Global Suicide

A Statistical Perspective

Jiatuan Luo, Conor Kennedy, John Montgomery
6414 Regression Analysis, Fall 2019
Dr. Nicoleta Serban
December 5, 2019

Summary

Every suicide is a tragedy and a mystery. The fact that anyone would want to take their own life is difficult to comprehend by those of us who have never entertained the thought. This study is an attempt to delve deeper, through statistical analysis, into what factors might contribute to someone's propensity to make that decision. We evaluate data that has been collected over a number of years from many countries in the world utilizing proven statistical methods that we are hopeful will shed additional light on causative effects.

Table of Contents

Reason for the Study

Suicide is a worldwide phenomena and occurs across all age groups. Gaining an understanding of the underlying causes allows health officials and policy makers to develop strategies that might reduce or minimize suicide in populations that are negatively impacted. Utilizing analytics to identify and test assumptions, reveal less than obvious factors, and predict outcomes based on analytical evaluation provides a scientific method to approach the problem and make decisions about how to best deploy available resources.

We have undertaken the effort to study the statistical underpinnings of suicide at a global level in an effort to better understand what might be driving the increase or decrease in suicide rates experienced in different parts of the world. Do environmental, economic, or demographic factors impact suicide rates globally? What factors can be controlled or influenced to have a positive effect on suicide rates? What approaches have been implemented that have proven effective or ineffective? What new approaches should be considered? The ultimate goal is to utilize data to help determine causation and provide guidance on where investments can be made to help reduce the loss of life and far reaching negative social and economic impact. It is estimated that nearly 1 million individuals are lost each year to suicide.

Suicide and Non-fatal Suicide Behavior have significant flow-on effects impacting the lives of any number of individuals—from family to friends, colleagues, clinicians, first responders, coronial staff, volunteers of bereavement support services, and other associates—who inevitably suffer intense and conflicted emotional distress in response to such behavior. Further, the economic impacts are significant and can include loss of productive capacity and earnings. For every suicide death in the U.S., 147 people are affected, and among those, 18 experience a major life disruption.[1] The financial cost of suicide in Australia is estimated to be $1.7M per suicide.[2] While these costs might not be representative across the globe, the social impact and financial cost is significant and provides strong impetus for financial investment in suicide prevention strategies.

---

[1] Cerel, J. (2015, April 18). We are all connected in suicidology: The continuum of "survivorship." Plenary presentation at the 48th annual conference of the American Association of Suicidology,Atlanta GA. [data from Cerel, Brown, Maple, Bush, van de Venne,Moore, & Flaherty, in progress]

[2] Kinchin, I., & Doran, C. M. (2017). The Economic Cost of Suicide and Non-Fatal Suicide Behavior in the Australian Workforce and the Potential Impact of a Workplace Suicide Prevention Strategy. *International journal of environmental research and public health*, *14*(4), 347. doi:10.3390/ijerph14040347

A Priori Expectations & Assumptions

Evaluating the data available, we have decided to limit the number of variables that we will use to perform our statistical analysis. The factors we have included are Dependent Variable - Suicide Rate (Suicides per 100K Population). Independent Variables - Country, Region, Age, Sex, socioeconomic factors such as Human Development Index (HDI) and Gini Coefficient as well as mental health factors represented in the Happiness Index.

We developed the following expectations before we began the statistical analyses process:

1. We expect countries on the lower end of the HDI to have higher suicide rates than well developed countries. Our initial thoughts are that populations that struggled with more 3rd world problems such as famine, hunger, disease, poor living conditions, etc. would be more susceptible to suicide.

2. We expect suicide rates to be higher for 15-24 and 75+ age groups with rates stabilized in between. Our thinking here is that the younger population is burdened with a plethora of challenges associated with the transition from childhood to young adult and would succumb more easily to suicidal thoughts. We also view the 75+ age group as having a higher probability due to advancing age, disease, financial challenges and other factors associated with aging.

3. We expect males to have a higher suicide rate than females. Males often choose more violent and thus effective methods of attempting suicide, and in those societies where men continue to be perceived as the breadwinner and provider for the family, any disruption in that role could have negative consequences.

