Global Suicide Mortality: A Statistical Analysis of Socioeconomic Markers on Suicide Counts

William-Elijah Clark

Department of Statistics

Florida State University

Tallahassee, FL 32306

STA 5167

Dr. Xu-Feng Niu

May 02, 2024

I: Introduction

Background Information

The term "*Deaths of Despair*" is a relatively newer term from the field of Economics, coined by Economists Anne Case and Angus Deaton in 2015 to describe the phenomenon of working-class Americans without a bachelor's degree dying from liver disease, drug overdoses, and/or suicides. This concept is explored in "*Rising Morbidity and Mortality in Midlife Among White Non-Hispanic Americans in the 21st Century*", "*Mortality and Morbidity in the 21st Century*", and "*Deaths of Despair and the Future of Capitalism*" *(Case and Deaton 2015, 2017, 2020).*

However, it should be noted that in the past nine years as of this writing, this concept has been criticized for varying reasons. These criticisms include but are not limited to whether it is due to economic policy disagreements *(Henderson 2020)*, how heterogenous this phenomenon has historically been with respect to racial/ethnic groups *(Zheng and Choi 2024; Friedman et al. 2023),* or methodological differences in using the Consumer Price Index (CPI) as opposed to the Personal Consumption Expenditures Price Index (PCE) for inflation calculations (*The Economist*, 2023; Johnson N., 2017). It can also be noted that the features of the phenomenon may have changed in the context of the COVID-19 pandemic (*Entrup et al. 2023*), and there are reports of narrowing suicide racial disparities within the United States *(Johnson S. 2024, Gold 2020).* Even further, research has been limited in scope to the United States and other high-income, developed nations within the scope of the original paper describing the "*Deaths of Despair*" phenomenon. *(Case and Deaton 2015).*

As made apparent, more quantitative analysis can be done on this topic to see if the concept of Deaths of Despair holds globally: research on this scale is even newer than research on the scale of the United States, with some analysis being done between the years 2000-2019 *(Ilic and Ilic 2022).* Further, determining whether certain socioeconomic predictors are statistically related to suicide mortality is useful. It should be noted that certain psychiatric/medical phenomena can also be tested in conjunction with economic factors (e.g., the Gender Paradox of Suicide (*Schrijvers et al. 2012, Tucker 2020*), the general rarity of child suicide, particularly before the 2010s (*Kingkade and Chuck 2021, Asarnow n.d.*), and regional differences in suicide trends (*Ilic and Ilic 2022*).

Data Description

Within this analysis, global data involving suicide rates and socioeconomic markers will be used, as compiled from both the World Health Organization and World Bank by Ronald Onyango on Kaggle *(Onyango, 2024)* in the *"suicide_rates_1990-2022.csv"* datafile of *n=118580* observations, with 18 variables. By analyzing global data, it should be possible to ascertain whether there is statistically significant evidence for low-income conditions being correlated with higher suicide rates and/or counts on an even broader macroeconomic scale.

Table 1: Variables in data in "suicide_rates_1990-2022.csv"

Variable	***Variable Type***	***Description***
RegionCode	*Categorical*	*Code for region (AF, AS, CSA, EU, NAC, OA)*
RegionName	*Categorical*	*Full name for region*
CountryCode	*Categorical*	*Code for country (101 nations total)*
CountryName	*Categorical*	*Full name for county*
Year	*Numeric*	*Year data was collected (1990-2022)*
Sex	*Categorical*	*Sex demographic for death count (M, F, Unknown)*
AgeGroup	*Categorical*	*Age demographic for death count (0-14, 15-24, 25-34, 35-54, 55-74, 75+)*
Generation	*Categorical*	*Generational label demographic*
SuicideCount	*Numeric*	*Number of Suicide Deaths Recorded*
CauseSpecificDeathPercentage	*Numeric*	*Percentage of deaths attributed to suicide*
DeathRatePer100K	*Numeric*	*Death Rate per 100K people in country*
Population	*Numeric*	*Population of country in given year*
GDP	*Numeric*	*Gross Domestic Product of country in USD in given year*
GDPPerCapita	*Numeric*	*Gross Domestic Product per capita country in USD in given year (GDP/Population)*
GrossNationalIncome	*Numeric*	*Gross National Income of country in USD in given year*
GNIPerCapita	*Numeric*	*Gross National Income per capita country in USD in given year (GNI/Population)*
InflationRate	*Numeric*	*Annual increase in prices per year (CPI or PCE not specified in source)*
EmploymentPopulationRatio	*Numeric*	*Percentage of population 15+ employed within country.*

The "Generation" variable from the original data file will be ignored in this paper due to being anachronistic (e.g., the data describes someone who was between the ages of 0-14 in the 1990s as "Generation Alpha". That demographic is defined as being born in the early 2010s at the earliest). It may also be noted that Region Code/Name and Country Code/Name are identical for analysis.

Research Questions and Objectives

This data is somewhat limited in that it cannot be used for some questions mentioned in the background. First, it cannot be used in context of drug overdose and/or alcoholic liver disease mortality rates. Secondly, it cannot be used to make *direct* inferences about specific racial groups, given that the data is divided by region and country only. Third, at the time of this writing, the type of inflation values used have not been specified by Onyango. Hence, the question of whether using CPI or PCE for inflation and determining if these two inflation measures change analysis cannot be ascertained here. However, there are still worthwhile research questions that can be formed in context of this data:

1. Which categorical and/or quantitative variables are correlated and/or statistically significant for suicide counts and rates, if at all?
2. Do the documented phenomena and conclusions of other researchers hold in context of the above-cited data?

Hence, this paper aims to answer these two questions to see if the economic concept of Deaths of Despair holds with this data sourced from the World Health Organization and World Bank, combined by Onyango in 2024.

II: Data Preprocessing

Data Cleaning and Imputation

As mentioned in the Data Description, the "*Generation*" variable from Onyango's data has been ignored in this analysis. However, additional data cleaning and imputation are required before proceeding to any analysis.

All *NA* variables for *any* variable were subsequently dropped from the final data, save for unknown gender. Further, all precise duplicates within the data were removed. However, the data still presented additional problems in that some rows from the original data file were near identical duplicates, save for SuicideCount, CauseSpecificDeathPercentage, and DeathRatePer100K.

Figure 1: Dataframe filtered by Sex, AgeGroup, and CountryName with two rows reporting two different SuicideCount values for the same year.

CountryName	Year	Sex	AgeGroup	Generation	SuicideCount
United States of America	1990	Male	25-34 years	Millennials	2667
United States of America	1990	Male	25-34 years	Millennials	2672
United States of America	1991	Male	25-34 years	Millennials	2563
United States of America	1991	Male	25-34 years	Millennials	2798
United States of America	1992	Male	25-34 years	Millennials	2435
United States of America	1992	Male	25-34 years	Millennials	2667
United States of America	1993	Male	25-34 years	Millennials	2500
United States of America	1993	Male	25-34 years	Millennials	2718
United States of America	1994	Male	25-34 years	Millennials	2555
United States of America	1994	Male	25-34 years	Millennials	2740

As having two conflicting reports for Suicide Counts for the same year, gender, age group, and country would create issues in calculation via artificially inflating the sample size, this was solved by imputing between the two rows for each combination of groups via the dplyr library in R.

Figure 2: Dataframe of the same demographic from Figure 1 where SuicideCount has been imputed as an average of multiple rows that share the same characteristics.

ct_MasterScript.R* | US_Deaths_Male_2534 | Death | unalived

Filter

erCapita	InflationRate	EmploymentPopulationRatio	SuicideCount	CauseSp
24100	4.2349640	60.343	2680.5	
25090	3.0288197	60.119	2551.0	
25910	2.9516570	60.412	2609.0	
27250	2.6074416	61.220	2647.5	
28450	2.8054197	61.613	2617.0	
29870	2.9312042	61.899	2424.0	

From a realistic perspective, it is not possible to have a fraction of a death (someone has either passed away by suicide or they have not, and this would generally be a Boolean or integer). However, it is realistic to have a scenario where varying government agencies within a nation or state where a centralized database with cohesive mortality and mental health information does not exist. Rather, there could be multiple, conflicting data points across several different government agencies, such as within the State of Florida (*Florida Commission on Mental Health & Substance Abuse 2001, Ogozalek 2023*). It thereby follows that imputing an approximate value when a nation is reporting multiple different numbers for the same years and demographics may be warranted, as no exact answer may be available across multiple reports.

An additional consideration regarding the data is the existence of zero values within the columns for *SuicideCount, CauseSpecificDeathPercentage*, and *DeathRatePer100K*. For instance, the five number summary for *SuicideCount* is min=0, Q1=1, Q2=9.25, Q3=45, and max=5584.5. This is also the case with *CauseSpecificDeathPercentage (*min=0, Q1= 0.2637654, Q2=1.6810092, Q3=7.4058928, and max=100), and *DeathRatePer100K (*min=0, Q1= 0.6396541, Q2=5.6412280, Q3=15.8183089, and max=210.0229319). Given that there are zero values for several entries, there are two possible assumptions one could make regarding those values.

The first potential assumption that could be made is that these are naturally occurring zeros. This would make sense in context of the age 0-14 factor within *AgeGroup*, as suicide in very young children is particularly rare. (*Kingkade and Chuck 2021, Asarnow n.d.*). It would thereby follow that eliminating these datapoints would be erroneous, as we would essentially fail to account for an entire age cohort.

The second potential assumption is that these zeros are a failure to report. This is also a realistic possibility, as mentioned in Ilic and Ilic's "*Worldwide Suicide Mortality Trends (2000-2019): A Joinpoint Regression Analysis*". In particular, the possibility of under-reporting or even non-reporting (especially from developing countries), variability of data quality from various countries beyond the above-mentioned issue of multiple national agencies, ill-defined causes of death, and scenarios where a non-suicide is documented as a suicide or *vice-versa* (*Ilic and Ilic 2022*). While the data bias for positive values may not be eliminated, bias from zero values that may be a result of under-reporting can be dropped.

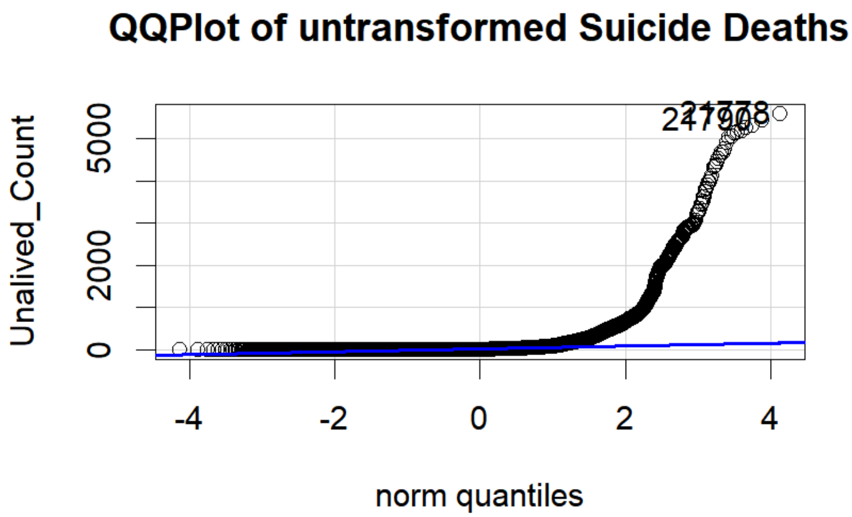
In practice, it would be difficult to determine exactly which zero counts would be a result of under-reporting and which ones would be from naturally occurring zeros. Hence, from this point, an analysis will be applied to two scenarios and referred to as such:

- *Scenario 1*, where the death reports of zero are assumed to be naturally occurring zeros. Zero death counts will be kept for this analysis. After cleaning, this leaves a final dataset of n=28187 for analysis.
- *Scenario 2*, where the death reports are assumed to be a consequence of under-reporting or non-reporting. Zero death counts will be dropped for this analysis. After cleaning, this leaves a final dataset of n=24661 for analysis.

