard often that it was a symptom of mostly first world countries and while I don’t think I found evidence that pointed to that idea being true save for one exception; I figured something out along the way that was pretty obvious looking back. The truth with suicide is that it is just as complex as any other problem and it goes far deeper beyond the scope of my data or at least my analysis capabilities. That being said, what I found in the data was beyond fascinating.

**Questions 1 & 2: How does GDP impact suicide rate and is there a global predictor for suicide?**

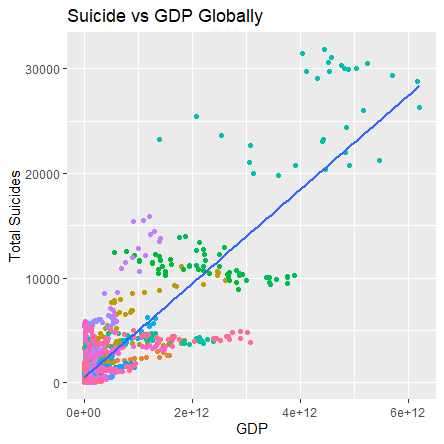
My plan initially was to create a model and see how the data worked with it. However, through hell and high water, and every damn data transformation I could think of, I couldn’t get the data to cooperate with me whatsoever. Not only that, but when I started looking at the different scatter plots for suicide I noticed that the relationships between the data were fairly unclear with a few exceptions. So instead of looking for the primary predictor of suicide I decided to look at the primary factors based off of the scatter plot relationships first starting with GDP and HDI and attempted to establish a global predictor.



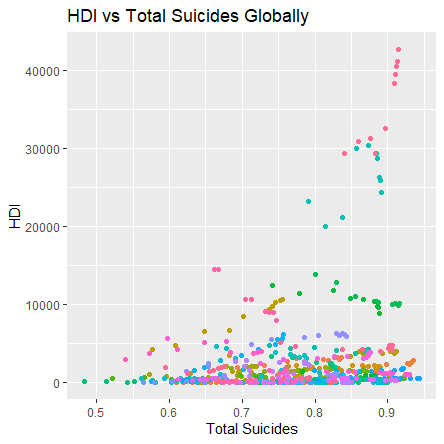
Here is the scatterplot that looks at suicide as a function of GDP across the globe. I removed the legend because it had some 256 countries, but don’t worry we only need to take note of a few. Firstly, if you look at the plot you will notice three fairly obvious irregularities.



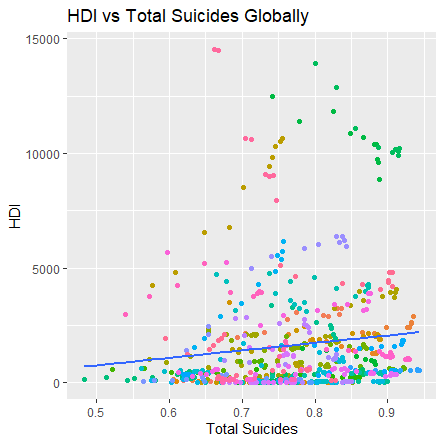
The purple spike is the Russian Federation and the spike below it is Ukraine. I’m not sure what exactly is going on here but comparatively to the rest of the world they seem to be pretty obviously not with the general trend. Additionally, the spike on the far right of the plot that extends out the furthest is the United States. While the US maybe be in the same direction as the rest of the world it is definitely far more pronounced. Now whether these are of statistical significance or mathematically egregious outliers is fairly irrelevant, because my goal was to see if there were common links to suicide globally, and actually with these exceptions from the plot removed it does seem like the rest of the world follows along the same trend.