4. We expect those countries/regions that rank higher on the Happiness Index to have a lower suicide rate. Our initial belief is that if a country's population is deemed to be happier through self-survey, then we would assume reduced mental health challenges and see a lower incidence of suicide.

5. We expect those countries that have a lower Gini Coefficient (higher wealth distribution) to have lower suicide rates. The broader the wealth distribution and smaller the gap between affluence and poverty would have a positive effect on suicide rate.

6. We expect countries that are more well off in terms of gdp per capita and life expectancy to have a lower suicide rate.

In the end, we will compare our analysis to these expectations to see if our initial assumptions hold.

Data Sources & Preparations

We acquired our data from various sources :

- Suicide rates dataset from Kaggle, which includes suicide rate, country, age, sex, region, etc.
- Happiness score that is based on surveys by the Gallup World Poll.
- Life expectancy, Gini Index from World bank.
- Household debt as a percent of GDP from tradingeconomics.com
- Median age from world.bymap.org

In areas where data was needed for both suicide rates and happiness level we developed a dataset where we had inner joined the two datasets on country and year. Because the only year for which the datasets overlapped and also had a usable amount of data was 2015, we only used this year for those areas of analysis.

The "World Region" data was manually coded, and the data we needed for our analysis was not available for all countries. This left some of our world regions with less countries than necessary for use in some of our analysis, which necessitated grouping some regions together. North America and Oceania were grouped with Western Europe and the Nordic States to create the region "Global West", the Baltic States and Central Asia were added to Eastern Europe, and Southeast Asia was added to Eastern Asia. Our rationale behind these groupings was trying to keep a cultural grouping at the expense of a regional one - Australia is much closer to Southeast Asia than Western Europe but it is culturally closer to the UK than it is to Thailand.

The dataset we used for regression analysis came from 6 different sources. Each country is an independent observation, and the goal was to explore the effect of socioeconomic factors on suicide rate. The data we used came from 2015. Time and effort were spent on data cleaning. Only rows with less than 3 missing values and columns with more than 70 percent of values were preserved. The country names on which the variables are joined together was conventionalized, and the missing data were imputed with values from other years, from similar observations, or column means if the other methods don't apply. In the end, we were able to arrive at 87 records to train our model on.

"Suicide Rates Overview 1985 to 2016"

https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

This compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio economic system.

Attributes:
country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, gdp_for_year, gdp_per_capita, generation (based on age grouping average).

"Human Development Index"
http://hdr.undp.org/en/content/human-development-index-hdi

The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.

The HDI is calculated by country. The range is 0 -1 with a higher value indicating higher achievement in human development.

"The World Happiness Index"

https://worldhappiness.report/ed/2019/

https://s3.amazonaws.com/happiness-report/2019/Chapter2OnlineData.xls

The World Happiness Report is a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be. 2019's World Happiness Report focuses on happiness and the community: how happiness has evolved over the past dozen years, with a focus on the technologies, social norms, conflicts and government policies that have driven those changes.

"The Gini Index"
https://data.worldbank.org/indicator/SI.POV.GINI

Gini index, or Gini ratio, is a measure of statistical dispersion intended to represent the income or wealth distribution of a nation's residents, and is the most commonly used measurement of inequality. 0 = perfect equality, 1 = maximum inequality

"Household Debt as a % of GDP"
https://tradingeconomics.com/country-list/households-debt-to-gdp

Statistical Analyses

Multiple Linear Regression:

A main focus of our project is to better inform health officials and policy makers in an effort to help them make fact based decisions regarding resource deployment in order to lower suicide rates globally. So, we selected socioeconomic indicators as variables to examine what are the factors that drive suicide rates on a macroeconomic scale.

To explore the effect of our predictors on suicide rate, we first performed multiple linear regression. The distribution of suicide rates appear to be right-skewed, so we tried different transformations on the response variable and discovered that a square root transformation can lead to a distribution that most resembles a normal distribution. (Figure 1.)