Data Transformations

Given the skewed distribution of our three death variables, as seen in the five number summaries and via the following QQ-Plot with a nearly flat reference line, it follows that a variable transformation of some kind would be warranted to normalize the data as much as possible.

Figure 3: A QQ-Plot for untransformed SuicideCount



However, it should be noted that in Scenario 1, attempts at finding a Box-Cox transformation will be impossible without further modification. For $\lambda \leq 0$, log(0) or any iteration of 1/0 will subsequently give an undefined answer. Additionally, it will be impossible for a computer to proceed with those calculations in this case. However, the existence of log(x+c) transformations, while problematic (Muldoon 2018), also provides a workaround for Box-Cox functions in general. Particularly, any $0<c\leq 1$ value for x+c should work to ensure that a Box-Cox transformation function does not result in any undefined values. All attempted Box-Cox functions run with *SuicideCount* were thus run with (*SuicideCount*+0.0001) as the actual variable. In the context of this paper, c=0.0001 was chosen to be as close to zero as possible within four decimal spaces. Furthermore, the resulting transformations should be a close approximation to transformations needed for Scenario 2 without having to use a constant.

Figure 4: Lambda value graph for GNI and Population

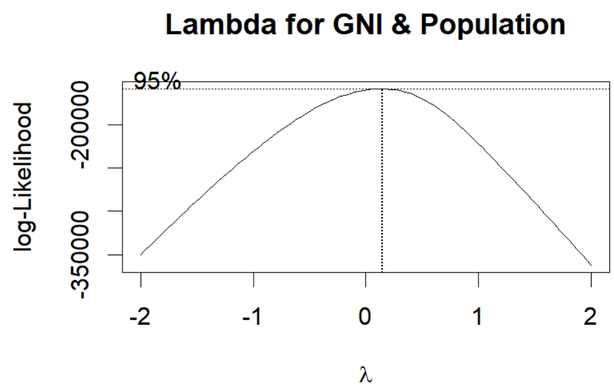


Figure 5: Lambda value graph for GDP and Population

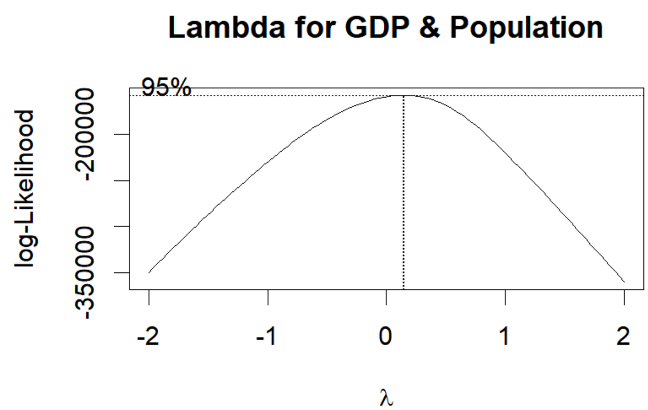
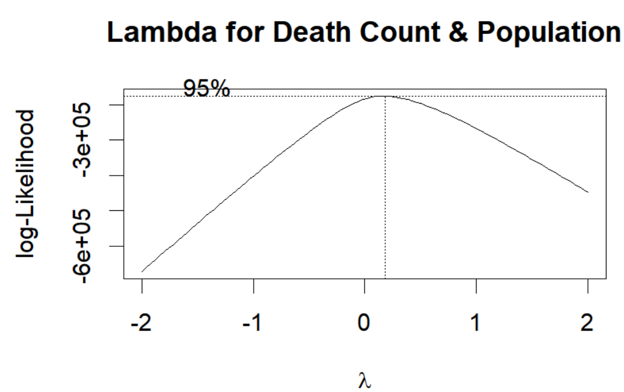


Figure 6: Lambda value graph for SuicideCount (referred to as Death Count) and Population



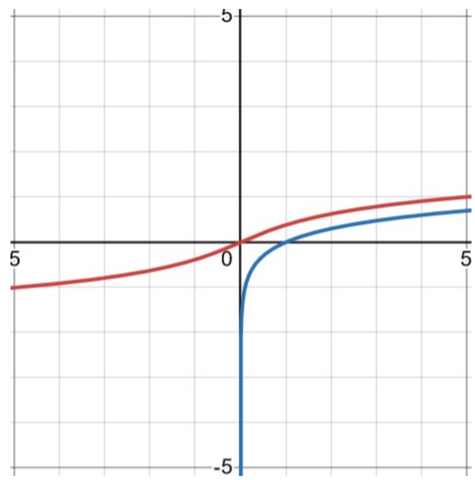
While it should be noted that the above graphs produce values close to λ=0, the precise values included λ =0.1414141, λ=0.1414141, and λ =0.1818182. Inverting the response and predictor variables produced similar λ ≈0.1 values. For the sake of simplicity, a log transformation of log(x+c) would be viable here.

In a context where log(x+c) transformations work, there also exists an alternative transformation referred to as the Inverse Hyperbolic Sine (IHS) transformation that functions similarly to a log transformation (*Aihounton and Henningsen 2021, Norton 2022*). The IHS transformation is defined as follows:

$$IHS(x) = arcsinh(x) \approx log(x + \sqrt{x^2 + 1})$$

The above function has the benefit of having the general shape of the log function for all values x≥0.

Figure 7: A comparison of the log(x) transformation (in blue) and the approximated IHS(x) transformation (in red)



However, the IHS transformation is not without drawbacks. First, it can overestimate responses in comparison to a log transformation. There are also additional drawbacks addressed by David McKenzie in "Interpreting Treatment Effects on an Inverse Hyperbolic Sine Outcome Variable and Alternatives". Informally, it is mentioned that explaining this in the context of discussions with policymakers can be difficult. (*McKenzie, 2023; Norton, 2022*). Further, there are formal problems to address with IHS transformations. Particularly, the transformation of zero-valued outcomes presents an issue where the three following assumptions about the data or any subsequent model cannot be simultaneously true:

- The resulting response variable is an average of individual-level treatment effects.
- The model is invariant to the scaling of the response.
- The model point-identified from the marginal distributions of the response.

In brief, it is assumed that the model is *not* normally distributed. Given that this has already been established with the untransformed version of the *SuicideCount* variable with respect to the QQ-Plot, it would make sense to proceed under the assumption that at least one of the three assumptions are violated in proceeding analysis. Furthermore, to quote econometrician David Card "…experiments with alternative functional forms (such as log(citations+1) or the inverse hyperbolic sine function) […] are quite robust", indicating that the ISH transformations still have the potential to yield some useful findings for regression models. (*Card and DellaVigna 2013*).

Regardless of the above issues with the log(x+x) method or the IHS method, both methods yielded QQ-Plots for SuicideCount that have a less flat reference line.

Figure 8: QQ Plot for the distribution of IHS(SuicideCount)

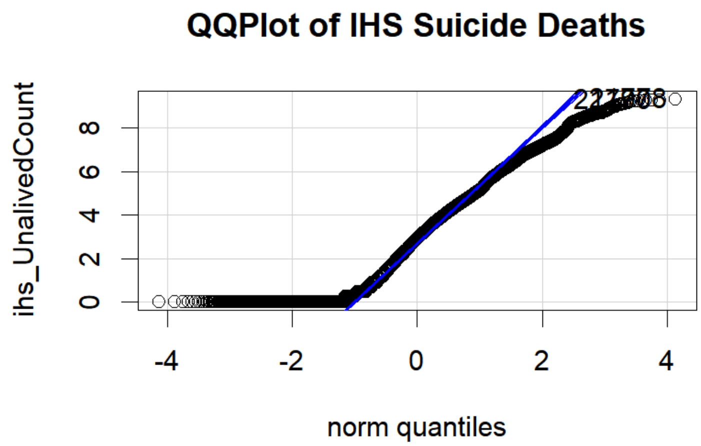
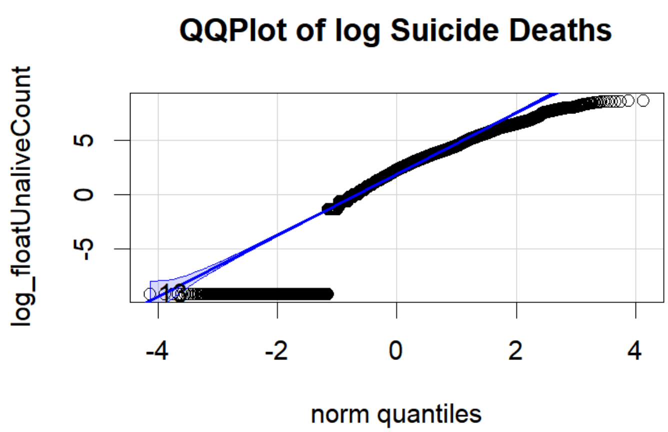


Figure 9: QQ Plot for the distribution of log(SuicideCount+.0001)



III: Statistical Methodology and Models

Exploration of Potential Variable Correlations for Numeric Variables

To find potential predictor variables that would have any correlation or statistical significance concerning *SuicideCount, CauseSpecificDeathPercentage, or DeathRatePer100K* involved, correlation matrices were made to discern which transformed and untransformed variables may be correlated to the three potential Suicide response variables.

