ENV872 Final Project

Keanu Valibia

Spring 2024

Contents

List of Tables

List of Figures

Rationale and Research Questions

This data project utilizes data from the World Bank spanning child mortality rates, total energy consumption, renewable energy consumption, and total greenhouse gas (ghg) emissions. The purpose of this lab is to understand any potential correlation between mortality rates and ghg emissions / energy consumption. Thus, the questions for this project would be:

Question 1: H0 = There is no observable effect of energy consumption, renewable energy consumption, and total ghg emissions on mortality rates (H0 = 0).

Question 2: Ha = There is observable effect of energy consumption, renewable energy consumption, and total ghg emissions on mortality rates (H0 != 0).

The World Bank API is used to pull data sources across the four mentioned variables. This data is used as a reliable source of data that spans across decades of data collection. Mortality rates, (specifically those for under the age of 5 per 1,000 deaths) are used as the independent variable as a proxy for development of a nation. GHG emissions are used as a very rough proxy for not just how pollutant a country may be, but also for level of development. Renewable energy consumption is used as a proxy for how sustainable a country may be, while total energy consumption may be used as another proxy for how productive or developed a country is.

Dataset Information

Table 1: WorldBank Dataset Information

Item	Value
Source	World Bank API
Date	1990 - 2020
Filename	World_Bank_EnergyUse_Mortality.csv

Table 2: WorldBank Dataset Information - Fields

Item	Value
iso2c	2 character country acronym
iso2c	2 character country acronym
country	Country name
date	Year row data was collected
Renewable_Consump	Estimated usage of renewable energies as a percentage of total final energy consumption
Energy_Use	Energy use per country in terms of kg of oil equivalent per capita
Total_GHG_Emissions	Total greenhouse gas emissions as kilotons of CO2 equivalent
Mortality_Rate	Child mortality rates under the age of 5 per 1,000 live births