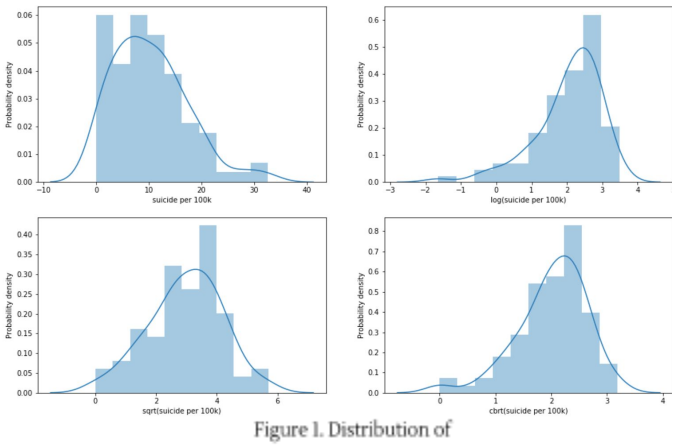


Figure 1. Distribution of suicide/100k Before & After transformations

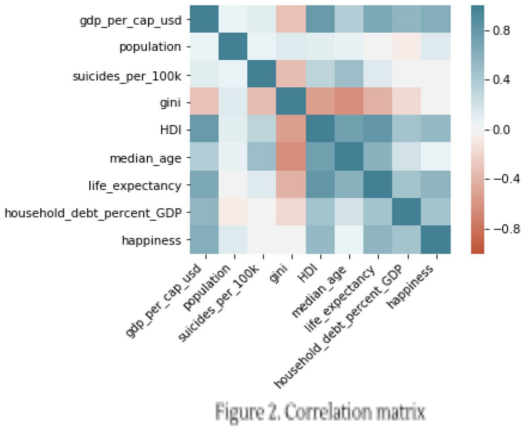


Figure 2. Correlation matrix

Next, we ran the full model on the eight predictors and computed the variance inflation factor for each coefficient. Neither the VIF nor the correlation matrix signaled a problem of multicollinearity. (Figure 2). The result of the model points to three statistically significant variables at 90% confidence - median age, happiness, and life expectancy with an adjusted coefficient of determination of 0.2759. To reduce model complexity, we performed stepwise regression and regularized regression. The variables selected are as below:

	Backward step <chr>	Forward step <chr>	Lasso <chr>	Elastic Net <chr>
gdp_per_cap_usd				
population				
gini				x
HDI				
median_age	x	x	x	x
life_expectancy	x	x		
household_debt_percent_GDP				
happiness	x	x		

Figure 3. Variable Selection

We then ran two reduced models for variables selected in stepwise regression and Elastic Net. We were able to get statistically significant results for all three variables from stepwise regression and only for median age from Elastic Net. However, the model possesses minimal explanatory power with adjusted coefficient of determinations of 0.3081 and 0.2663 respectively. Despite the low variance explained, we were able to maintain or even increase R^2 while reducing model complexity through variable selection. The result of the first submodel shows that median age (0.13712) and happiness scores (0.33840) have a positive effect on the response, and life expectancy (-0.10425) has a negative effect. In the second model, the coefficient for median age is 0.092773 with a p-value close to 0, while the Gini Coefficient does not have a statistically significant effect on the response.

To assess the goodness-of-fit, we performed residual analysis for both submodels (Figure 4.) The results show good fits and no clear violations of the assumptions of multiple linear regression.

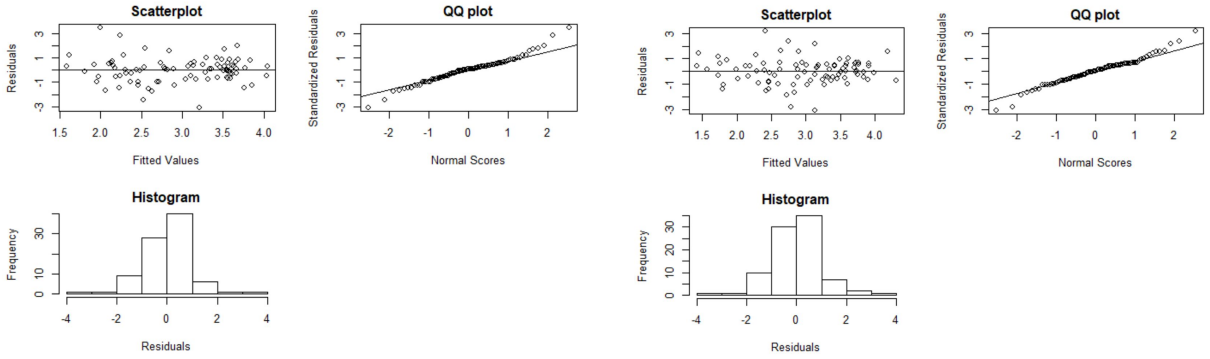


Figure 4. Residual Analysis For LM Models

Poisson Regression:

We were not overly satisfied with the results we got from multiple linear regression partly due to the exclusion of the majority of variables and the low explanatory power. We subsequently performed Poisson regression, which we believed would be better suited for modeling suicide rate. Since suicide rates may follow a Poisson distribution as opposed to normal.

To perform Poisson regression, we used the total number of suicide rather than suicides per 100k as the response variable. And we added an offset to the model to use population as an exposure variable. In this case, we are modeling the log of suicide rate of a country, with population being the unit.

We performed Poisson regression with the predictors GDP per capita, Gini index, HDI, median age, life expectancy household debt as a percent of GDP, and happiness score. To our surprise, in contrast to the results from multiple linear regression, all 7 predictors were shown to be statistically significant with near 0 p-values (Figure 5.). We also performed stepwise regression with AIC for the model, and in both forward and backward selection, the full model was selected.

```
Call:
glm(formula = suicide_no ~ gdp_per_cap_usd + gini + HDI + median_age +
    life_expectancy + household_debt_percent_GDP + happiness +
    offset(log(population)), family = "poisson", data = glmdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-57.707   -7.683   -0.879    7.670   67.536

Coefficients:
                             Estimate Std. Error  z value Pr(>|z|)
(Intercept)                -9.291e+00  7.595e-02 -122.327  < 2e-16 ***
gdp_per_cap_usd            -6.309e-06  3.143e-07  -20.074  < 2e-16 ***
gini                       -2.339e-02  4.574e-04  -51.145  < 2e-16 ***
HDI                         5.687e+00  1.089e-01   52.236  < 2e-16 ***
median_age                  3.382e-02  7.996e-04   42.295  < 2e-16 ***
life_expectancy            -6.542e-02  8.658e-04  -75.563  < 2e-16 ***
household_debt_percent_GDP  4.171e-04  1.270e-04    3.284  0.00102 **
happiness                   3.448e-02  4.611e-03    7.477 7.63e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74577  on 86  degrees of freedom
Residual deviance: 29409  on 79  degrees of freedom
AIC: 30102

Number of Fisher Scoring iterations: 5
```

Figure 5. Summary Poisson Regression

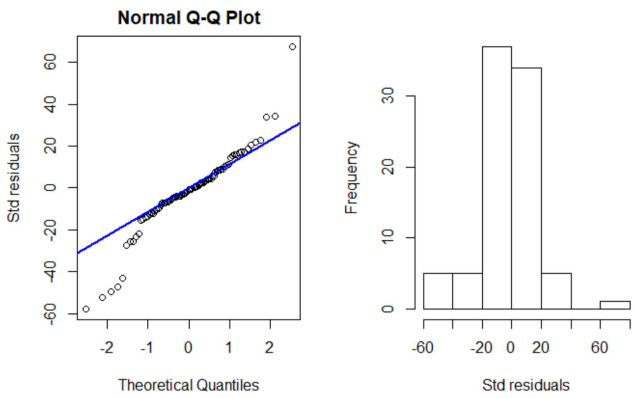


Figure 6. Distribution of Deviance Residual

We subsequently performed goodness-of-fit tests using both deviance and Pearson residuals, it turns the null hypothesis of good fit is rejected in both cases. We guessed that the variability in our data is too large to satisfy the mean equals variance assumption of Poisson distribution. And according to the dispersion parameter we computed, it is indeed an overdispersed model, with a value many times larger than 2. Besides, it appears that the deviance residual does not adhere very well to normal distribution, with most values aggregating towards the middle and curvatures on the two sides of the Q-Q plot. (Figure.6)