Figure 10: Correlation Matrix for untransformed data from Scenario 1

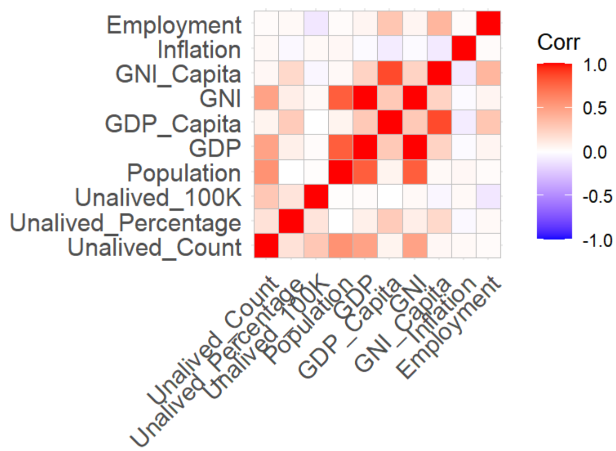
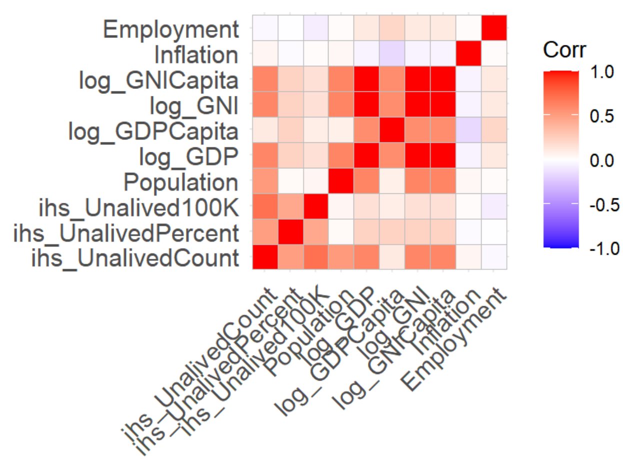


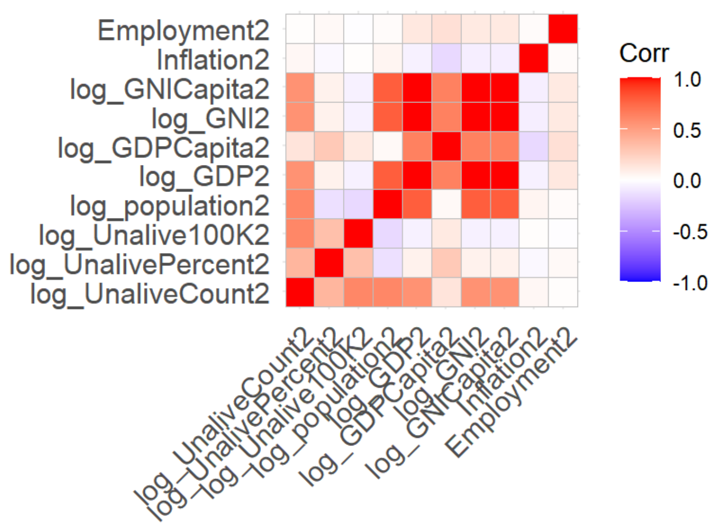
Figure 11: Correlation Matrix for transformed data from Scenario 1



As seen in both Figure 8 and 9, there are stronger correlations for transformed data than untransformed data, justifying the transformations established via Figures 4, 5, and 6. Furthermore, it appears that the *SuicideCount* variable shows stronger correlations than the *CauseSpecificDeathPercentage, or DeathRatePer100K* variables, indicating that for modeling purposes, using the response variable of *IHS(SuicideCount)* would be the best choice for regression modeling over IHS(*CauseSpecificDeathPercentage), or IHS(DeathRatePer100K)*. Similar correlation matrices occur for Scenario 2, where all zero-counts for deaths are dropped due to assumed under-reporting or bias.

Figure 12: Correlation Matrix for untransformed data from Scenario 2

Figure 13: Correlation Matrix for transformed data from Scenario 2



It should be noted that the correlation between *log(SuicideCount)* and other economic predictors such as log(GDP) and log(GNI) is even more stark in comparison to the correlation with *log(CauseSpecificDeathPercentage)* or *log(DeathRatePer100K)* when zero values are removed, although the general correlations do not change much beyond log(Population) having some negative correlation to *log(CauseSpecificDeathPercentage)* and *log(DeathRatePer100K)*.

As an additional note, *GDP, GNI, GDPPerCapita,* and *GNIPerCapita* will always be correlated and have issues with multicollinearity due to how these economic metrics are calculated. Specifically:

- *GDP = Consumption + (Government Expenditures)+ Investment + (Net Exports)*
- *GNI = GDP + ((Money from Foreign Countries) – (Money to Foreign Countries))*

Further, *GDPPerCapita* and *GNIPerCapita* will have multicollinearity with *Population*. It can also be noted that whether GDP or GNI is a better economic predictor is contextual (e.g., over time *("National income - Gross national income - OECD Data" n.d.),* or whether a country has a notable amount of foreign investment/aid (*The Investopedia Team 2024*)). Hence, testing for both will be useful.

<u>*ANOVA Models*</u>

The first model type to determine whether *SuicideCount, CauseSpecificDeathPercentage, or DeathRatePer100K* were correlated to any categorical variables involved one-way, two-way, and three-way ANOVA modeling for *AgeGroup*, *Sex*, and *Region* to see if the gender paradox of suicide, regional differences, and age group differences were present within the data. Untransformed, log(x+c), and IHS-transformed versions of *SuicideCount* were modeled. It can be noted that the untransformed data and both transformations of SuicideCount all result in p<2e-16 values, whether on their own or in interaction models. The main difference between the transformed and untransformed response variables were the exact values for Sums of Squares, Mean Squares, and F-values.

For the sake of visual clarity and brevity, the IHS transformation was used for plotting below one way interaction models below, and will

The one-way model for *AgeGroup* tested was $Y_{IHS(SuicideCount)} = \mu_{Age\_Group} + \varepsilon_i$, with the hypothesis of $H_0: \mu_1 = \cdots = \mu_6$, versus $H_1: Not\ all\ \mu_i\ equal$. With p<2e-16, we may reject the null hypothesis that the means between age groups are the same in favor of the alternative hypothesis that the means between age groups are different at an α=0.001 [99.9%] confidence level. Further, TukeyHSD was also performed. The null hypothesis that almost all pairs of age groups are the same can be rejected in favor of the alternative hypothesis that these groups are different at the α=0.001 [99.9%] confidence level. (The only null hypothesis that we fail to reject at any confidence level is for the 25-34 and 35-54 age group pairing)

The one-way model for *Region* was $Y_{IHS(SuicideCount)} = \mu_{Region} + \varepsilon_j$, with the hypothesis of $H_0: \mu_1 = \cdots = \mu_6$, versus $H_1: Not\ all\ \mu_j\ equal$. With p<2e-16, we may reject the null hypothesis that the means between ag e groups are the same in favor of the alternative hypothesis that the means between age groups are different at an α=0.001 [99.9%] confidence level. Further, TukeyHSD was also performed. The null hypothesis that almost all pairs of region groups are the same can be rejected in favor of the alternative hypothesis that these groups a re different at the α=0.001 [99.9%] confidence level. (The only null hypothesis that we fail to reject at any confi dence level is for the North America and the Caribbean-Africa region pairing). One unusual finding was that thi s ANOVA model found fewer deaths in Africa, contrary to Ilia and Ilia's analysis.

The one-way model for Sex was $Y_{IHS(SuicideCount)} = \mu_{Sex} + \varepsilon_k$, with the hypothesis of $H_0: \mu_{Female} = \mu_{Male} = \mu_{Unknown}$, versus $H_1: Not\ all\ \mu_k\ equal$. With p<2e-16, we may reject the null hypothesis that the means between Female, Male, and Unknown gender deaths are the same in favor of the alternative hypothesis that the means between these three groups have differing death rates at an α=0.001 [99.9%] confidence level.With p<2e-16, we may reject the null hypothesis that the means between Female, Male, and "Unknown" gender deaths are the same in favor of the alternative hypothesis that the means between these three groups have differing death rates at an α=0.001 [99.9%] confidence level.

Figure 14: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu_{Age\_Group} + \varepsilon_i$ Model

```
> age_deathcount3 <- aov(ihs_UnalivedCount ~ Factored_AgeGroup)
> summary(age_deathcount3)
                     Df Sum Sq Mean Sq F value Pr(>F)
Factored_AgeGroup     6  54114    9019    2613 <2e-16 ***
Residuals         41197 142179       3
```

Figure 15: ANOVA Plot for $Y_{IHS(SuicideCount)} = \mu_{Age\_Group} + \varepsilon_i$ Model

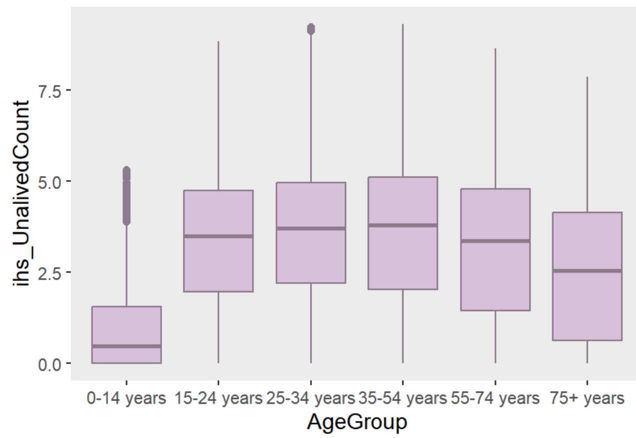


Figure 16: TukeyHSD p-value adj output for $Y_{IHS(SuicideCount)} = \mu_{Age\_Group} + \varepsilon_i$ Model

```
                                p adj
15-24 years-0-14 years  0.0000000
25-34 years-0-14 years  0.0000000
35-54 years-0-14 years  0.0000000
55-74 years-0-14 years  0.0000000
75+ years-0-14 years    0.0000000
Unknown-0-14 years      0.0000000
25-34 years-15-24 years 0.0000007
35-54 years-15-24 years 0.0000001
55-74 years-15-24 years 0.0000021
75+ years-15-24 years   0.0000000
Unknown-15-24 years     0.0000000
35-54 years-25-34 years 0.9999141
55-74 years-25-34 years 0.0000000
75+ years-25-34 years   0.0000000
Unknown-25-34 years     0.0000000
55-74 years-35-54 years 0.0000000
75+ years-35-54 years   0.0000000
Unknown-35-54 years     0.0000000
75+ years-55-74 years   0.0000000
Unknown-55-74 years     0.0000000
Unknown-75+ years       0.0000000
```

Figure 17: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu_{Region} + \varepsilon_j$ *Model*

```
> region_deathcount3 <- aov(ihs_UnalivedCount ~ RegionName)
> summary(region_deathcount3)
               Df Sum Sq Mean Sq F value Pr(>F)
RegionName      5  13250  2650.0   596.4 <2e-16 ***
Residuals   41198 183044     4.4
```

Figure 18: ANOVA Plot for $Y_{IHS(SuicideCount)} = \mu_{Region} + \varepsilon_j$ *Model*

Figure 19: TukeyHSD p-value adj output for $Y_{IHS(SuicideCount)} = \mu_{Age\_Group} + \varepsilon_j$ *Model*

```
                                                         p adj
Asia-Africa                                          0.0000000
Central and South America-Africa                     0.0000000
Europe-Africa                                        0.0000000
North America and the Caribbean-Africa               0.0733345
Oceania-Africa                                       0.0000000
Central and South America-Asia                       0.0022774
Europe-Asia                                          0.0000000
North America and the Caribbean-Asia                 0.0000000
Oceania-Asia                                         0.0000002
Europe-Central and South America                     0.0000000
North America and the Caribbean-Central and South America 0.0000000
Oceania-Central and South America                    0.0012009
North America and the Caribbean-Europe               0.0000000
Oceania-Europe                                       0.0064714
Oceania-North America and the Caribbean              0.0000000
```

Figure 20: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu_{Gender} + \varepsilon_k$ *Model*

```
> gender_deathcount3 <- aov(ihs_UnalivedCount ~ Gender)
> summary(gender_deathcount3)
                Df Sum Sq Mean Sq F value Pr(>F)    
Gender           2  10386    5193    1151 <2e-16 ***
Residuals   41201 185907       5                   
---
```

Figure 21: ANOVA Plot for $Y_{IHS(SuicideCount)} = \mu_{Gender} + \varepsilon_k$ *Model*