Exploratory Analysis

```r
#Review first few values of dataframe
glimpse(energyUse_mort)
```

```
## Rows: 186
## Columns: 8
## $ iso2c               <chr> "AE", "AE", "AE", "AE", "AE", "AE", "AE", "AE", "A~
## $ iso3c               <chr> "ARE", "ARE", "ARE", "ARE", "ARE", "ARE", "ARE", "~
## $ country             <chr> "United Arab Emirates", "United Arab Emirates", "U~
## $ date                <dbl> 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 19~
## $ Renewable_Consump   <dbl> 0.00, 0.00, 0.19, 0.15, 0.12, 0.11, 0.08, 0.08, 0.~
## $ Energy_Use          <dbl> 10748.655, 11694.922, 10564.510, 10576.052, 11186.~
## $ Total_GHG_Emissions <dbl> 78601.84, 87044.85, 85014.50, 89207.17, 97918.57, ~
## $ Mortality_Rate      <dbl> 16.5, 15.7, 14.9, 14.2, 13.6, 13.1, 12.6, 12.2, 11~
```

```r
#Fetch dimensions of df
dim(energyUse_mort)
```

```
## [1] 186   8
```

```r
#Fetch summary statistics of df across fields
summary(energyUse_mort)
```

```
##     iso2c              iso3c              country               date     
##  Length:186         Length:186         Length:186         Min.   :1990  
##  Class :character   Class :character   Class :character   1st Qu.:1997  
##  Mode  :character   Mode  :character   Mode  :character   Median :2005  
##                                                           Mean   :2005  
##                                                           3rd Qu.:2013  
##                                                           Max.   :2020  
##                                                                         
##  Renewable_Consump   Energy_Use    Total_GHG_Emissions Mortality_Rate  
##  Min.   : 0.000    Min.   : 1411   Min.   :  78602     Min.   : 2.400  
##  1st Qu.: 1.400    1st Qu.: 2830   1st Qu.: 328642     1st Qu.: 4.725  
##  Median : 4.825    Median : 3711   Median : 520126     Median : 7.150  
##  Mean   : 5.677    Mean   : 4869   Mean   :1525375     Mean   : 8.549  
##  3rd Qu.: 8.805    3rd Qu.: 7662   3rd Qu.:1262850     3rd Qu.:10.025  
##  Max.   :18.690    Max.   :11695   Max.   :6810656     Max.   :28.800  
##                    NA's   :32
```

```r
#Create distrubtion plots of independent and dependent variables and review for heteroskedasticity.

mortality_dist <- ggplot(energyUse_mort, aes(x = Mortality_Rate)) +
  geom_histogram() +
  labs(title = "Energy Consumption Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

mortality_dist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

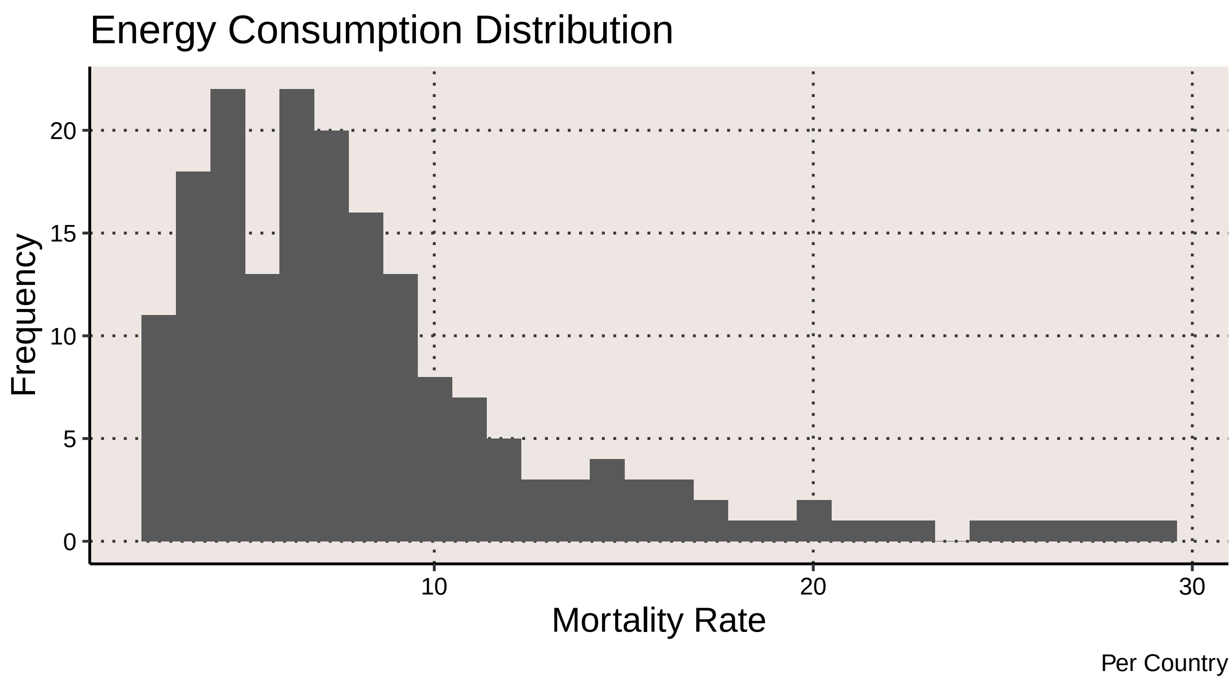

Figure 1: Mortality Distribution Analysis

```r
ghg_dist <- ggplot(energyUse_mort, aes(x = Total_GHG_Emissions)) +
  geom_histogram() +
  labs(title = "GHG Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

ghg_dist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
EnergyUse_dist <- ggplot(energyUse_mort, aes(x = Energy_Use)) +
  geom_histogram() +
  labs(title = "Total Energy Use",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