We noticed the inadequacy of our model and conducted further research on how to deal with the problem of overdispersion. Some viable techniques we discovered include using the quasi-family to augment the normal family by adding a dispersion parameter, also known as Quasi-Poisson; try different distributions such as negative binomial; or using observation-level random effects which allows the expectation to vary more than a Poisson distribution would suggest. (From *Overdispersion, and how to deal with it in R and JAGS* by Carsten F. Dormann)

```
Call:
glm(formula = suicide_no ~ gdp_per_cap_usd + gini + HDI + median_age +
    life_expectancy + household_debt_percent_GDP + happiness +
    offset(log(population)), family = quasipoisson, data = glmdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-57.707   -7.683   -0.879    7.670   67.536

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -9.291e+00  1.445e+00  -6.432 8.91e-09 ***
gdp_per_cap_usd            -6.309e-06  5.978e-06  -1.055 0.294416
gini                       -2.339e-02  8.699e-03  -2.689 0.008735 **
HDI                         5.687e+00  2.071e+00   2.747 0.007456 **
median_age                  3.382e-02  1.521e-02   2.224 0.029014 *
life_expectancy            -6.542e-02  1.647e-02  -3.973 0.000156 ***
household_debt_percent_GDP  4.171e-04  2.416e-03   0.173 0.863368
happiness                   3.448e-02  8.770e-02   0.393 0.695295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 361.7171)

    Null deviance: 74577  on 86  degrees of freedom
Residual deviance: 29409  on 79  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

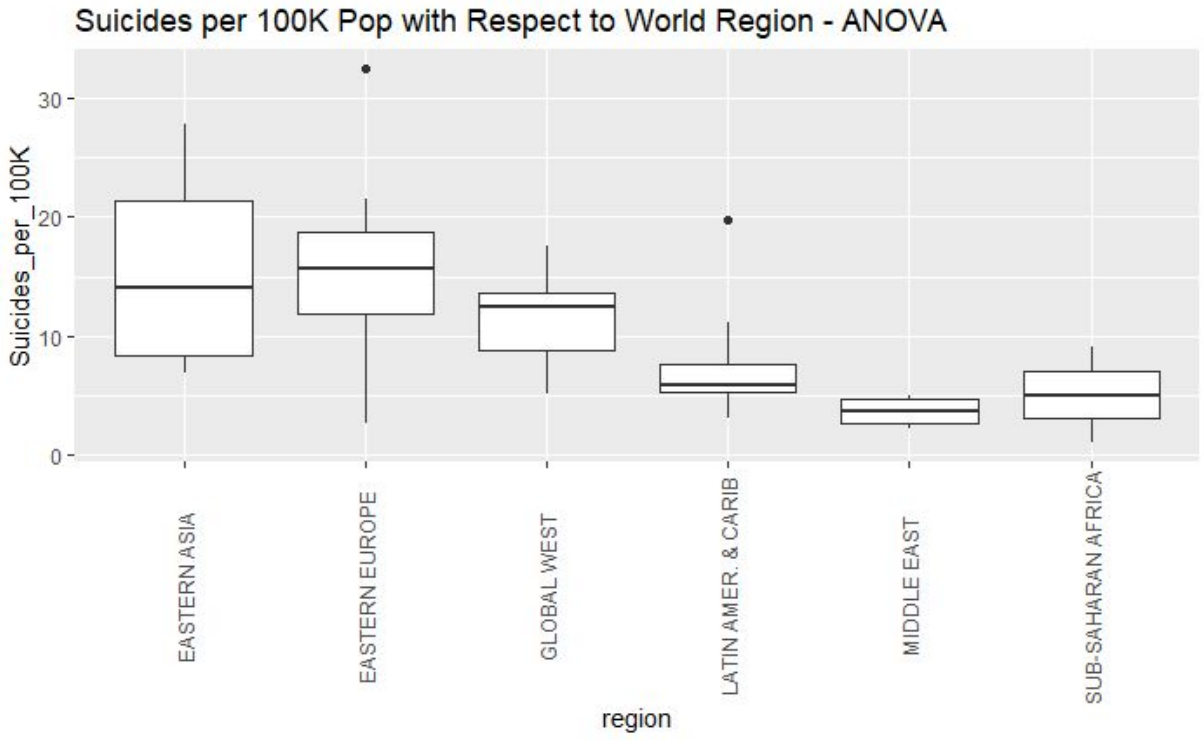
Figure 7. Summary Quasi-Poisson

In the end, we adopted Quasi-Poisson and ended up with vastly different coefficients and significance. (Figure 7.) In our new model, the Dispersion parameter for Quasi-poisson family was taken to be 361.7171, and the result showed that the response log(suicide rate) had a statistically significant positive relationship with median age (3.382e-02) and HDI (5.687) and a negative relationship with Gini index (-2.339e-02) and life expectancy (-6.542e-02). This indicates that countries that are higher developed or with higher median age are more likely to have high suicide rates, while low life expectancy and high equality (low Gini index) are also associated with high suicide rates. Both household debt and happiness are positively correlated with suicide rate, but not on a statistically significant level. The interpretations are under the condition that all other variables are fixed constant.

Analysis of Variance (ANOVA):

The first analysis of variance we conducted was determining whether the regions of the world had statistically significantly different means from each other. Our null hypothesis was that the regions did not have significantly different means at a 90% confidence interval. The result was that, after performing an analysis of variance on Suicides per 100K Population with respect to World Region, we saw a p- value of 0.001, indicating that at least one region's mean was significantly different from the others. A

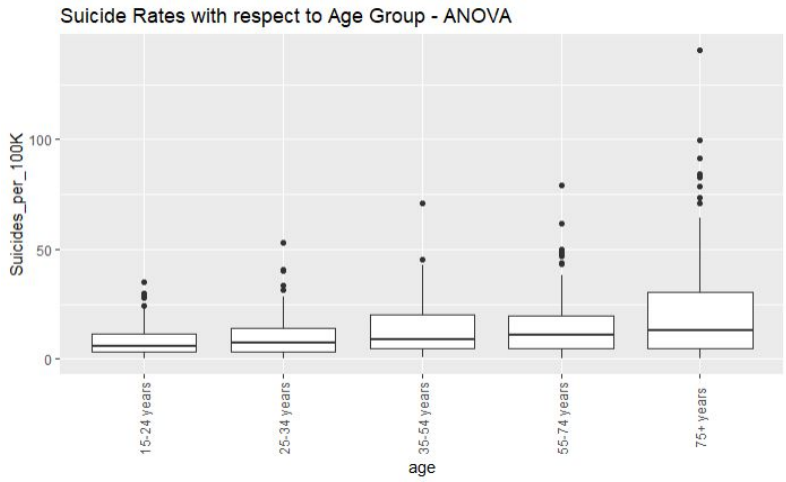
graph of this relationship can be seen below.