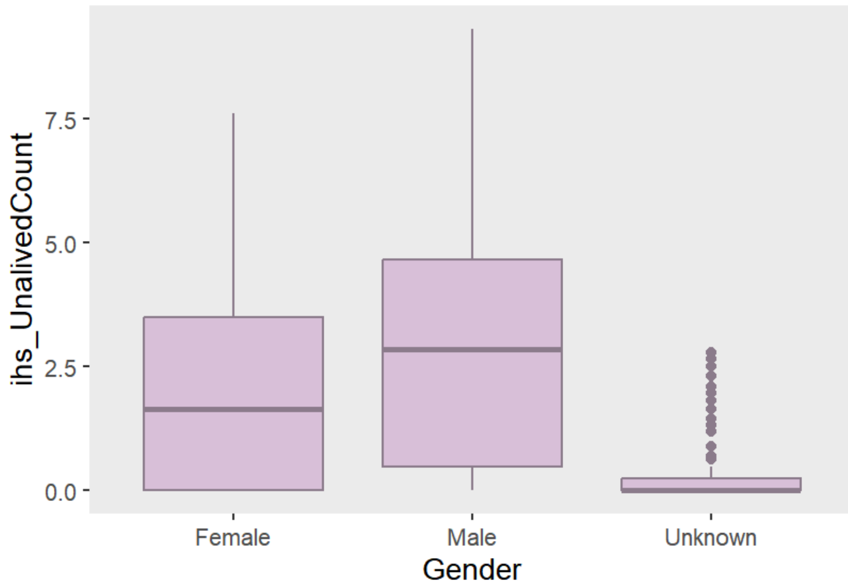


Figure 22: TukeyHSD p-value adj output for $Y_{IHS(SuicideCount)} = \mu_{Gender} + \varepsilon_k$ *Model*

```
> TukeyHSD(gender_deathcount3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ihs_UnalivedCount ~ Gender)

$Gender
                     diff        lwr        upr p adj
Male-Female     0.8427692  0.7933122  0.8922261     0
Unknown-Female -1.7751280 -1.9703349 -1.5799210     0
Unknown-Male   -2.6178971 -2.8131040 -2.4226902     0
```

Interaction terms for three two-way ANOVA models and one three-way ANOVA model were also statistically significant at an at an α=0.001 [99.9%] confidence level. Hence, we can also reject the null hypothesis that the means between interaction terms are not different in favor of the alternative hypothesis that they are different.

Figure 23: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu + \alpha_{Gender} + \beta_{AgeGroup} + (\alpha\beta)_{AgeGroup:Gender} + \varepsilon_{ijk}$ *Model*

```
> ihsagegender_deathcount <- aov(ihs_UnalivedCount ~ Factored_AgeGr
oup * Gender)
> summary(ihsagegender_deathcount)
                             Df Sum Sq Mean Sq F value Pr(>F)
Factored_AgeGroup             6  54114    9019 2874.91 <2e-16
Gender                        2  10392    5196 1656.30 <2e-16
Factored_AgeGroup:Gender     12   2589     216   68.76 <2e-16
Residuals                 41183 129198       3

Factored_AgeGroup        ***
Gender                   ***
Factored_AgeGroup:Gender ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 24: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu + \alpha_{AgeGroup} + \beta_{Region} + (\alpha\beta)_{AgeGroup:Region} + \varepsilon_{ijk}$ *Model*

```
> ihsregionage_deathcount <- aov(ihs_UnalivedCount ~ Factored_Regio
nName * Factored_AgeGroup)
> summary(ihsregionage_deathcount)
                                        Df Sum Sq Mean Sq
Factored_RegionName                      5  13250    2650
Factored_AgeGroup                        6  53931    8989
Factored_RegionName:Factored_AgeGroup   30   8902     297
Residuals                            41162 120210       3
                                     F value Pr(>F)
Factored_RegionName                    907.4 <2e-16 ***
Factored_AgeGroup                     3077.8 <2e-16 ***
Factored_RegionName:Factored_AgeGroup  101.6 <2e-16 ***
Residuals
```

Figure 25: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu + \alpha_{Gender} + \beta_{Region} + (\alpha\beta)_{Gender:Region} + \varepsilon_{ijk}$ *Model*

```
> ihsgenderregion_deathcount <- aov(ihs_UnalivedCount ~ Gender *Fac
tored_RegionName)
> summary(ihsgenderregion_deathcount)
                               Df Sum Sq Mean Sq F value Pr(>F)
Gender                          2  10386    5193 1238.65 <2e-16
Factored_RegionName             5  12814    2563  611.28 <2e-16
Gender:Factored_RegionName      9    412      46   10.92 <2e-16
Residuals                   41187 172681       4

Gender                     ***
Factored_RegionName        ***
Gender:Factored_RegionName ***
Residuals
```

Figure 26: ANOVA Table for $Y_{IHS(SuicideCount)} = \mu + \alpha_{Gender} + \beta_{Region} + (\alpha\beta)_{Gender:Region} + (\alpha\gamma)_{Gender:AgeGroup} + (\beta\gamma)_{Region:AgeGroup} + (\alpha\beta\gamma)_{Gender:Region:AgeGroup} + \varepsilon_{ijkt}$ *Model*

```
> IHS_RegionAgeGend_deathcount <- aov(ihs_UnalivedCount ~ Factored_RegionName * Factored_Gende
r * Factored_AgeGroup)
> summary(IHS_RegionAgeGend_deathcount)
                                                       Df Sum Sq Mean Sq  F value
Factored_RegionName                                     5   6632    1326  437.058
Factored_Gender                                         1   6441    6441 2122.262
Factored_AgeGroup                                       5  24851    4970 1637.754
Factored_RegionName:Factored_Gender                     5    176      35   11.589
Factored_RegionName:Factored_AgeGroup                  25   5273     211   69.503
Factored_Gender:Factored_AgeGroup                       5    946     189   62.350
Factored_RegionName:Factored_Gender:Factored_AgeGroup  25    207       8    2.728
Residuals                                           28115  85323       3
                                                      Pr(>F)
Factored_RegionName                                    < 2e-16 ***
Factored_Gender                                        < 2e-16 ***
Factored_AgeGroup                                      < 2e-16 ***
Factored_RegionName:Factored_Gender                   3.31e-11 ***
Factored_RegionName:Factored_AgeGroup                  < 2e-16 ***
Factored_Gender:Factored_AgeGroup                      < 2e-16 ***
Factored_RegionName:Factored_Gender:Factored_AgeGroup 7.25e-06 ***
Residuals
```

Stepwise Regression Models

Within the data, there are three potential response variables *SuicideCount, CauseSpecificDeathPercentage,* and *DeathRatePer100K*. Further, transformations of those three response variables are also available. Additionally, *AgeGroup, Gender, RegionName, Year, log(GDP), log(GDPCapita), log(GNI), log(GNICapita), log(population), Inflation,* and *Employment* were all potential predictors. Due to potential multi-collinearity issues previously mentioned, as well as 32 years being problematic for regression analysis, 48 separate stepwise regressions were computed to determine which predictor variables (if any) were statistically significant in the context of one of our three death variables.

One issue with the stepwise regression models is that the process would attempt to remove all predictors.

Figure 27: Stepwise Regression removing all predictors in a model, with increasing AIC for each predictor variable added for Scenario 1

```
> stepwise_10 <- step(stepwise_10, direction = "both")
Start:  AIC=4954.09
ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored
_RegionName +
    log_population + log_GDP + Inflation + Employment

                      Df Sum of Sq   RSS     AIC
<none>                             33565  4954.1
- Employment           1       4.0 33569  4955.5
- Inflation            1      37.1 33602  4983.2
- log_GDP              1     239.4 33804  5152.4
- Factored_RegionName  5    4290.6 37856  8334.8
- Factored_Gender      1    6442.7 40008  9901.4
- log_population       1   12164.0 45729 13668.9
- Factored_AgeGroup    5   24848.5 58414 20561.5
```

However, it should be noted that R^2 and adjusted R^2 are not very different from each other for the full model, despite the above-mentioned issue with stepwise regression eliminating all predictors. Furthermore, in some of these models, all p-values for predictors had a p-value ≤ 0.1. For at least $\alpha=0.1$, we would be able to reject the null hypothesis of $H_0: E(Y|X_i = X_1 + \cdots + X_n) = \beta_0, i = 1, 2, \ldots, n$ in favor of the alternative hypothesis of $H_1: E(Y|X_i = X_1 + \cdots + X_n) = \beta_0 + \cdots + \beta_n U_n, i = 1, 2, \ldots, n$.

Figure 28: Model for Scenario 1 where R^2 is 0.7415 and Adjusted R^2 is 0.7414

```
Factored_RegionNameOceania                              < 2e-16 ***
log_population                                          < 2e-16 ***
log_GDP                                                 < 2e-16 ***
Inflation                                              2.41e-08 ***
Employment                                               0.0666 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.092 on 28171 degrees of freedom
Multiple R-squared:  0.7415,    Adjusted R-squared:  0.7414
F-statistic:  5387 on 15 and 28171 DF,  p-value: < 2.2e-16
```

Figure 29: Model for Scenario 1 where R^2 is 0.716 and Adjusted R^2 is 0.7159

```
log_population                                          < 2e-16 ***
log_GDP                                                 < 2e-16 ***
Inflation                                              3.09e-11 ***
Employment                                             2.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.173 on 34297 degrees of freedom
  (7181 observations deleted due to missingness)
Multiple R-squared:  0.716,     Adjusted R-squared:  0.7159
F-statistic:  5087 on 17 and 34297 DF,  p-value: < 2.2e-16
```

Figure 30: Model for Scenario 2 where R^2 is 0.7134 and Adjusted R^2 is 0.7132