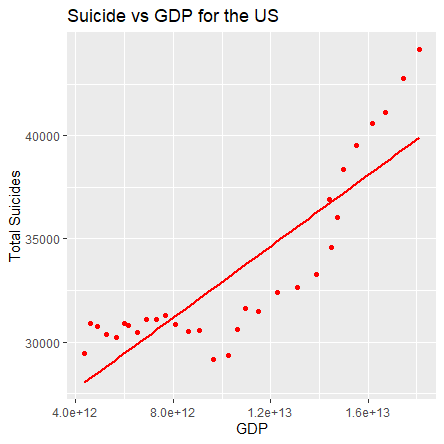
While it’s hard to say, GDP is one of the best measures of a country’s “world-level” that I know of. I’m no expert in this area so there are probably better metrics or perhaps the above statement is true only in some cases, but another measure that I would consider along the same lines would be the Human Development Index, or HDI. HDI intakes life expectancies, schooling and standard of living and creates a general measure based off of those inputs. In combination with GDP, they create a solid measurement of a country’s “world-level”. However, when I looked at the relationship between total suicides and HDI, the trend seen in GDP had dissipated almost completely.

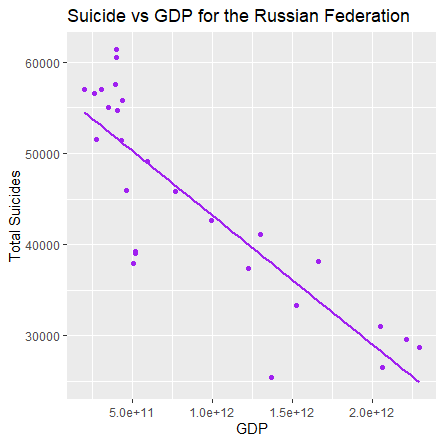


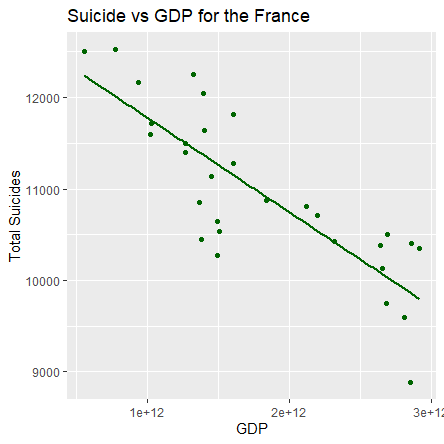
If we remove the few outliers in this plot that are above 20000, the relationship between the two variables is becomes extremely weak.



While the line may not be completely flat, I am hesitant to even say that there is a relationship here. Yes there are large values further to the right of the plot but there are so many values on the far right closer to zero that I am willing to bet that if all outliers were removed that you would have no relationship between the variables. That being said, the outliers are the exact problem that I had with this data and they are ultimately what lead me to my conclusions. It’s not just the outliers too, it’s that fact that when you look at different countries and the different trends relative to GDP vs Total Suicides the picture changes completely depending on the country that you are looking at. For example lets look at three different countries. The United States, The Russian Federation, and France.



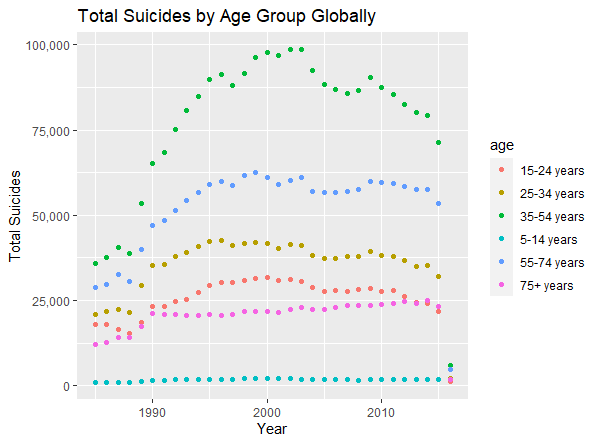




You’ll notice that the US is completely different from France and Russia. If you were to dig deeper and start plotting more of these you would see that they would vary on a case by case basis just as they do above. What’s more, is that you if look at the HDI vs Suicide plots for the respective countries above (Russian Federation does not have an HDI value) they follow the same trend as GDP, despite that trend weakening extremely when observed globally. Which leads me to my summary point for this question which was originally designed to find a global metric. I don’t think that there is such thing as a global metric for suicide, at least not from the data that I was working with. It varies so widely depending on the country that you are looking even if you had a value that was a fairly strong predictor globally, the question that I would ask, “What is the point if the individual variation by country is so high?” It seems that suicide is something that needs to be analyzed on a country by country basis for whatever reason that may be. With that said, there were a few categories that when aggregated painted a pretty glaring picture.