EnergyUse_dist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

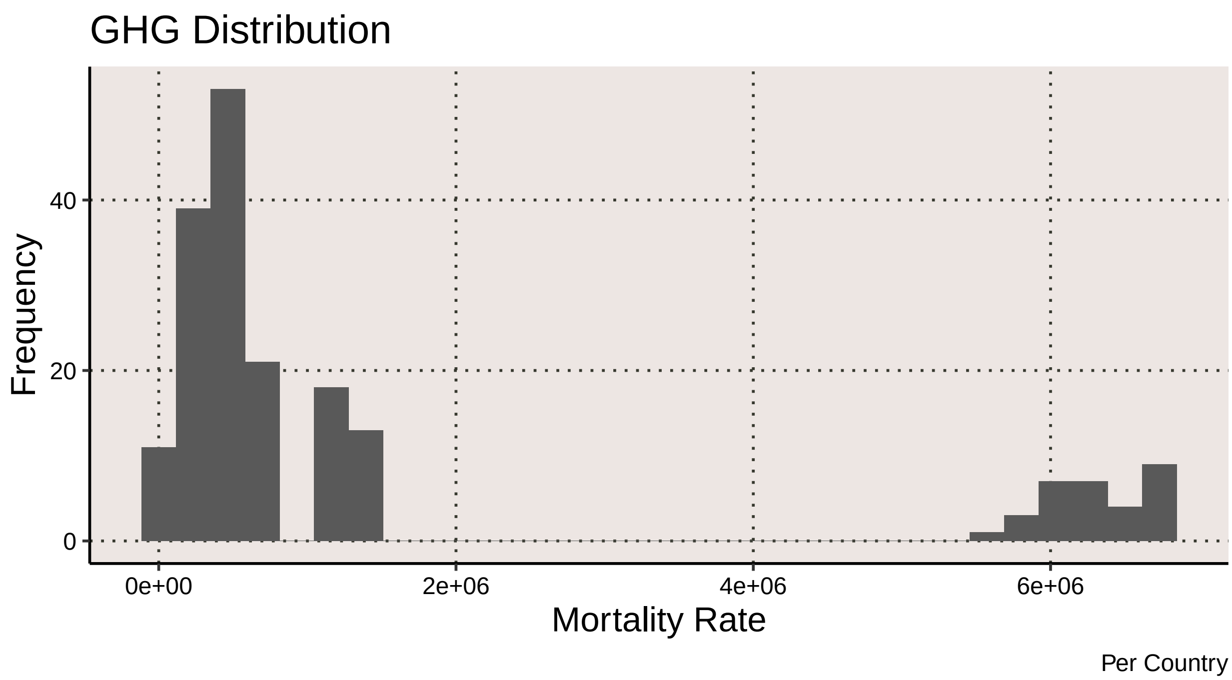

Figure 2: GHG Distribution Analysis

```
## Warning: Removed 32 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

```
Renewable_Consump_dist <- ggplot(energyUse_mort, aes(x = Renewable_Consump)) +
  geom_histogram() +
  labs(title = "Renewable Energy Use Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

Renewable_Consump_dist
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
#Create logged versions of histograms and review for normalized graphs
mortality_dist <- ggplot(energyUse_mort, aes(x = log(Mortality_Rate))) +
  geom_histogram() +
  labs(title = "Logged Mortality Rate Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