```
log_GDP2                                           8.747  < 2e-16 ***
log_population2                                  103.193  < 2e-16 ***
Inflation2                                         4.144 3.43e-05 ***
Employment2                                        4.999 5.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.123 on 26357 degrees of freedom
  (4248 observations deleted due to missingness)
Multiple R-squared:  0.7134,    Adjusted R-squared:  0.7132
F-statistic:  3859 on 17 and 26357 DF,  p-value: < 2.2e-16
```

It should be noted that the above models where $R^2>0.7$ were specifically for Scenario 1 where zero-values are assumed to be naturally occurring and the response variable is IHS(SuicideCount). For Scenario 2, where zero values are dropped and log(SuicideCount) is used, R^2 goes down significantly.

Hence, the following model is proposed to account for demographic dummy variables:

$$IHS(\widehat{SuicideCount}) = -.01301 + 2.409U_{Age15-24} + 2.615U_{Age25-34} + 2.638U_{Age35-54} + 2.240U_{Age55-74} + 1.642U_{Age75+} - 0.518U_{GenderUnkown} + 0.8715U_{GenderMale} + 0.924U_{RegionAsia} + 1.151U_{RegionCentral/SouthAmerica} + 1.616U_{RegionEurope} + 1.40U_{RegionNorthAmerica/Carribean} + 1.440U_{RegionOceania} + 0.6806(log(Population)) + 0.04964(log(GDP)) + 0.0002729(Inflation) + 0.004(Employment) + \varepsilon$$

IV: Conclusions

<u>*Main Research Questions*</u>

For research question one, it appears that all categorical and quantitative variables appeared to be related to SuicideCounts. In all ANOVA models, *sex*, *age*, and *region* had statistically significant impacts on suicide counts, regardless of the transformation method used. Further, for IHS transformations used in regression models, *log(GDP), log(GNI), population, employment*, and *inflation* were correlated with *IHS(SuicideCount)*. Depending on models, employment was or not statistically significant at α=0.05, but was consistently significant at α=0.1. However, inflation, log(GDP), and log(GNI) were consistently significant, indicating that these predictors may be more relevant for predicting suicides. It can also be noted that either log(GDP) or log(GNI) can be used with minimal differences.

Hence, for research question two, there does appear to be credence to the *Deaths of Despair* concept outlined by Case and Deaton. Further, the data analyzed indicates that the Gender Paradox of Suicide, the general rarity of child suicide, and regional differences in suicide trends were all present. However, this analysis indicates that Africa had a *lower* death rate as opposed to a higher death rate, in comparison to Ilia and Ilia's research. Questions about whether zero counts earlier in time are related to this phenomenon would be worthwhile to address, given differing time frames.

<u>*Further Questions: IHS and Natural Zero Assumptions*</u>

Models that used IHS transformations under the assumption that zero data entries for *SuicideCount* were organically occurring events consistently had better R^2 values than models with *log(SuicideCount+0.001)* transformations, all other factors remained the same. Further, *log(SuicideCount+0.001)* transformations where zero values were kept had lower R^2 and adjusted R^2 values than an equivalent *log(SuicideCount)* model where zero-values were removed. Most notably of all, the IHS transformation applied to Scenario 1 produced similar R^2 values to log transforms applied to Scenario 2.

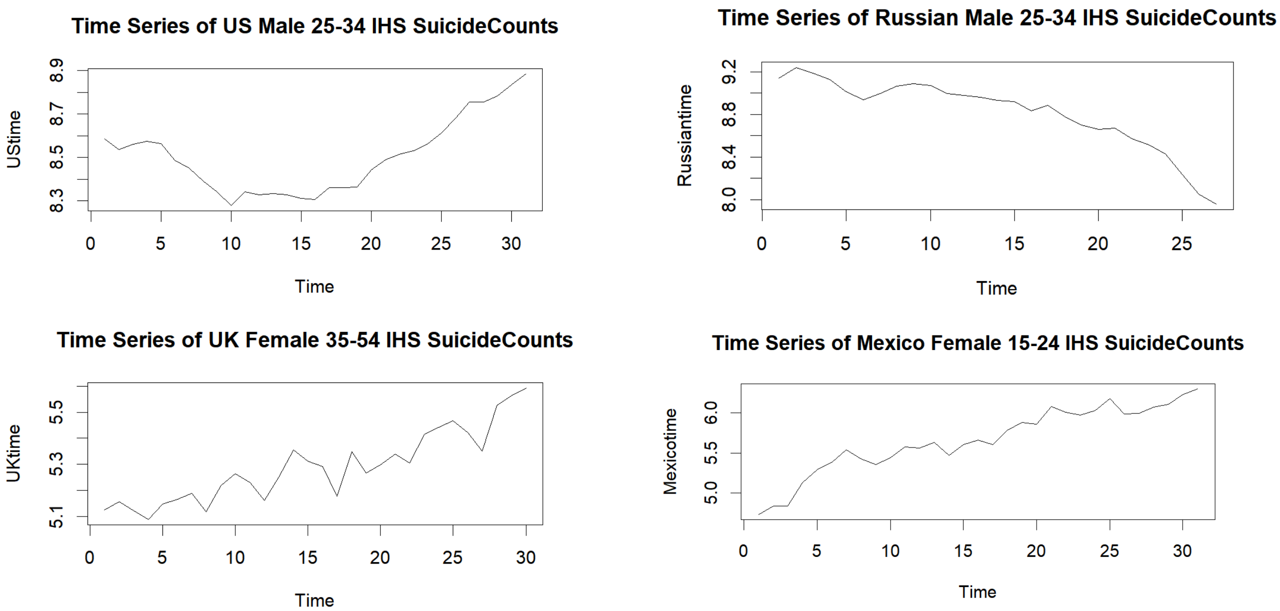
For practicality, it would be difficult to determine which individual zero datapoints are naturally occurring and which are due to a failure to report. However, the fact that these assumptions change which transformation works well is worth further exploration.

<u>*Further Questions: Time Series Feasibility and Graphs*</u>

With some of the stepwise regression models where factored(Years) were added, some factored years were statistically significant at varying points and varying confidence levels, particularly with years>2000. However, it should be noted that suicide mortality reporting is not consistent across countries, with some countries not reporting until further in time than the 1990s. Furthermore, only 101 countries were represented in the original dataset (most notably, India was missing despite being one of the most populated countries on the globe). It is generally advised to have at least n=50 observations for a time series analysis (Box et al. 2008). However, it is theoretically possible to do ARIMA with sample sizes n=15, n=25, and n=35. The main consideration is that there will be an increase in prediction errors, which would need to be accounted for…but should be reasonably sufficient in the context of these data being unavailable past 1990. *(Hassouna and Al-Sahili 2020).*

Hence, it may be possible to do this for a select number of countries, to see whether there is a trend related to economic downturns (e.g., the 2001, 2008, and 2020 recessions may be related to increased suicide mortality in the United States, and could be accounted for as intervention events). However, it should also be mentioned that differing countries may have other intervention events that would need to be modeled differently, based on context (e.g., the Russo-Ukrainian war may be correlated to a decrease in suicide death counts for Russian men between 25-34 due to an increase in military deaths.) Further data cleaning and filtering would need to be performed to determine which year has *SuicideCount* data for all countries present in a region to implement ARIMA models for assorted regions.

Figures 31a, 31b, 31c, 31d: Time Series Plots of US 25-34 Men, Russian 25-34 Men, UK 35-54 Women, and Mexican 15-24 Women.



Further Questions: Additional Variables for Future Research

Given that the dataset used for this project may or may not use CPI or PCE inflation as its *Inflation* variable, this would be useful to test for further research. Additional variables to include might also involve GINI coefficients for each country to account for income inequality or the Human Development Index to account for the average quality of life in a country. These values are generally available for most countries. Within the United States, newer underemployment metrics could also be used to see what correlation those have to suicide rates.

V: Appendix A – Citations

Aihounton, G. B. D., and Henningsen, A. (2021), "Units of measurement and the inverse hyperbolic sine transformation," *The Econometrics Journal*, 24, 334–351. https://doi.org/10.1093/ectj/utaa032.

Asarnow, J. (n.d.). "Suicide: Pediatric Mental Health Minute Series," *American Academy of Pediatrics*, .org, , Available athttps://www.aap.org/en/patient-care/mental-health-minute/suicide/.

Bosson, J. K., Buckner, C. E., and Vandello, J. A. (2022), *The psychology of sex and gender*, Los Angeles: SAGE.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008), *Time series analysis: forecasting and control*, Wiley series in probability and statistics, Hoboken, N.J.: John Wiley.

Canetto, S. S., and Sakinofsky, I. (1998), "The gender paradox in suicide," *Suicide & Life-Threatening Behavior*, 28, 1–23.

Card, D., and DellaVigna, S. (2013), "Nine Facts about Top Journals in Economics," *Journal of Economic Literature*, 51, 144–161. https://doi.org/10.1257/jel.51.1.144.

Case, A., and Deaton, A. (2015), "Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century," *Proceedings of the National Academy of Sciences*, 112, 15078–15083. https://doi.org/10.1073/pnas.1518393112.

Case, A., and Deaton, A. (2017), "Mortality and Morbidity in the 21st Century," *Brookings Papers on Economic Activity*, 2017, 397–476. https://doi.org/10.1353/eca.2017.0005.

Case, A., and Deaton, A. (2020), *Deaths of despair and the future of capitalism*, Princeton: Princeton University Press.

"CDC releases 2019 Youth Risk Behavior Survey Results | CDC" (2023), Available athttps://www.cdc.gov/healthyyouth/data/yrbs/feature/index.htm.

Florida Commission on Mental Health & Substance Abuse (2001), "Florida Commission on Mental Health & Substance Abuse Final Report," *FMHI Publications*.

Friedman, J., Hansen, H., and Gone, J. P. (2023), "Column: What Does the 'Deaths of Despair' Narrative Leave Out?," *www.uclahealth.org*, .org, , Available athttps://www.uclahealth.org/news/column-what-does-deaths-despair-narrative-leave-out.

Gold, M. S. (2020), "The Role of Alcohol, Drugs, and Deaths of Despair in the U.S.'s Falling Life Expectancy," *Missouri Medicine*, 117, 99–101.

Hassouna, F. M. A., and Al-Sahili, K. (2020), "Practical Minimum Sample Size for Road Crash Time-Series Prediction Models," *Advances in Civil Engineering*, (V. Vignali, ed.), 2020, 1–12. https://doi.org/10.1155/2020/6672612.

Henderson, D. R. (2020), "Deaths of Despair," *cato.org*, .org, , Available athttps://www.cato.org/regulation/fall-2020/deaths-despair#.

Ilic, M., and Ilic, I. (2022), "Worldwide suicide mortality trends (2000-2019): A joinpoint regression analysis," *World Journal of Psychiatry*, 12, 1044–1060. https://doi.org/10.5498/wjp.v12.i8.1044.

Introcaso, D. (2021), "Deaths of despair: the unrecognized tragedy of working class immiseration," *STAT*.

Johnson, N. (2017), "A Comparison of PCE and CPI: Methodological Differences in U.S. Inflation Calculation and Their Implications : U.S. Bureau of Labor Statistics," *Bureau of Labor Statistics*, Available athttps://www.bls.gov/osmr/research-papers/2017/st170010.htm.

Johnson, S. R. (2024), "Suicide Rates Have Risen Among People of Color," *U.S. News and World Report*, .com, , Available athttps://www.usnews.com/news/health-news/articles/2023-02-09/suicide-rates-have-risen-among-people-of-color.

Kingkade, T., and Chuck, E. (2021), "Suicidal thoughts are increasing in young kids, experts say. It began before the pandemic.," *NBC News*, .com, , Available athttps://www.nbcnews.com/news/us-news/suicidal-thoughts-are-increasing-young-kids-experts-say-it-began-n1263347.

McKenzie, D. (2023), "Interpreting treatment effects on an inverse hyperbolic sine outcome variable and alternatives," *World Bank Blogs*, .org, , Available athttps://blogs.worldbank.org/en/impactevaluations/interpreting-treatment-effects-inverse-hyperbolic-sine-outcome-variable-and.

Muldoon, A. (2018), “The log-0 problem: analysis strategies and options for choosing c in log(y + c),” *Very statisticious*, Available athttps://aosmith.rbind.io/2018/09/19/the-log-0-problem/.
“National income - Gross national income - OECD Data” (n.d.). *theOECD*, Available athttp://data.oecd.org/natincome/gross-national-income.htm.
Norton, E. (2022), *The Inverse Hyperbolic Sine Transformation and Retransformed Marginal Effects*, Cambridge, MA: National Bureau of Economic Research, p. w29998. https://doi.org/10.3386/w29998.
Ogozalek, S. (2023), “A lot of thought, little action: Proposals about mental health go unheeded,” *Health News Florida*, Available athttps://health.wusf.usf.edu/health-news-florida/2023-03-22/a-lot-of-thought-little-action-proposals-about-mental-health-go-unheeded.
Ogozalek, S. (n.d.). “Thousands of Floridians struggle to access mental health services in complex, disjoined system,” *Tampa Bay Times*, Available athttps://www.tampabay.com/news/health/2023/03/21/mental-health-florida-reforms-report-flagged-problems-parkland/.
Onyango, R. (2024), “Suicide Rates & Socioeconomic Factors (1990 - 22),” kaggle.com. https://doi.org/10.34740/KAGGLE/DS/4597596.
Schrijvers, D. L., Bollen, J., and Sabbe, B. G. C. (2012), “The gender paradox in suicidal behavior and its impact on the suicidal process,” *Journal of Affective Disorders*, 138, 19–26. https://doi.org/10.1016/j.jad.2011.03.050.
“Suicide Data and Statistics | Suicide Prevention | CDC” (2023), Available athttps://www.cdc.gov/suicide/suicide-data-statistics.html.
The Investopedia Team (2024), “Gross National Income (GNI) Definition, With Real-World Example,” *Investopedia*, Available athttps://www.investopedia.com/terms/g/gross-national-income-gni.asp.
Tucker, R. P. (2020), “Gender Paradox of Suicide: Relationship Between Gender & Suicide,” *CAMS-care*, Available athttps://cams-care.com/resources/educational-content/the-gender-paradox-of-suicide/.
Zheng, H., and Choi, Y. (2024), “Reevaluating the ‘deaths of despair’ narrative: Racial/ethnic heterogeneity in the trend of psychological distress-related death,” *Proceedings of the National Academy of Sciences*, 121, e2307656121. https://doi.org/10.1073/pnas.2307656121.

V: Appendix B – R Code