We performed a Tukey test on the results of the ANOVA to get the pairwise comparisons of the regions, and received the result that the only two regions which had statistically significantly different means at the 90% confidence level were the Middle East and Eastern Europe. The expected difference between the two was that the Middle East has 11.66 less suicides per 100K population than Eastern Europe, with lower and upper bounds of -21.14 and -2.17 respectively. The p-value associated with this pairwise comparison was 0.008.

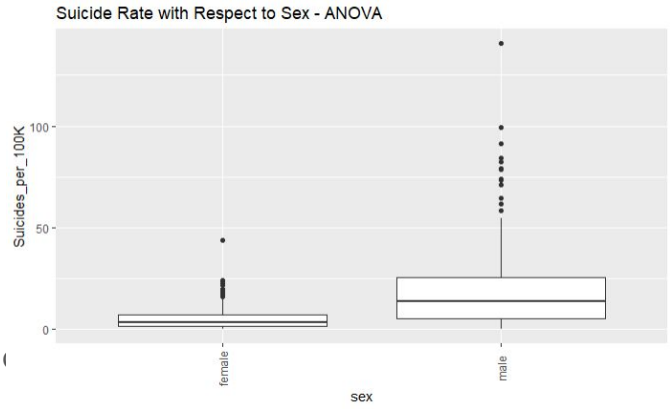
We would expect that reality differs from the results of this particular analysis, as we don't have data on many countries from the Middle East and this may be artificially depressing its number of suicides per 100K population. Given that the only significant pairwise comparison was that between Eastern Europe and the Middle East, while it may appear from absolute values of suicides per 100K population that regions of the world differ, it may be that none differ significantly from a statistical perspective at the 90% confidence interval. It would be possible to ascertain these numbers piecemeal to fill out the category, but the risk of mixing sources on these numbers which are somewhat sensitive and thus prone to manipulation outweighs the benefit of a more complete dataset.

The second analysis of variance we conducted was determining whether the different age groups represented in our data had statistically significantly different means for Suicides per 100K Population. The result was that, after performing an analysis of variance on Suicides per 100K Population with respect to age group, we saw a p-value that approached zero. However, after looking at a graph of the different age groups with respect to Suicides per 100K population, we realized that the 5-14 age group was skewing the results as it was abnormally low as compared to the other age groups. This makes sense, as most of those years are early childhood - a time where most individuals are unlikely to be mentally developed enough to understand suicide, much less take action towards it. We dropped this age group and re-ran the ANOVA, and while the p-value was not approaching zero as before, it was still very small (3e-10) indicating strongly that at least one age group's mean differed significantly at the 90% confidence level from anothers. A graph of this relationship can be seen below.



We performed a Tukey test on the results of the ANOVA to get the pairwise comparisons of the age groups and received the result that, at the 90% confidence level, the 75+ age group had a significantly higher average number of suicides per 100K population than every other age group, and that with higher p-values but still significant at the 90% confidence level the 35-54 and 55-74 age groups had higher average suicides per 100K population than the 15-24 age group.

This differed slightly from our expectations, as we expected higher rates of suicide amongst the 15-24 age group. This difference in expectations can be explained by the much higher visibility of suicides of young people as compared to other age groups. As the adage goes - a single death is a tragedy, but a million deaths is a statistic.



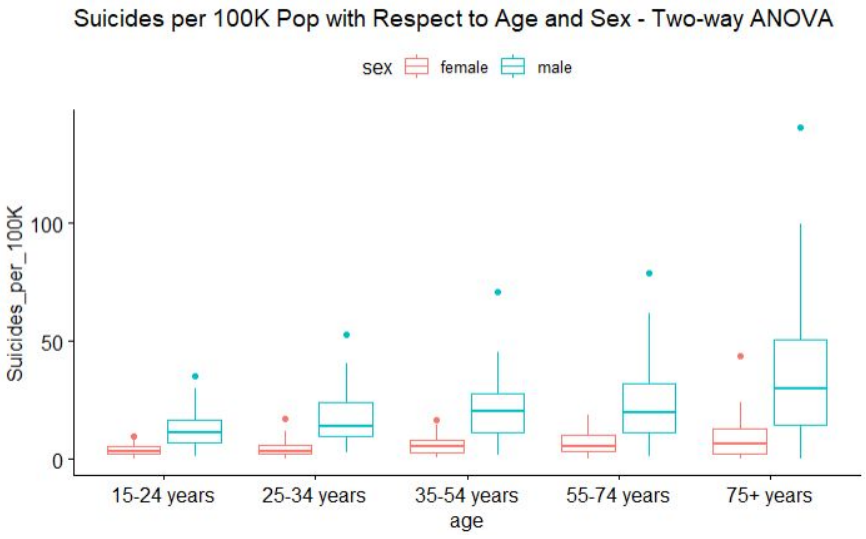