mortality_dist
```

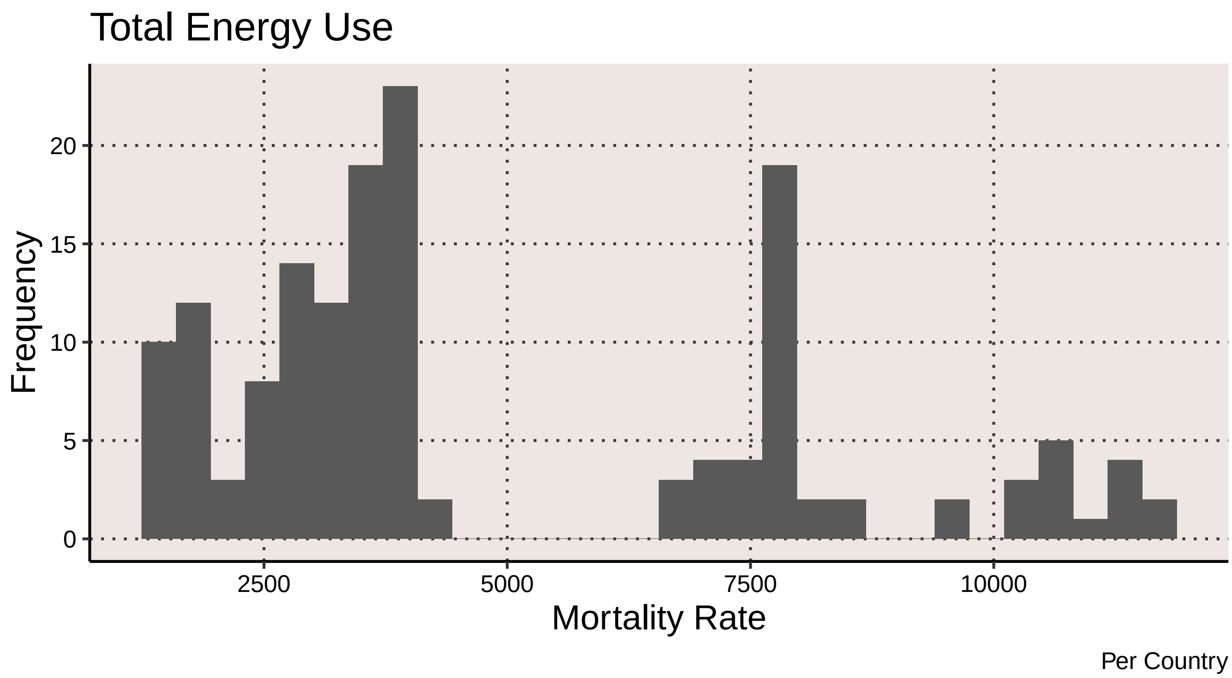

Figure 3: Total Energy Use Distribution Analysis

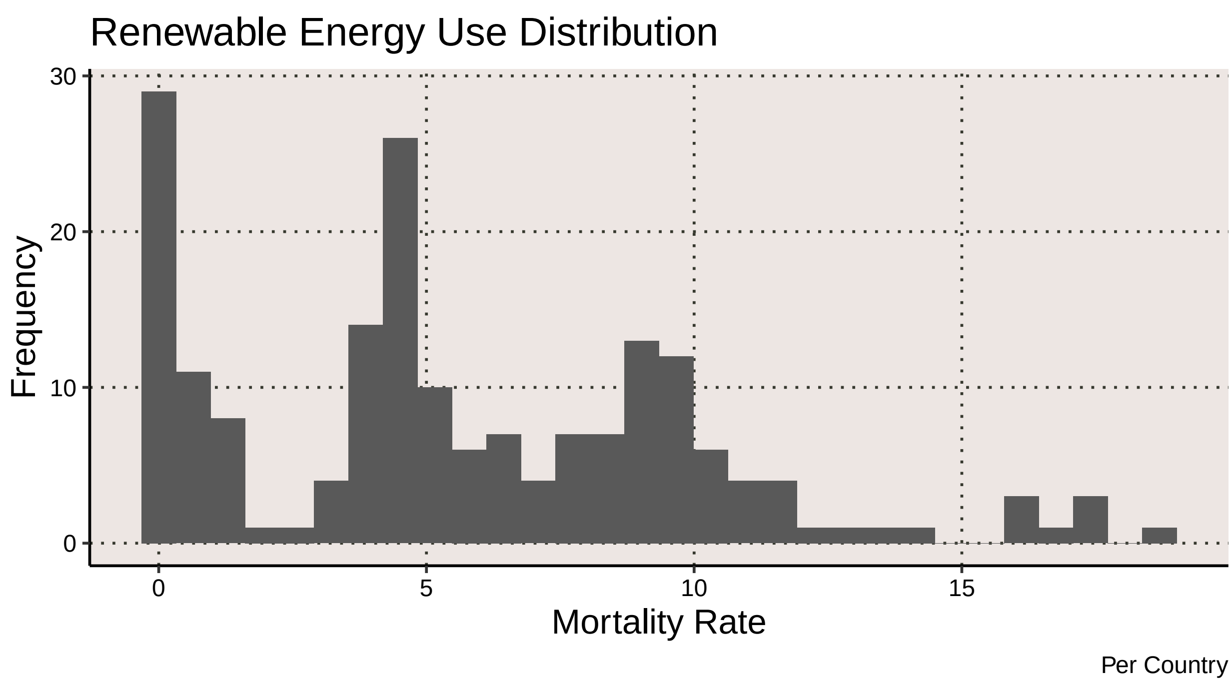