```
#######################################

#W. Elijah Clark STA 5167 Project Code#

#######################################

#Note: Some Code Discrepancies have occurred with summary statements between the presentation and the final paper.

#All Libraries Used

#########

library(ggplot2)

library(readr)

library(MASS)

library(dplyr)

library(car)

library(multcompView)

library(Rfit)

library(NSM3)

library(corrr)

library(ggcorrplot)

library(factoextra)

library(data.table)

library(GLDEX)

#########


#Custom Functions Used

#########

ihs <- function(x) {

  y <- log(x + sqrt(x ^ 2 + 1))
```

```
  return(y)

}

#########



#Data Import

#########

library(readr)

Death <- read_csv("All Academic Files/All Graduate Dox/Grad School Course Files/Statistics in Applications 2/Project/New_Project_Post_Withdraw/suicide_rates_1990-2022.csv")



#Removing Duplicates and NAs

library(dplyr)

Death %>% distinct()

Death <- na.omit(Death)



#Old Imputation Method with aggregate() (Did not impute all semi-duplicate rows.)

#unalived1 <- aggregate(Death$SuicideCount,by=list(RegionName=Death$RegionName, CountryName=Death$CountryName, Year=Death$Year, Sex=Death$Sex, AgeGroup=Death$AgeGroup, CauseSpecificDeathPercentage=Death$CauseSpecificDeathPercentage, DeathRatePer100K=Death$DeathRatePer100K, Population=Death$Population, GDP=Death$GDP, GDPPerCapita=Death$GDPPerCapita, GrossNationalIncome=Death$GrossNationalIncome, GNIPerCapita=Death$GNIPerCapita, InflationRate=Death$InflationRate, EmploymentPopulationRatio=Death$EmploymentPopulationRatio),FUN=mean)

#unalived2 <- aggregate(unalived1$CauseSpecificDeathPercentage,by=list(RegionName=unalived1$RegionName, CountryName=unalived1$CountryName, Year=unalived1$Year, Sex=unalived1$Sex, AgeGroup=unalived1$AgeGroup, SuicideCount=unalived1$x, DeathRatePer100K=unalived1$DeathRatePer100K, Population=unalived1$Population, GDP=unalived1$GDP, GDPPerCapita=unalived1$GDPPerCapita, GrossNationalIncome=unalived1$GrossNationalIncome, GNIPerCapita=unalived1$GNIPerCapita, InflationRate=unalived1$InflationRate, EmploymentPopulationRatio=unalived1$EmploymentPopulationRatio),FUN=mean)

#unalived3 <- aggregate(unalived2$DeathRatePer100K,by=list(RegionName=unalived2$RegionName, CountryName=unalived2$CountryName, Year=unalived2$Year, Sex=unalived2$Sex, AgeGroup=unalived2$AgeGroup, SuicideCount=unalived2$SuicideCount, CauseSpecificDeathPercentage=unalived2$x, Population=unalived2$Population, GDP=unalived2$GDP,
```

```
GDPPerCapita=unalived2$GDPPerCapita, GrossNationalIncome=unalived2$GrossNationalIncome,
GNIPerCapita=unalived2$GNIPerCapita, InflationRate=unalived2$InflationRate,
EmploymentPopulationRatio=unalived2$EmploymentPopulationRatio),FUN=mean)

#unalived4 <- na.omit(unalived3)

#unalived <- unalived4



#Newer Data Cleaning (Imputes all rows with dplyr)

#Note: One time I ran this removed Unknown genders.

#Re-running it again with post-presentation feedback somehow keeps that factor.

#I do not think I changed anything, so I do not know why that happened.

Death <- Death %>%

 group_by(

  RegionName,

  CountryName,

  Year,

  Sex,

  AgeGroup,

  Population,

  GDP,

  GDPPerCapita,

  GrossNationalIncome,

  GNIPerCapita,

  InflationRate,

  EmploymentPopulationRatio

 ) %>%

 summarise(

  SuicideCount = mean(SuicideCount),
```

```
    CauseSpecificDeathPercentage = mean(CauseSpecificDeathPercentage),
    DeathRatePer100K = mean(DeathRatePer100K),
    .groups = "drop"
  )

#How to test if the dplyr data cleaning worked:
Death %>%
  filter(
    CountryName == "United States of America",
    AgeGroup == "25-34 years",
    Year == 2020,
    Sex == "Male"
  ) %>%
  nrow()

#Natural Zero Death versus Removing Zero Deaths
#Remaking the data frame for natural zeros versus removing zeros

#Scenario 1: Zero deaths assumed to be naturally occurring values
unalived <- Death

#Variables Declared Version 1: Imputed
#########
RegionCode <- unalived$RegionCode
Factored_RegionCode <- factor(RegionCode)
RegionName <- unalived$RegionName
```

Factored_RegionName <- factor(RegionName)

Year <- unalived$Year

Factored_Year <- factor(Year)

Gender <- unalived$Sex

Factored_Gender <- unalived$Sex

AgeGroup <- unalived$AgeGroup

Factored_AgeGroup <- unalived$AgeGroup

Unalived_Count <- unalived$SuicideCount

Unalived_Percentage <- unalived$CauseSpecificDeathPercentage

Unalived_100K <- unalived$DeathRatePer100K

Population <- unalived$Population

GDP <- unalived$GDP

GDP_Capita <- unalived$GDPPerCapita

GNI <- unalived$GrossNationalIncome

GNI_Capita <- unalived$GNIPerCapita

Inflation <- unalived$InflationRate

Employment <- unalived$EmploymentPopulationRatio

#Transformed Variables

float_UnalivedCount <- Unalived_Count+.0001

float_UnalivedPercent <- Unalived_Percentage+.0001

float_Unalived100K <- Unalived_100LK+.0001

log_floatUnaliveCount <- log(float_UnalivedCount)

log_floatUnalivePercent <- log(float_UnalivedPercent)

log_floatUnalive100K <- log(float_Unalived100K)

ihs_UnalivedCount <- ihs(Unalived_Count)

```
ihs_UnalivedPercent <- ihs(Unalived_Percentage)

ihs_Unalived100K <- ihs(Unalived_100K)

log_population <- log(Population)

log_GDP <- log(GDP)

log_GDPCapita <- log(GDP_Capita)

log_GNI <- log(GNI)

log_GNICapita <- log(GNI)

#########


#Summary Statistics

#######

fivenum(Unalived_Count)

boxplot(fivenum(Unalived_Count))

fivenum(Unalived_Percentage)

boxplot(fivenum(Unalived_Percentage))

fivenum(Unalived_100K)

boxplot(fivenum(Unalived_100K))

fivenum(log_floatUnaliveCount)

boxplot(fivenum(log_floatUnaliveCount))

#######


#Some QQ Plots

########

qqPlot(Unalived_Count) + title("QQPlot of untransformed Suicide Deaths")

qqPlot(log_floatUnaliveCount) + title("QQPlot of log Suicide Deaths")

qqPlot(ihs_UnalivedCount) + title("QQPlot of IHS Suicide Deaths")
```

```
########


#PCA and Corr Matrix for Scenario 1
#######
#Different Data Frame and Correlation Matrix
testDF1 <- data.frame(Unalived_Count,
                      Unalived_Percentage,
                      Unalived_100K,
                      Population,
                      GDP,
                      GDP_Capita,
                      GNI,
                      GNI_Capita,
                      Inflation,
                      Employment)

data_normalized <- scale(testDF1)
corr_matrix <- cor(testDF1)
print(corr_matrix)
ggcorrplot(corr_matrix)

#PCA Stuff
data.pca <- princomp(corr_matrix)
summary(data.pca)
```

```
data.pca$loadings[, 1:2]

fviz_eig(data.pca, barfill = "thistle2", barcolor = "thistle", addlabels = TRUE)


#Note: cos2 entails qualities of representation

fviz_pca_var(data.pca,  col.var = "cos2",

        gradient.cols = c("midnightblue", "aquamarine", "coral"),

        repel = TRUE)



#Corr Matrix with Transformed Variables

testDF2 <- data.frame(ihs_UnalivedCount,

            ihs_UnalivedPercent,

            ihs_Unalived100K,

            Population,

            log_GDP,

            log_GDPCapita,

            log_GNI,

            log_GNICapita,

            Inflation,

            Employment)


data_normalized2 <- scale(testDF2)

corr_matrix2 <- cor(testDF2)

ggcorrplot(corr_matrix2)
```

```
#######



#Attempted Box Cox Transforms

#######

library(MASS)

boxcox(lm(float_UnalivedCount ~ Population)) + title("Lambda for Death Count & Population")

boxcox(lm(GDP ~ Population)) + title("Lambda for GDP & Population")

boxcox(lm(GNI ~ Population)) + title("Lambda for GNI & Population")


boxcox(lm(Population ~ float_UnalivedCount)) + title("Lambda for Death Count & Population")

boxcox(lm(Population ~ GDP)) + title("Lambda for GDP & Population")

boxcox(lm(Population ~ GNI)) + title("Lambda for GNI & Population")


b <- boxcox(lm(GNI ~ Population))

# Exact lambda

#Note: DO NOT CHANGE FORMULA BELOW. Replacing y and x with real object names BREAKS this.

exact_lambda <- b$x[which.max(b$y)]

print(exact_lambda)


b2 <- boxcox(lm(GDP ~ Population))

# Exact lambda

#Note: DO NOT CHANGE FORMULA BELOW. Replacing y and x with real object names BREAKS this.

exact_lambda <- b2$x[which.max(b2$y)]

print(exact_lambda)
```

```
b3 <- boxcox(lm(float_UnalivedCount ~ Population))

# Exact lambda

#Note: DO NOT CHANGE FORMULA BELOW. Replacing y and x with real object names BREAKS this.

exact_lambda <- b3$x[which.max(b3$y)]

print(exact_lambda)

