Final Project: Analyzing Suicide Mortality Trends by Demographic Factors

by Joseph Kim and Jun Tianzhong

2024-04-23

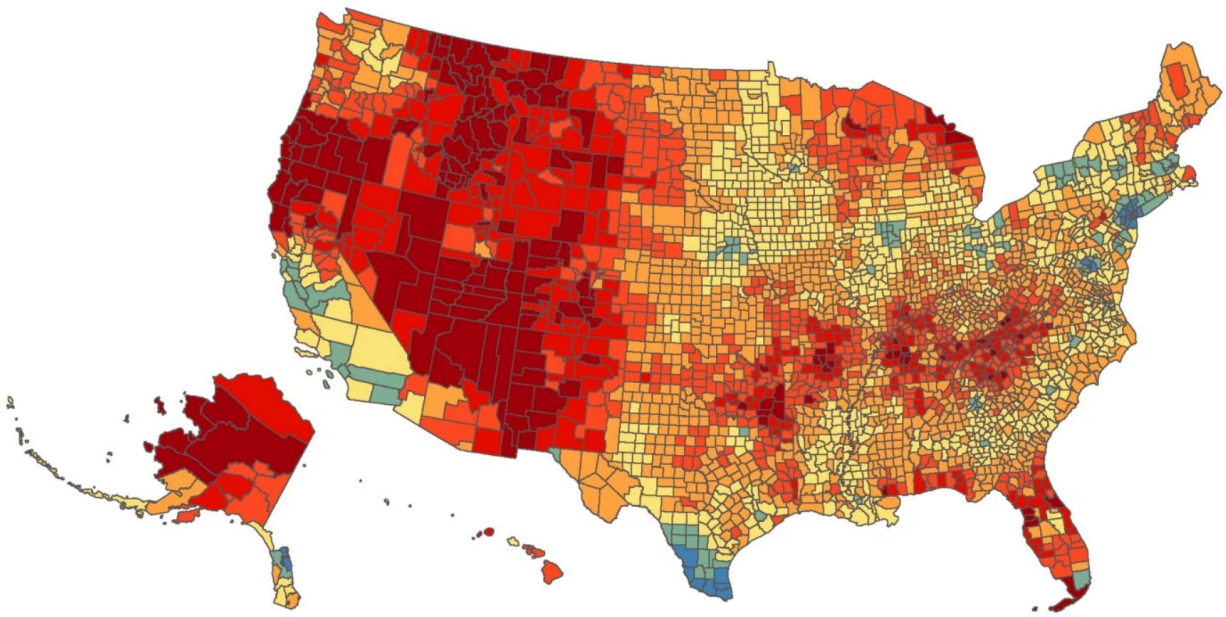

Figure 1: Suicide Rate by County in the United States in 2016.

Abstract

Introduction

Understanding the dynamics of suicide rates across demographic categories is paramount in addressing the complexities of mental health and societal well-being. Over the decades, the United States has witnessed fluctuations in suicide rates (increased trend), influenced by a multitude of factors including societal changes, economic conditions, and mental health awareness. The exploration of these trends sheds light on the historical patterns of suicide but also guides targeted interventions and awareness to mitigate risk factors associated with suicide.

Recent years we have seen a growing recognition of the urgency to address mental health issues, including suicide prevention, at both the nation and global level. With increasing awareness, there has been an increase in research efforts aimed to dissecting the interrelationship between demographic variables and suicide rates.This includes the examination of disparities across sex, race, Hispanic origin, and age groups, acknowledging the nuanced experiences and demographic categories that are the most vulnerable.

With a data set provided by the U.S. Department of Health & Human Services titled "Death rates for suicide, by sex, race, Hispanic origin, and age: United States", this study seeks to contribute to this ongoing research by comprehensively analyzing death rates for suicide in the United States spanning nearly seven decades, from 1950 to 2018. By analyzing into historical data, we aim to discern long-term trends and identify significant shifts in suicide rates among different demographic groups. Moreover, we want to uncover potential underlying factors contributing to these trends, informing targeted interventions and policy initiatives tailored to address the unique needs of diverse populations like the United States. In essence, we aspire to contribute towards fostering a society that has increased mental health awareness through this research.