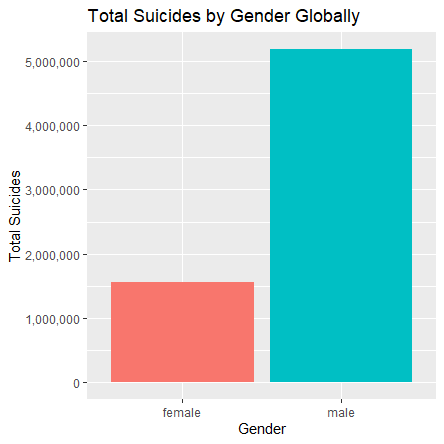
**Questions 3 & 4: Which age groups are at greater risk of suicide and how does gender influence suicide rate?**

The first grouping that I looked at was the grouping by age variables. I won’t go too far into it because it will be self-explanatory when you see the legend on the chart.



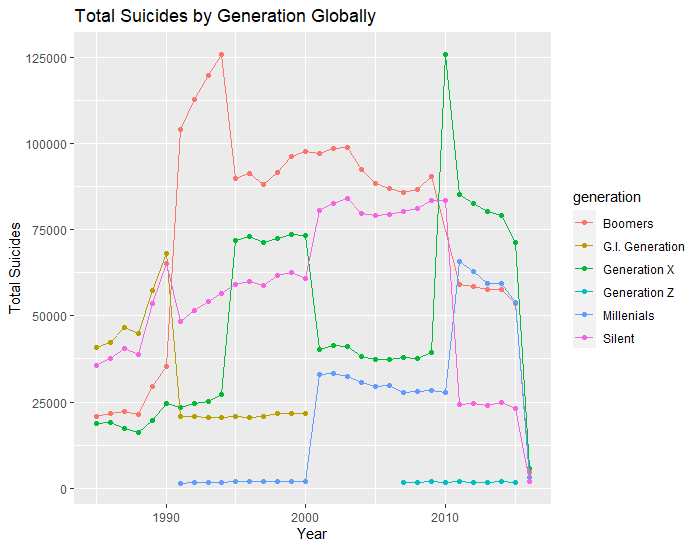
Interestingly enough there was a clear hierarchy established within the data that we can see. Before we get into it I wanted to point out a disclaimer with this data. It is aggregated from all countries and that means it’s possible that the outlier countries skewed this data to be this way and this trend may not be representative of the countries with lower amounts of suicides. I attempted to observe this on a country by country level but it was more or less impossible to read visually. Moving on to the data, I think it is a pretty clear answer for the question of which age groups are at greater risk. 35-54 commit the most suicides, as we’d expect and hope to be the case 5-14 year-olds commit the least and surprisingly enough the 15-24 age group is only the 3rd highest group. I expected them to be the highest.

Next was to observe suicides by gender on an aggregate scale just as I looked at age group. Fortunately enough this was also pretty clear and it was expected, because I had always heard that suicide affect men far greater than women. The same disclaimer for the age group variable applies here as well. This data was aggregated so it may not be representative for all countries in the data.



That being said, with the age data, some of the age groups were a lot closer together, it’s much more likely that the trend from the aggregate data is not the case for some countries, however in the case of gender this data is not even close. The male suicides are over 3x greater than the female suicides in the aggregate so it is much less likely that there are countries that will not follow this trend.