#######


#Step-wise Regressions: Unalived Counts

#######

#Log Transforms



#GDP with Years

stepwise_1 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_population + log_GDP + Inflation + Employment)

stepwise_1 <- step(stepwise_1, direction = "both")

summary(stepwise_1)



#GDP without Years

stepwise_2 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_population + log_GDP + Inflation + Employment)

stepwise_2 <- step(stepwise_2, direction = "both")

summary(stepwise_2)



#GNI with Years

stepwise_3 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_population + log_GNI + Inflation + Employment)

stepwise_3 <- step(stepwise_3, direction = "both")
```

summary(stepwise_3)

#GNI without Years

stepwise_4 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GNI + Inflation + Employment)

stepwise_4 <- step(stepwise_4, direction = "both")

summary(stepwise_4)

#GDP/Capita with Years

stepwise_5 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GDPCapita + Inflation + Employment)

stepwise_5 <- step(stepwise_5, direction = "both")

summary(stepwise_5)

#GDP/Capita without Years

stepwise_6 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GDPCapita + Inflation + Employment)

stepwise_6 <- step(stepwise_6, direction = "both")

summary(stepwise_6)

#GNI/Capita with Years

stepwise_7 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GNICapita + Inflation + Employment)

stepwise_7 <- step(stepwise_7, direction = "both")

summary(stepwise_7)

#GNI/Capita without Years

stepwise_8 <- lm(log_floatUnaliveCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GNICapita + Inflation + Employment)

stepwise_8 <- step(stepwise_8, direction = "both")

summary(stepwise_8)

#IHS Transforms

#GDP with Years

stepwise_9 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_population + log_GDP + Inflation + Employment)

stepwise_9 <- step(stepwise_9, direction = "both")

summary(stepwise_9)

#GDP without Years

#This appears to be one of the two best models

stepwise_10 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GDP + Inflation + Employment)

stepwise_10 <- step(stepwise_10, direction = "both")

summary(stepwise_10)

#GNI with Years

stepwise_11 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_population + log_GNI + Inflation + Employment)

stepwise_11 <- step(stepwise_11, direction = "both")

summary(stepwise_11)

#GNI without Years

#This appears to be the other of the two best models

stepwise_12 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GNI + Inflation + Employment)

stepwise_12 <- step(stepwise_12, direction = "both")

summary(stepwise_12)

#GDP/Capita with Years

stepwise_13 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GDPCapita + Inflation + Employment)

stepwise_13 <- step(stepwise_13, direction = "both")

summary(stepwise_13)

#GDP/Capita without Years

stepwise_14 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GDPCapita + Inflation + Employment)

stepwise_14 <- step(stepwise_14, direction = "both")

summary(stepwise_14)

#GNI/Capita with Years

stepwise_15 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GNICapita + Inflation + Employment)

stepwise_15 <- step(stepwise_15, direction = "both")

summary(stepwise_15)

#GNI/Capita without Years

stepwise_16 <- lm(ihs_UnalivedCount ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GNICapita + Inflation + Employment)

stepwise_16 <- step(stepwise_16, direction = "both")

summary(stepwise_8)

```
#######

#Step-wise Regressions: Unalived Percentages

#######

#Log Transforms


#GDP with Years

stepwise_17 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_population + log_GDP + Inflation + Employment)

stepwise_17 <- step(stepwise_17, direction = "both")

summary(stepwise_17)


#GDP without Years

stepwise_18 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_population + log_GDP + Inflation + Employment)

stepwise_18 <- step(stepwise_18, direction = "both")

summary(stepwise_18)


#GNI with Years

stepwise_19 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_population + log_GNI + Inflation + Employment)

stepwise_19 <- step(stepwise_19, direction = "both")

summary(stepwise_19)


#GNI without Years

stepwise_20 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_population + log_GNI + Inflation + Employment)
```

```
stepwise_20 <- step(stepwise_20, direction = "both")

summary(stepwise_20)




#GDP/Capita with Years

stepwise_21 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_GDPCapita + Inflation + Employment)

stepwise_21 <- step(stepwise_21, direction = "both")

summary(stepwise_21)



#GDP/Capita without Years

stepwise_22 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_GDPCapita + Inflation + Employment)

stepwise_22 <- step(stepwise_22, direction = "both")

summary(stepwise_22)



#GNI/Capita with Years

stepwise_23 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_GNICapita + Inflation + Employment)

stepwise_23 <- step(stepwise_23, direction = "both")

summary(stepwise_23)



#GNI/Capita without Years

stepwise_24 <- lm(log_floatUnalivePercent ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_GNICapita + Inflation + Employment)

stepwise_24 <- step(stepwise_24, direction = "both")

summary(stepwise_24)
```

#IHS Transforms

#GDP with Years

stepwise_25 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_population + log_GDP + Inflation + Employment)

stepwise_25 <- step(stepwise_25, direction = "both")

summary(stepwise_25)

#GDP without Years

stepwise_26 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GDP + Inflation + Employment)

stepwise_26 <- step(stepwise_26, direction = "both")

summary(stepwise_26)

#GNI with Years

stepwise_27 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_population + log_GNI + Inflation + Employment)

stepwise_27 <- step(stepwise_27, direction = "both")

summary(stepwise_27)

#GNI without Years

stepwise_28 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GNI + Inflation + Employment)

stepwise_28 <- step(stepwise_28, direction = "both")

summary(stepwise_28)

#GDP/Capita with Years

stepwise_29 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GDPCapita + Inflation + Employment)

stepwise_29 <- step(stepwise_29, direction = "both")

summary(stepwise_29)

#GDP/Capita without Years

stepwise_30 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GDPCapita + Inflation + Employment)

stepwise_30 <- step(stepwise_30, direction = "both")

summary(stepwise_30)

#GNI/Capita with Years

stepwise_31 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GNICapita + Inflation + Employment)

stepwise_31 <- step(stepwise_31, direction = "both")

summary(stepwise_31)

#GNI/Capita without Years

stepwise_32 <- lm(ihs_UnalivedPercent ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GNICapita + Inflation + Employment)

stepwise_32 <- step(stepwise_32, direction = "both")

summary(stepwise_32)

#######

#Step-wise Regressions: Unalived 100LK

#######

#Log Transforms

#GDP with Years

stepwise_33 <- lm(log_floatUnalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_population + log_GDP + Inflation + Employment)

stepwise_33 <- step(stepwise_33, direction = "both")

summary(stepwise_33)

#GDP without Years

stepwise_34 <- lm(log_floatUnalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GDP + Inflation + Employment)

stepwise_34 <- step(stepwise_34, direction = "both")

summary(stepwise_34)

#GNI with Years

stepwise_35 <- lm(log_floatUnalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_population + log_GNI + Inflation + Employment)

stepwise_35 <- step(stepwise_35, direction = "both")

summary(stepwise_35)

#GNI without Years

stepwise_36 <- lm(log_floatUnalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_population + log_GNI + Inflation + Employment)

stepwise_36 <- step(stepwise_36, direction = "both")

summary(stepwise_36)

#GDP/Capita with Years

stepwise_37 <- lm(log_floatUnalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GDPCapita + Inflation + Employment)

stepwise_37 <- step(stepwise_37, direction = "both")

```
summary(stepwise_37)



#GDP/Capita without Years

stepwise_38 <- lm(log_floatUnalive100K ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_GDPCapita + Inflation + Employment)

stepwise_38 <- step(stepwise_38, direction = "both")

summary(stepwise_38)



#GNI/Capita with Years

stepwise_39 <- lm(log_floatUnalived100K ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + Factored_Year + log_GNICapita + Inflation + Employment)

stepwise_39 <- step(stepwise_39, direction = "both")

summary(stepwise_39)



#GNI/Capita without Years

stepwise_40 <- lm(log_floatUnalived100K ~ Factored_AgeGroup + Factored_Gender +
Factored_RegionName + log_GNICapita + Inflation + Employment)

stepwise_40 <- step(stepwise_40, direction = "both")

summary(stepwise_40)



#IHS Transforms



#GDP with Years

stepwise_41 <- lm(ihs_Unalived100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName +
Factored_Year + log_population + log_GDP + Inflation + Employment)

stepwise_41 <- step(stepwise_41, direction = "both")

summary(stepwise_41)
```

```
#GDP without Years

stepwise_42 <- lm(ihs_Unalived100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName +
log_population + log_GDP + Inflation + Employment)

stepwise_42 <- step(stepwise_42, direction = "both")

summary(stepwise_42)



#GNI with Years

stepwise_43 <- lm(ihs_Unalived100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName +
Factored_Year + log_population + log_GNI + Inflation + Employment)

stepwise_43 <- step(stepwise_43, direction = "both")

summary(stepwise_43)



#GNI without Years

stepwise_44 <- lm(ihs_Unalived100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName +
log_population + log_GNI + Inflation + Employment)

stepwise_44 <- step(stepwise_44, direction = "both")

summary(stepwise_44)




#GDP/Capita with Years

stepwise_45 <- lm(ihs_Unalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName +
Factored_Year + log_GDPCapita + Inflation + Employment)

stepwise_45 <- step(stepwise_45, direction = "both")

summary(stepwise_45)



#GDP/Capita without Years

stepwise_46 <- lm(ihs_Unalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName +
log_GDPCapita + Inflation + Employment)

stepwise_46 <- step(stepwise_46, direction = "both")
```

```
summary(stepwise_46)

#GNI/Capita with Years

stepwise_47 <- lm(ihs_Unalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + Factored_Year + log_GNICapita + Inflation + Employment)

stepwise_47 <- step(stepwise_47, direction = "both")

summary(stepwise_47)

#GNI/Capita without Years

stepwise_48 <- lm(ihs_Unalive100K ~ Factored_AgeGroup + Factored_Gender + Factored_RegionName + log_GNICapita + Inflation + Employment)

stepwise_48 <- step(stepwise_48, direction = "both")

summary(stepwise_48)

#######

#All ANOVAS

#Rudimentary One-Way ANOVA plots for visual comparisons: Unalived_Count

#######

#Gender

ggplot(unalived, aes(Gender, Unalived_Count)) + geom_boxplot(fill="thistle", color="thistle4") +theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())

ggplot(unalived, aes(Gender, log_floatUnaliveCount)) + geom_boxplot(fill="thistle", color="thistle4") +theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())

ggplot(unalived, aes(Gender, ihs_UnalivedCount)) + geom_boxplot(fill="thistle", color="thistle4") +theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())

#AgeGroup
```

```
ggplot(unalived, aes(AgeGroup, Unalived_Count)) + geom_boxplot(fill="thistle", color="thistle4")
+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())

ggplot(unalived, aes(AgeGroup, log_floatUnaliveCount)) + geom_boxplot(fill="thistle", color="thistle4")
+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())

ggplot(unalived, aes(AgeGroup, ihs_UnalivedCount)) + geom_boxplot(fill="thistle", color="thistle4")
+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())



#RegionName

ggplot(unalived, aes(RegionName, Unalived_Count)) + geom_boxplot(fill="thistle", color="thistle4")

ggplot(unalived, aes(RegionName, log_floatUnaliveCount)) + geom_boxplot(fill="thistle", color="thistle4")

ggplot(unalived, aes(RegionName, ihs_UnalivedCount)) + geom_boxplot(fill="thistle", color="thistle4")