The third analysis of variance we conducted was determining whether the different sexes had statistically significantly different levels of Suicides per 100K Population. The result of the ANOVA was that the two means were

different, with the p-value of the ANOVA approaching zero. The resulting graph can be seen above.

Males as a group were found to have an average of 13.48 more suicides per 100K population than females, with lower and upper bounds of 11.71 and 15.26 respectively.

The results of these last two analyses of variance, that between age and that between sex, made us question the relationship between age and sex with respect to suicides per 100K population. Males and older people were found to have higher rates of suicide, but it was not clear whether male and female suicide rates both increased with time nor whether one sex was more heavily contributing to a particular age group's suicide rate. We conducted a two-way ANOVA using age and sex as the categorical variables, the results are visualized below.

The results of this were very interesting, as it appeared that males had a higher suicide rate across all age groups, but that while the average suicides per 100K for males went up over time they appeared fairly static for women - though the variance for both sexes visibly increased.



We performed a Tukey test on the results of this two-way ANOVA to determine which age-group/sex combos had statistically significantly different means from each other at the 90% confidence level. The result was that nearly all male age groups had higher average suicides per 100K population than all female age groups. The exceptions were that males age 15-24 did not have significantly higher suicides per 100K population than any female age group, and that males age 25-34 did not have higher suicides per 100K population than females age 75+. For all other combinations of sex and age group, males had higher suicides per 100K population than females. Additionally, it was found that no female age group had a statistically higher number of suicides per 100K population than any other female age group.

Findings & Implications

In reviewing our initial expectations compared to statistical outcomes, we made the following observations:

1) Underdeveloped countries do not demonstrate a higher probability for suicide. In fact, those countries considered to be "3rd World" actually exhibited a lower mean when compared to countries with higher Human Development Indexes. This brings into question certain biases we, as a group, may have harbored in our initial assessment due to our countries of origin. This also indicates that there are factors inherent in populations connected to low HDI that should be evaluated to gain clearer insight into how they might contribute to lower suicide rates.
2) Evaluating our initial expectation of seeing a higher rate of suicide among 15-24 year old age group, we found that this did not hold true. In fact, the mean for this age group was actually lower and the variance narrower than the other age groups. Our expectation of seeing higher suicide rates among the 75+ age group was confirmed with a higher average mean and broader variance than any of the other groups. Our initial expectations appear to have been impacted by the prevalence of media exposure to younger suicides and the bias that exposure creates.