Data Analysis & Results

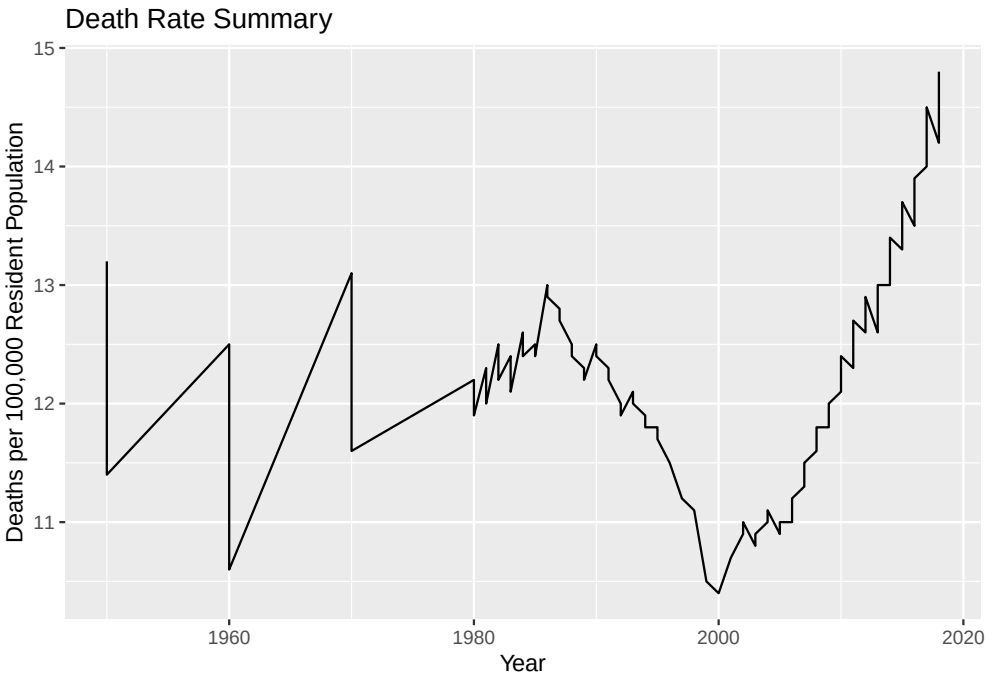

Figure 2: Suicide Death Rate 1960-2018

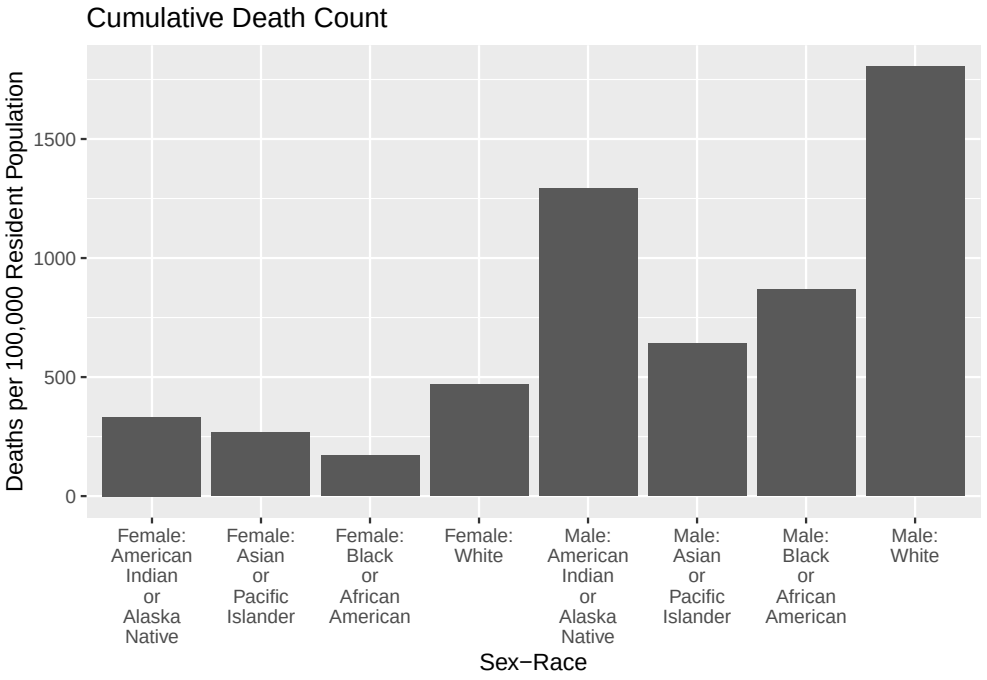

Figure 3: Cumulative Death Count from Suicide in the U.S.