#######


#Actual ANOVA Models


#One Way ANOVAs

######

#Untransformed

#Age

age_deathcount <- aov(Unalived_Count ~ Factored_AgeGroup)

summary(age_deathcount)

TukeyHSD(age_deathcount)


#Gender

gender_deathcount <- aov(Unalived_Count ~ Gender)

summary(gender_deathcount)
```

```
TukeyHSD(gender_deathcount)


#Region

region_deathcount <- aov(Unalived_Count ~ RegionName)

summary(region_deathcount)

TukeyHSD(region_deathcount)


#Log Transforms

age_deathcount2 <- aov(log_floatUnaliveCount ~ Factored_AgeGroup)

summary(age_deathcount2)

TukeyHSD(age_deathcount2)


#Gender

gender_deathcount2 <- aov(log_floatUnaliveCount ~ Gender)

summary(gender_deathcount2)

TukeyHSD(gender_deathcount2)


#Region

region_deathcount2 <- aov(log_floatUnaliveCount ~ RegionName)

summary(region_deathcount2)

TukeyHSD(region_deathcount2)


#IHS Transforms

age_deathcount3 <- aov(ihs_UnalivedCount ~ Factored_AgeGroup)

summary(age_deathcount3)

TukeyHSD(age_deathcount3)
```

```
#Gender
gender_deathcount3 <- aov(ihs_UnalivedCount ~ Gender)
summary(gender_deathcount3)
TukeyHSD(gender_deathcount3)

#Region
region_deathcount3 <- aov(ihs_UnalivedCount ~ RegionName)
summary(region_deathcount3)
TukeyHSD(region_deathcount3)
#####

#Two-Way Models
#####
#Age and Gender
agegender_deathcount <- aov(Unalived_Count ~ Factored_AgeGroup * Gender)
summary(agegender_deathcount)
TukeyHSD(agegender_deathcount)

#Gender and Region
genderregion_deathcount <- aov(Unalived_Count ~ Gender *Factored_RegionName)
summary(genderregion_deathcount)
TukeyHSD(genderregion_deathcount)

#Region and Age
regionage_deathcount <- aov(Unalived_Count ~ Factored_RegionName * Factored_AgeGroup)
```

```
summary(regionage_deathcount)

TukeyHSD(regionage_deathcount)


#Log Transforms

#Age and Gender

logagegender_deathcount <- aov(log_floatUnaliveCount ~ Factored_AgeGroup * Gender)

summary(logagegender_deathcount)

TukeyHSD(logagegender_deathcount)


#Gender and Region

loggenderregion_deathcount <- aov(log_floatUnaliveCount ~ Gender *Factored_RegionName)

summary(loggenderregion_deathcount)

TukeyHSD(loggenderregion_deathcount)


#Region and Age

logregionage_deathcount <- aov(log_floatUnaliveCount ~ Factored_RegionName * Factored_AgeGroup)

summary(logregionage_deathcount)

TukeyHSD(logregionage_deathcount)



#IHS Transforms


#Age and Gender

ihsagegender_deathcount <- aov(ihs_UnalivedCount ~ Factored_AgeGroup * Gender)

summary(ihsagegender_deathcount)

TukeyHSD(ihsagegender_deathcount)
```

#Gender and Region

*ihsgenderregion_deathcount <- aov(ihs_UnalivedCount ~ Gender *Factored_RegionName)*

summary(ihsgenderregion_deathcount)

TukeyHSD(ihsgenderregion_deathcount)

#Region and Age

*ihsregionage_deathcount <- aov(ihs_UnalivedCount ~ Factored_RegionName * Factored_AgeGroup)*

summary(ihsregionage_deathcount)

TukeyHSD(ihsregionage_deathcount)

#####

#Three-Way Model:

#####

#Untransformed

*RegionAgeGend_deathcount <- aov(Unalived_Count ~ Factored_RegionName * Factored_Gender * Factored_AgeGroup)*

summary(RegionAgeGend_deathcount)

TukeyHSD(RegionAgeGend_deathcount)

#Log

*Log_RegionAgeGend_deathcount <- aov(log_floatUnaliveCount ~ Factored_RegionName * Factored_Gender * Factored_AgeGroup)*

summary(Log_RegionAgeGend_deathcount)

TukeyHSD(Log_RegionAgeGend_deathcount)

```
#IHS

IHS_RegionAgeGend_deathcount <- aov(ihs_UnalivedCount ~ Factored_RegionName * Factored_Gender *
Factored_AgeGroup)

summary(IHS_RegionAgeGend_deathcount)

TukeyHSD(IHS_RegionAgeGend_deathcount)


######




#Time Series Attempts

######


#United States

US_Deaths <- unalived[unalived$CountryName %in% c("United States of America"),]

US_Deaths_Male <- US_Deaths[US_Deaths$Sex %in% c("Male"),]

US_Deaths_Male_2534 <- US_Deaths_Male[US_Deaths_Male$AgeGroup %in% c("25-34 years"),]


USM2534Deaths <- US_Deaths_Male_2534$SuicideCount

IHS_USM2534Deaths <- ihs(USM2534Deaths)


UStime <- ts(IHS_USM2534Deaths)

plot.ts(UStime) + title("Time Series of US Male 25-34 IHS SuicideCounts")


#Russia

Russian_Deaths <- unalived[unalived$CountryName %in% c("Russian Federation"),]

Russian_Deaths_Male <- Russian_Deaths[Russian_Deaths$Sex %in% c("Male"),]
```

```
Russian_Deaths_Male_2534 <- Russian_Deaths_Male[Russian_Deaths_Male$AgeGroup %in% c("25-34 years"),]

RussianM2534Deaths <- Russian_Deaths_Male_2534$SuicideCount

IHS_RussianM2534Deaths <- ihs(RussianM2534Deaths)

Russiantime <- ts(IHS_RussianM2534Deaths)

plot.ts(Russiantime) + title("Time Series of Russian Male 25-34 IHS SuicideCounts")

#UK+Ireland

UK_Deaths <- unalived[unalived$CountryName %in% c("United Kingdom of Great Britain and Northern Ireland"),]

UK_Deaths_Female <- UK_Deaths[UK_Deaths$Sex %in% c("Female"),]

UK_Deaths_Female_3554 <- UK_Deaths_Female[UK_Deaths_Female$AgeGroup %in% c("35-54 years"),]

UKF3554Deaths <- UK_Deaths_Female_3554$SuicideCount

IHS_UKF3554Deaths <- ihs(UKF3554Deaths)

UKtime <- ts(IHS_UKF3554Deaths)

plot.ts(UKtime) + title("Time Series of UK Female 35-54 IHS SuicideCounts")

#Mexico
```

```
Mexico_Deaths <- unalived[unalived$CountryName %in% c("Mexico"),]

Mexico_Deaths_Female <- Mexico_Deaths[Mexico_Deaths$Sex %in% c("Female"),]

Mexico_Deaths_Female_1524 <- Mexico_Deaths_Female[Mexico_Deaths_Female$AgeGroup %in% c("15-
24 years"),]



MexicoF1524Deaths <- Mexico_Deaths_Female_1524$SuicideCount

IHS_MexicoF1524Deaths <- ihs(MexicoF1524Deaths)


Mexicotime <- ts(IHS_MexicoF1524Deaths)

plot.ts(Mexicotime) + title("Time Series of Mexico Female 15-24 IHS SuicideCounts")




#Scenario 2: Zero deaths assumed to be underreports or nonreports

library(dplyr)

unalived2 <- filter(unalived, SuicideCount != 0)


#Variables Declared from unalived2

#########

RegionCode2 <- unalived2$RegionCode

Factored_RegionCode2 <- factor(RegionCode2)

RegionName2 <- unalived2$RegionName

Factored_RegionName2 <- factor(RegionName2)

CountryName2 <- unalived2$CountryName

Factored_CountryName2 <- factor(CountryName2)
```

```
Year2 <- unalived2$Year

Factored_Year2 <- factor(Year2)

Gender2 <- unalived2$Sex

Factored_Gender2 <- unalived2$Sex

AgeGroup2 <- unalived2$AgeGroup

Factored_AgeGroup2 <- unalived2$AgeGroup

Unalived_Count2 <- unalived2$SuicideCount

Unalived_Percentage2 <- unalived2$CauseSpecificDeathPercentage

Unalived_100K2 <- unalived2$DeathRatePer100K

Population2 <- unalived2$Population

GDP2 <- unalived2$GDP

GDP_Capita2 <- unalived2$GDPPerCapita

GNI2 <- unalived2$GrossNationalIncome

GNI_Capita2 <- unalived2$GNIPerCapita

Inflation2 <- unalived2$InflationRate

Employment2 <- unalived2$EmploymentPopulationRatio


#Transformed Variables (Note: IHS not necessary in this version.)

log_UnaliveCount2 <- log(Unalived_Count2)

log_UnalivePercent2 <- log(Unalived_Percentage2)

log_Unalive100K2 <- log(Unalived_100K2)

log_population2 <- log(Population2)

log_GDP2 <- log(GDP2)

log_GDPCapita2 <- log(GDP_Capita2)

log_GNI2 <- log(GNI2)

log_GNICapita2 <- log(GNI2)
```

```
#Summary Statistics
#######
fivenum(Unalived_Count2)
boxplot(fivenum(Unalived_Count2))
fivenum(Unalived_Percentage2)
boxplot(fivenum(Unalived_Percentage2))
fivenum(Unalived_100K2)
boxplot(fivenum(Unalived_100K2))
fivenum(log_UnaliveCount)
boxplot(fivenum(log_UnaliveCount))
#######


#PCA and Corr Matrix for Scenario 2
#######
#Different Data Frame and Correlation Matrix
testDF3 <- data.frame(Unalived_Count2,
           Unalived_Percentage2,
           Unalived_100K2,
           Population2,
           GDP2,
           GDP_Capita2,
           GNI2,
           GNI_Capita2,
```

```
                Inflation2,
                Employment2)


data_normalized <- scale(testDF3)
corr_matri3 <- cor(testDF3)
print(corr_matri3)
ggcorrplot(corr_matri3)


#PCA Stuff
data.pc3 <- princomp(corr_matri3)
summary(data.pca3)


data.pca3$loadings[, 1:2]
fviz_eig(data.pca3, barfill = "thistle2", barcolor = "thistle", addlabels = TRUE)


#Note: cos2 entails qualities of representation
fviz_pca_var(data.pca3,  col.var = "cos2",
         gradient.cols = c("midnightblue", "aquamarine", "coral"),
         repel = TRUE)



#Corr Matrix with Transformed Variables
testDF4 <- data.frame(log_UnaliveCount2,
                log_UnalivePercent2,
                log_Unalive100K2,
                log_population2,
```

```
        log_GDP2,
        log_GDPCapita2,
        log_GNI2,
        log_GNICapita2,
        Inflation2,
        Employment2)


data_normalized4 <- scale(testDF4)

corr_matrix4 <- cor(testDF4)

ggcorrplot(corr_matrix4)

#######


#Regression Model for Scenario2

#Based on better models from S1

stepwise_s2 <- lm(log_UnaliveCount2 ~ Factored_AgeGroup2 + Factored_Gender2 +
Factored_RegionName2 + log_GDPCapita2 + Inflation2 + Employment2)

stepwise_s2 <- step(stepwise_s2, direction = "both")

summary(stepwise_s2)
```