3) Our expectation that males have a higher suicide rate than females was confirmed through our analysis and was consistent across the globe. Our analysis did not explore the potential reasons behind this but begs for additional scrutiny to determine and leverage the source of this gap.
4) Although the models gave us mixed results on the significance of happiness score (significant in multiple linear, not in Quasi-Poisson), the coefficient of happiness is positive across the board. In this case, our hypothesis that the Happiness Index is negatively correlated with suicide rate was not confirmed, which lead us to infer that being in an environment surrounded by people who are happy will lead to more instances of suicide. This is an interesting finding that we would be eager to explore in the future.
5) Countries that are deemed to have lower Gini rates experience lower suicide rates. The models also gave us mixed results on the significance of Gini rates. (insignificant in multiple linear but significant in Quasi-Poisson) However both indicated negative correlation and therefore lead us to the idea that low income disparity (low Gini index) is associated with high suicide rate. Which is a rather counterintuitive finding.
6) In this case, we coupled GDP per capita and life expectancy to assume higher and longer equated to lower probability of suicide. We noticed that life expectancy did

have a negative effect on rates but GDP per Capita did not necessarily influence probability.

Conclusions & Further Questions Raised

It seems that suicide is a more complicated matter than we assumed prior to the study. There is not a definite answer as to what social-economic factors drives the suicide rate of a country, as models seem to give us mixed results, and the data also comes with a variety of limitations in terms of quality and size. It seems like there is more noise than a clear pattern that we could identify. However, our ANOVA tests have indeed discovered a significant difference between suicide occurrence in different demographic groups. With all that in mind, here are the topics that we propose to dig deeper to further our understanding of the issue:

1. Would the results of our region-based analysis of variance be the same if we had reliable suicide data for all countries?
2. What other factors can we explore that affect suicide rates? (e.g Does a country's religiosity impact suicide rate?)
3. Assuming our result from Quasi-Poisson is reliable, what attributes do populations in lower Human Development countries possess that contribute to lower suicide rates?
4. What is different between male and females that leads to such a drastic difference in the means and variances of their suicide rates?
5. What leads to the pronouced disparity in suicide rates between different age groups?
6. How to determine the overall economic impact suicide has on a country leading to justification for resource allocation that drives a postive economic outcome?
7. Is there a difference in the reporting of suicide cases between countries that leads to inaccuracy in suicide rates of certain countries? (e.g. Third world countries' data might be significantly deflated due to the lack of reliable infrastructures that collect such data.)