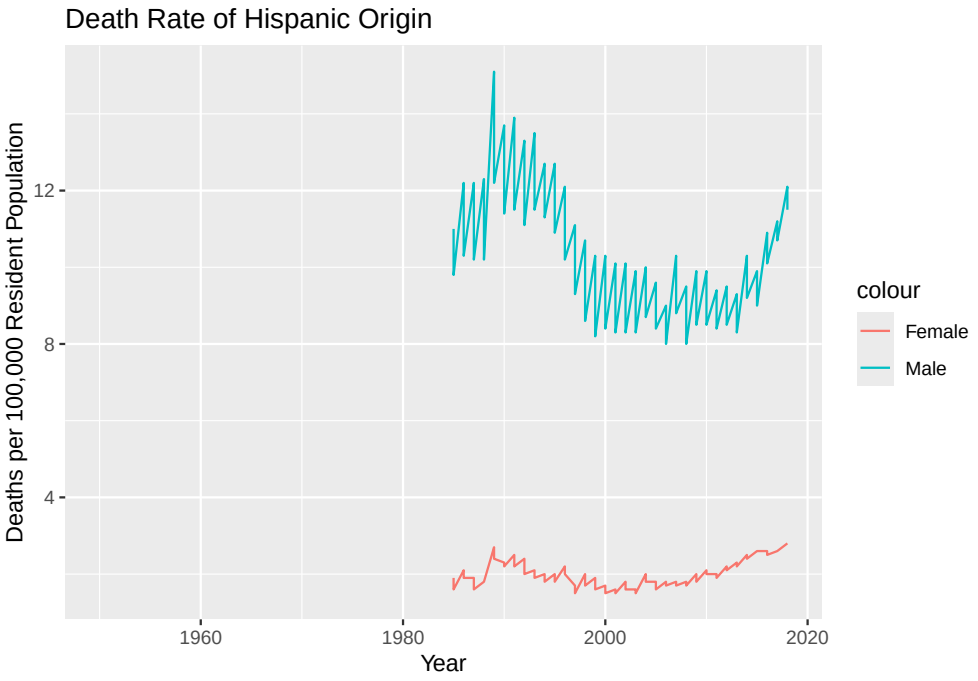

Figure 4: Suicide Rate of Hispanic Origin by Sex

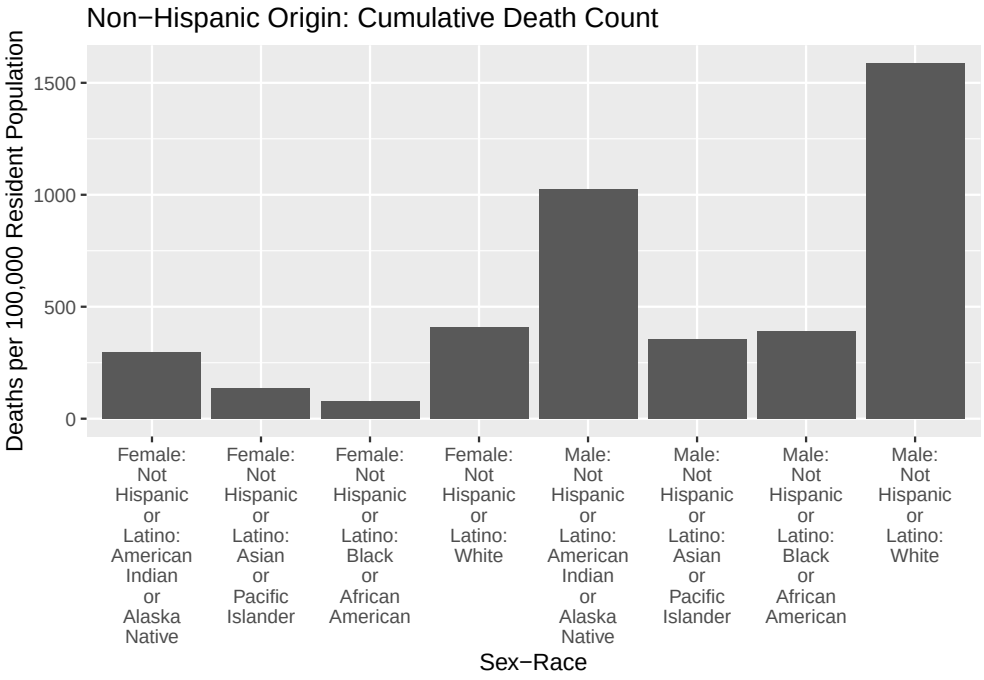

Figure 5: Cumulative Death Count of Non-Hispanic Origin

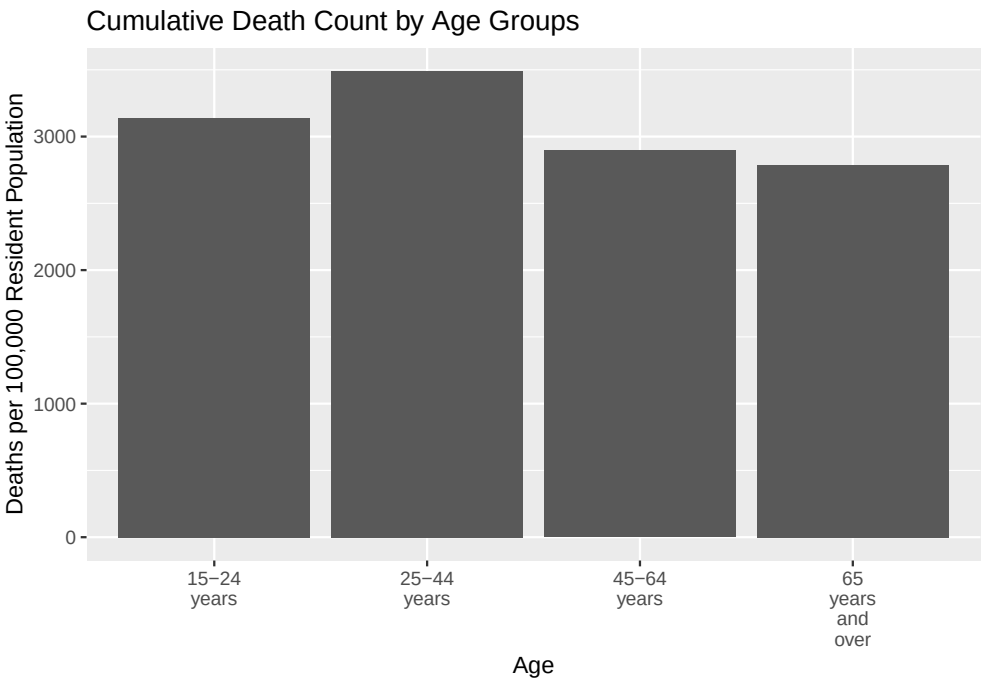

Figure 6: Cumulative Death Count of Different Age Groups