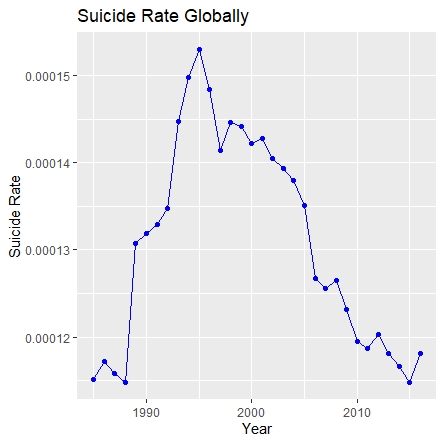
**Question 5: Is there a difference between generations for suicide?**



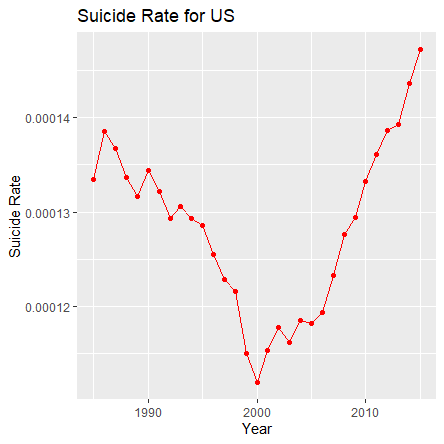
At first this graph is incredibly confusing, it seems that one thing that was true for all generations was that their suicide rates all spiked seemingly at random. Unfortunately, I don’t know what ages constitute these age groups and I don’t think any of them are set in stone either so it is hard to determine mathematically, but it seems that the spikes occur when each generation enters a different age group. I thought that there would be some pattern here but it is very difficult to discern one for the generation groups, what is interesting to note however is the spiking pattern is universal to all generations. At the very least this implies that the way the suicide permeates from generation to generation isn’t all too dissimilar.

**Question 6: What are the general trends for suicide?**

I wanted to ask this question originally for some peace of mind to think that things are getting better and globally that is resoundingly the case, but as per the motif of this project, global metrics are more or less irrelevant because it is completely different on a case by case basis. The following chart is the rate of suicide (total suicides/total population) globally.



Here we have some pretty good news it looks like, overall across the globe, suicide rates are decreasing. I used total\_suicides/total\_population here because total suicides doesn’t account for a change in population. Unfortunately, despite the global trend this is not the case at all in some places, namely the United States, which has been the largest outlier country in this dataset by a mile.



While not the “peace of mind” I was looking for it did confirm what was seen in almost all of the other charts where the US was the leading outlier.

**The implications of the analysis.**

While it’s a shame that the US data is the way that it is, it illustrates what I think to be the most important lesson that I have learned from this data and for anyone reading. Universal metrics are a cool idea but in practice, the amount of variation from group to group or in this case country to country is going to determine what measures are actually viable for modeling and prediction. Additionally, even if you could find a global metric that was somewhat viable I don’t understand what the point of it would be since in most cases you are going to be modeling on a more specific level anyway.

**Limitations of the Analysis**

There were a lot of limitations with problems that I encountered and things that I didn’t understand. Easily the most glaring problem that I encountered was my lack of knowledge on what to do what the data isn’t linear. I tried with great futility to transform the data in various ways but I wasn’t successful on finding one that would have been viable. I think in this case if there was someone more capable of dealing with data that isn’t linear or more experienced with how transformations work, I think they could have done a lot more with the data particularly with regards to building a linear model.

Another area of limitation that I found was my inability to produce a heat map to view suicide rates geographically. Given the implications of the analysis above, I’m not sure how useful this would have been but regardless it goes as a lead that was never checked.

Further, I think my lack of experience with dealing with outliers was majorly problematic in my analysis with this data. It seemed that the outliers were insanely obvious and also just radically large when compared with the larger body of the data. A lot of times I wanted to remove them and look at the data in a less skewed manner but I found myself constantly going back and forth and whether or not I should, because I know that dealing with them can be an issue of debate.

Lastly, I think from the perspective of results and output, my analysis on this data was fairly poorly done and in terms of the tools used. I realized that towards the end of this the majority of the work that I did was data slicing and visualizing relationships. I checked some correlation tables here and there but ultimately for the amount of tools that I had available I feel as if though I used almost none of them because I didn’t know where or how to use them.

**Conclusion**

Ultimately, I think an individual with more experience in the areas above could have really done some neat things with this data. With a greater skill set, a better understanding of what to do with non-linear data, and where to apply tools when necessary, I think there is far more to be discovered in this data. That being said, I feel like I was able to answer my questions fairly, and I learned an incredible amount for the seemingly little that I did.