Figure 4: Renewable Distribution Analysis

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

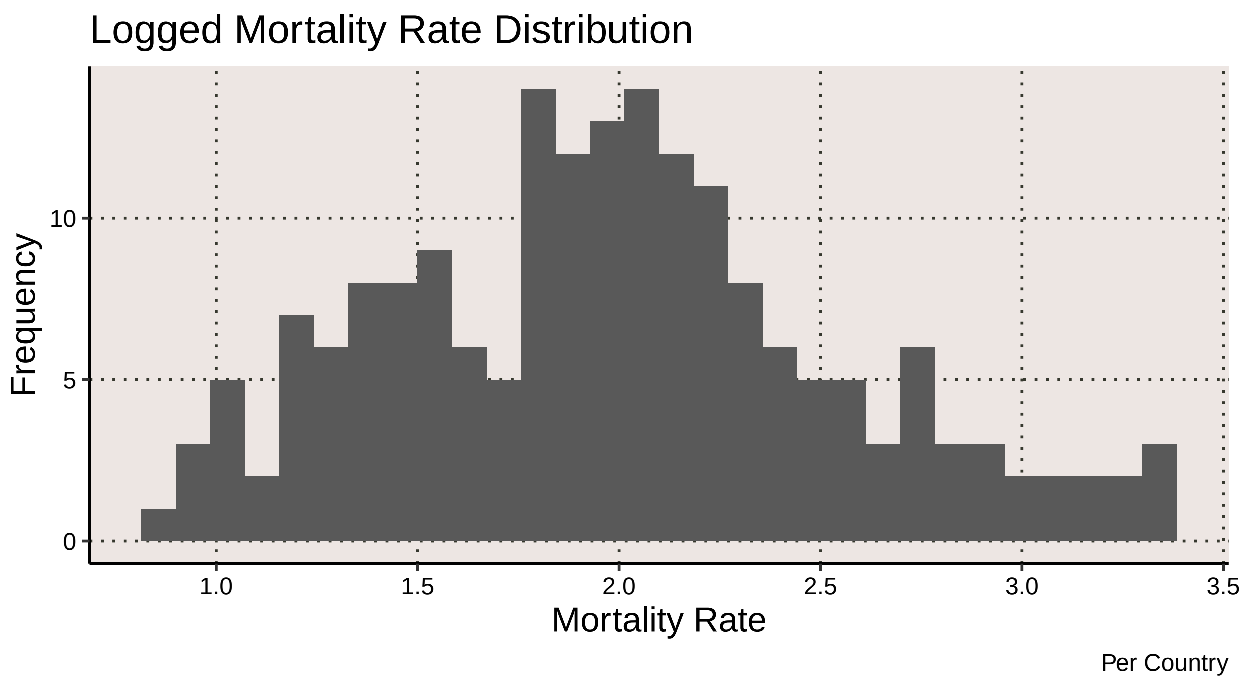

Figure 5: Mortality Logged Distribution Analysis