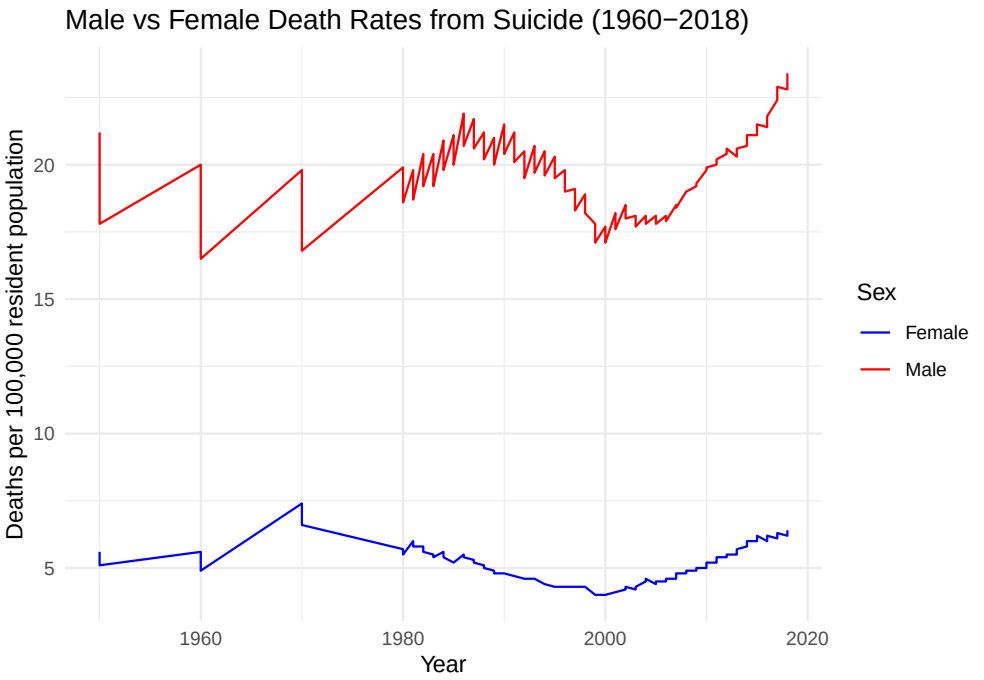

Figure 7: Suicide Rates of Each Sex

ANOVA Test Across Demographics

Testing for the difference in mean population of death rates observed across different age groups, races, Hispanic origins, and sexes:

- **Null Hypothesis:** H_0 : There are no significant differences in death rates among different age groups, races, Hispanic origins, and sexes.
- **Alternative Hypothesis:** H_a : At least one of the factors (age, race, hispanic origin, or sex) has a significant effect on death rates.
- $\alpha = 0.05$

Below is a subset of data achieved for the next set of tests and forms of analysis:

```
df <- data.frame(age = character(), race = character(), hispanic_origin = character(), sex = character()

for (i in 1:nrow(ageDeathRate)) {
  row <- ageDeathRate[i, ]
  df[i, "age"] <- row$AGE
  string <- row$STUB_LABEL
  split <- strsplit(string, ":")
  df[i, "sex"] <- trimws(split[[1]][1])
  df[i, "race"] <- trimws(split[[1]][3])
  df[i, "hispanic_origin"] <- trimws(split[[1]][2])
  df[i, "death"] <- row$ESTIMATE
}
```

```
anova_result <- aov(death ~ age + race + hispanic_origin + sex, data = df)
anova_table <- anova(anova_result)
print(anova_table)
```

```
## Analysis of Variance Table
##
## Response: death
##             Df Sum Sq Mean Sq   F value    Pr(>F)
## age          3    664     221    6.9676 0.0001224 ***
## race         4  37582    9396  295.6210 < 2.2e-16 ***
## sex          1  51688   51688 1626.3166 < 2.2e-16 ***
## Residuals 943  29971      32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table above provides insights into the significant effects of various factors on the outcome variable, death. This analysis was conducted to understand the influence of age, race, and sex on mortality rates.

Age: The ANOVA results indicate a significant effect of age on suicide rates ($F(3, 943) = 6.9676$, $p < 0.001$). Post-hoc tests revealed that mortality rates vary significantly across different age groups, with older age groups showing higher mortality rates compared to younger age groups.

Race: Race also emerges as a significant predictor of suicide rates ($F(4, 943) = 295.6210$, $p < 0.001$). Subsequent analyses suggest that mortality rates differ significantly across racial groups, with certain races experiencing higher suicide rates compared to others.

Sex: The ANOVA results demonstrate a significant effect of sex on suicide rates ($F(1, 943) = 1626.3166$, $p < 0.001$). Further examination reveals that mortality rates differ significantly between males and females, with one gender exhibiting higher mortality rates than the other.

All the p-values are less than 0.001 therefore these findings underscore the importance of considering demographic factors such as age, race, and sex when examining mortality rates.

Correlation Matrix

Below is the formation of a correlation matrix between the following demographics: age, race, Hispanic origin, and sex.

```
df_numeric <- df %>%
  mutate_if(is.character, as.factor) %>%
  mutate_all(as.numeric)
correlation_matrix <- cor(select(df_numeric, -death))
correlation_matrix
```

```
##                          age          race hispanic_origin           sex
## age              1.000000e+00 -5.816310e-18    2.752008e-18  0.000000e+00
## race            -5.816310e-18  1.000000e+00    8.044853e-01 -4.677704e-16
## hispanic_origin  2.752008e-18  8.044853e-01    1.000000e+00 -9.830500e-16
## sex              0.000000e+00 -4.677704e-16   -9.830500e-16  1.000000e+00
```

The correlation matrix provides insights into the relationships between various demographic variables in our dataset. Each cell in the matrix represents the correlation coefficient between two variables.

For this correlation matrix, we see that there is no off-diagonal correlation except between hispanic_origin and race of a factor of 0.8044853, which is expected due to the overlap.

In our analysis, it's essential to acknowledge that the data might not fully capture the intricate relationships between demographics due to potential missing or incomplete data. A much larger amount of data is required for a more accurate correlation matrix.

Distribution by Race and Age Groups

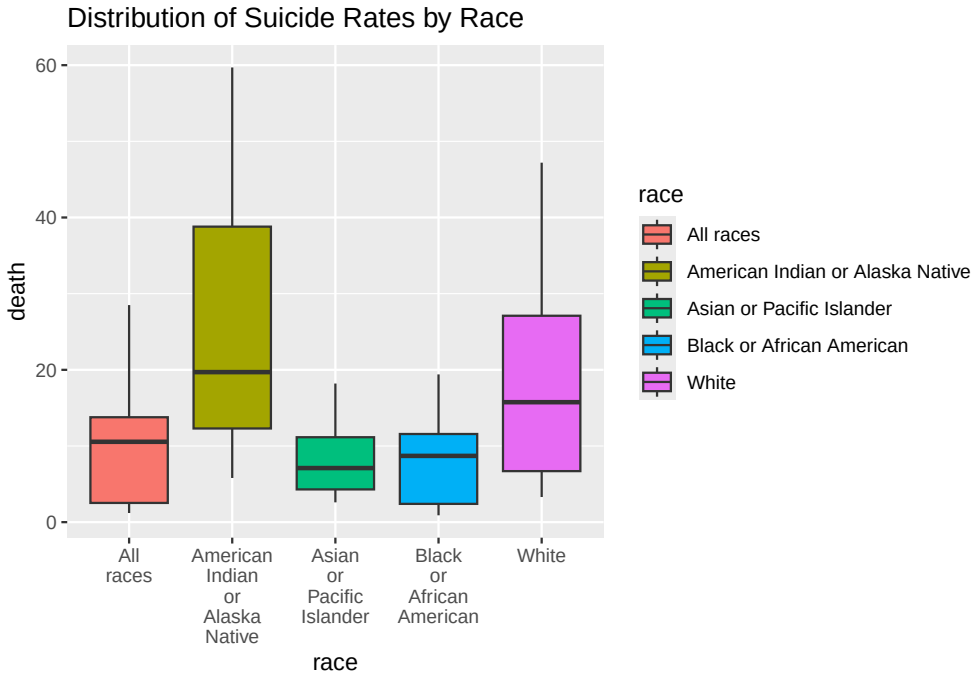

Figure 8: Distribution of Suicide Rates by Race

Distribution of Suicide Rates by Race: reveals varying rates across different racial demographics. Notably, the American Indian or Alaska Native group exhibits the highest suicide rate overall, whereas the Asian or Pacific Islander group demonstrates the lowest.

Distribution of Suicide Rates by Age Group Within Each Race: across racial categories, the age group of 65 and older displays the highest suicide rate, particularly evident within the Asian or Pacific Islander and White demographics. Conversely, the Black or African American community shows its highest suicide rate distribution within the 25-44 age group. Furthermore, the American Indian or Alaska Native group exhibits its peak suicide rate in the 15-24 age bracket.

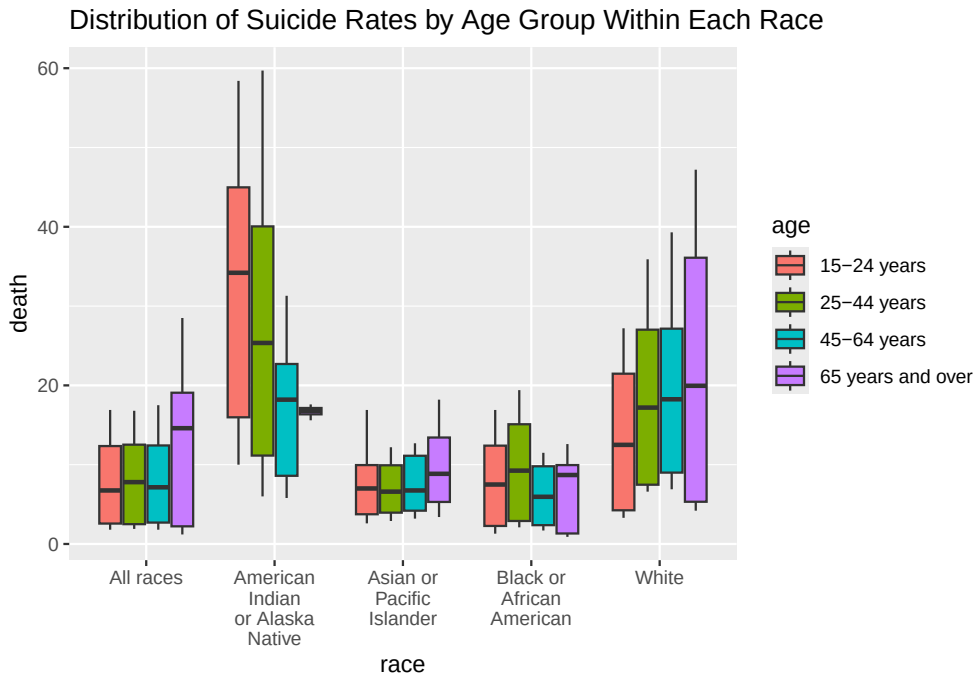

Figure 9: Distribution of Suicide Rates by Age Group Within Each Race

Longitudinal Model Across Demographics

The following formulates a linear mixed-effects model to explore how demographic factors relate to suicide rates while including variations:

```
df <- na.omit(df)
longitudinal_model <- lmer(death ~ age + sex + race + (1 | age) + (1 | sex) + (1 | race), data = df)
summary(longitudinal_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: death ~ age + sex + race + (1 | age) + (1 | sex) + (1 | race)
##    Data: df
##
## REML criterion at convergence: 5984.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3680 -0.6794 -0.0395  0.5257  4.8747
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  race     (Intercept)  33.4188  5.7809
##  age      (Intercept)   0.7391  0.8597
##  sex      (Intercept) 378.5197 19.4556
##  Residual              31.7824  5.6376
## Number of obs: 952, groups:  race, 5; age, 4; sex, 2
##
## Fixed effects:
```

```
##                                     Estimate Std. Error t value
## (Intercept)                           0.3375    20.3207   0.017
## age25-44 years                        1.1382     1.3149   0.866
## age45-64 years                       -0.4932     1.3171  -0.374
## age65 years and over                  2.0277     1.3304   1.524
## sexMale                              14.7910    27.5168   0.538
## raceAmerican Indian or Alaska Native 15.9519     8.2012   1.945
## raceAsian or Pacific Islander        -0.4795     8.1952  -0.059
## raceBlack or African American        -1.2435     8.1956  -0.152
## raceWhite                             9.5797     8.1901   1.170
##
## Correlation of Fixed Effects:
##             (Intr) a25-4y a45-6y a65yao sexMal rAIoAN rcAoPI rcBoAA
## age25-44yrs -0.032
## age45-64yrs -0.032  0.500
## ag65yrsando -0.032  0.495  0.495
## sexMale     -0.677  0.000  0.000  0.000
## rcAmrcnIoAN -0.201  0.000  0.001  0.003  0.000
## rcAsnorPcfI -0.201  0.000  0.000 -0.001  0.000  0.499
## rcBlckorAfA -0.201  0.000  0.000  0.000  0.000  0.499  0.499
## raceWhite   -0.202  0.000  0.000 -0.001  0.000  0.499  0.500  0.500
## optimizer (nloptwrap) convergence code: 0 (OK)
## unable to evaluate scaled gradient
##  Hessian is numerically singular: parameters are not uniquely determined
```

Considering the summary above, the model aims to account for variations within groups, capturing random variations across different ages, sexes, and races. Notably, the variance components for each grouping variable indicate considerable variability within race, age, and sex categories, suggesting heterogeneity in death rates. Regarding fixed effects, the coefficients provide estimates of the impact of each demographic factor on death rates. For instance, the positive coefficient for individuals aged 65 years and over suggests a higher death rate in this age group compared to the reference category. Additionally, the correlation of fixed effects table highlights potential collinearity issues between predictor variables. However, it's essential to note the model's convergence issues, indicated by the failure to converge due to a degenerate Hessian matrix with negative eigenvalues.

Along with our longitudinal model, we augmented a new data frame using US census data collected in 2022 to estimate the population count of different demographic groups by gender. This data frame is inputted into the longitudinal model to make predictions on the death counts and used to create a "predicted vs. observed plot" as shown in 10. From the plot, it can be clearly seen that our model has some errors in fitting the data, since the the line of best fit is close to being horizontal, indicating that on average, the model's predictions do not accurately capture the variability in the observed data across different time points or subjects. Additionally, it can be seen that there is a vertical scatter across death counts, which indicates that there are variability or inconsistency in how well they truly matched the observed values for individual data points. In retrospective, we found the reason behind such prediction errors due to both our newly augmented data frame and tested data frame lacking a time variable (augmented data frame is data from 2022 and tested data frame is a summary of data from 1980-2018, which means both data frames lack time-varying variables to represent how the predictor variables changed over time). Additionally, this unexpected result can also be explained by our finding of the longitudinal model being heteroscedastic based on the results obtained from 11, as the fitted vs. residual graph showing three separate clusters, wth one dense cluster and two sparser clusters at higher values, suggesting that the variance of the residuals varied across different levels of the fitted value. From our findings and the warning given by the linear mixed model fit of the predictor variables of being possibly collinear, we have come to conclude that we misrepresented the longitudinal model for prediction as it is unable to capture the true underlying relationship between the predictors and the outcome (death) due to the lack of time-varying variables in our data frames. For possible

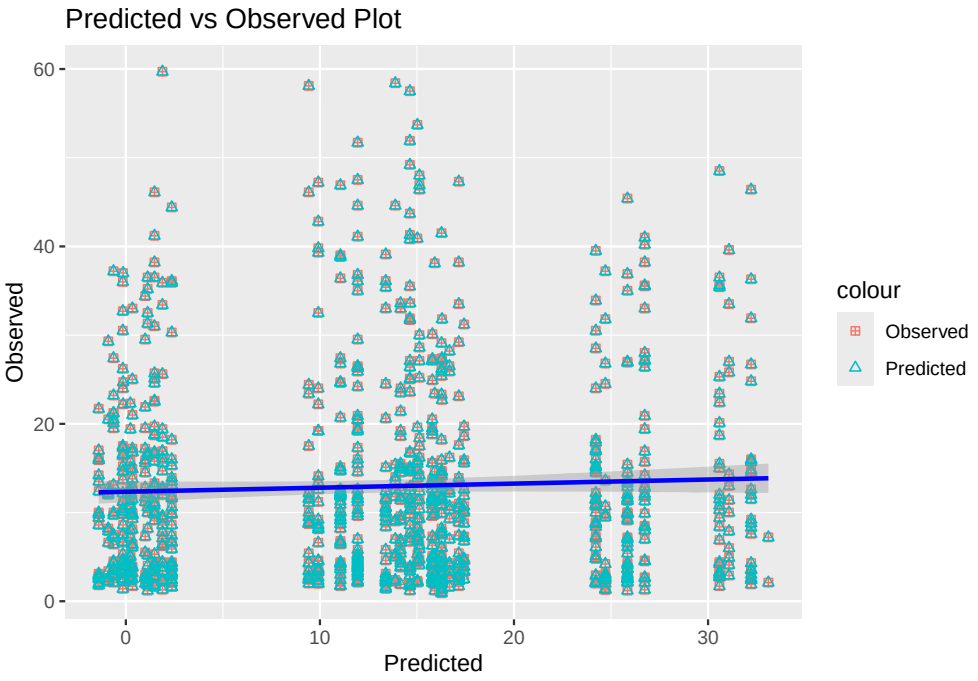

Figure 10: Predicted vs Observed Plot

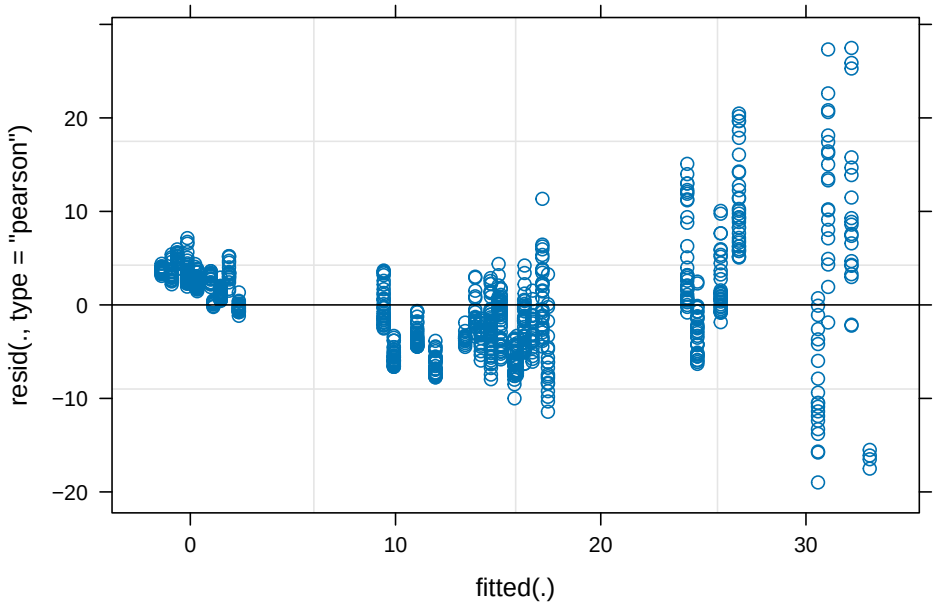

Figure 11: Longitudinal Model Visualization

solutions to fitting this model, we considered addressing the issue of our augmented data frame lacking time-varying variables and then performing the same procedure to test for collinearity and heteroscedasticity. If similar results are obtained, we did more research on the Lasso Regression and Ridge Regression to perform regularization techniques to add penalty terms to shrink the estimated coefficients towards zero to reduce the variances of the parameter estimates and possibily mitigate the effects of collinearity.

Conclusion

Team Contribution Statement