```r
ghg_dist <- ggplot(energyUse_mort, aes(x = log(Total_GHG_Emissions))) +
  geom_histogram()  +
  labs(title = "Logged GHG Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

ghg_dist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
EnergyUse_dist <- ggplot(energyUse_mort, aes(x = log(Energy_Use))) +
  geom_histogram() +
  labs(title = "Total Energy Used Logged Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

EnergyUse_dist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

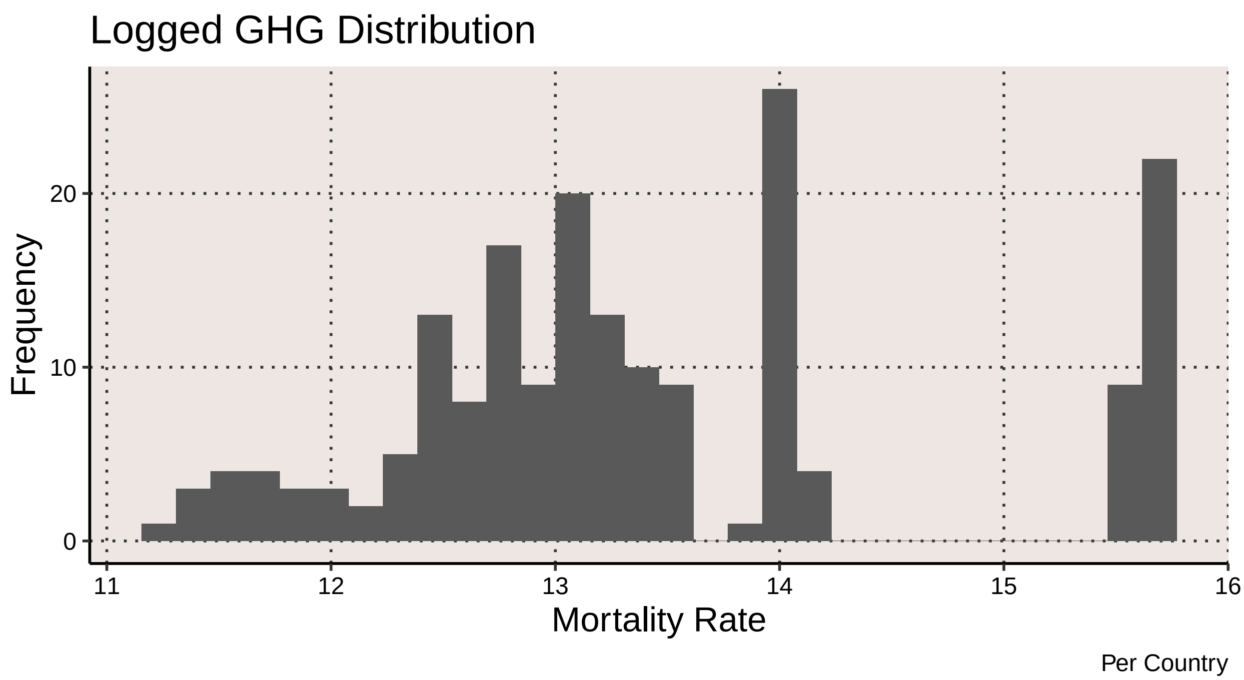

Figure 6: GHG Logged Distribution Analysis

```
## Warning: Removed 32 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

```
Renewable_Consump_dist <- ggplot(energyUse_mort, aes(x = log(Renewable_Consump))) +
  geom_histogram() +
  labs(title = "Renewable Energy Logged Distribution",
       caption = "Per Country",
       color = "Country") +
  xlab("Mortality Rate") +
  ylab("Frequency")

Renewable_Consump_dist
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

```
#Create scatterplots to review changes in independent variables over time.

renewable_consump_time <- ggplot(energyUse_mort, aes(x = date, y = Renewable_Consump, color = country))
  geom_point() +
  labs(title = "Renewable Energy Consumption Over Time",
       caption = "Per Country",
       color = "Country") +
```

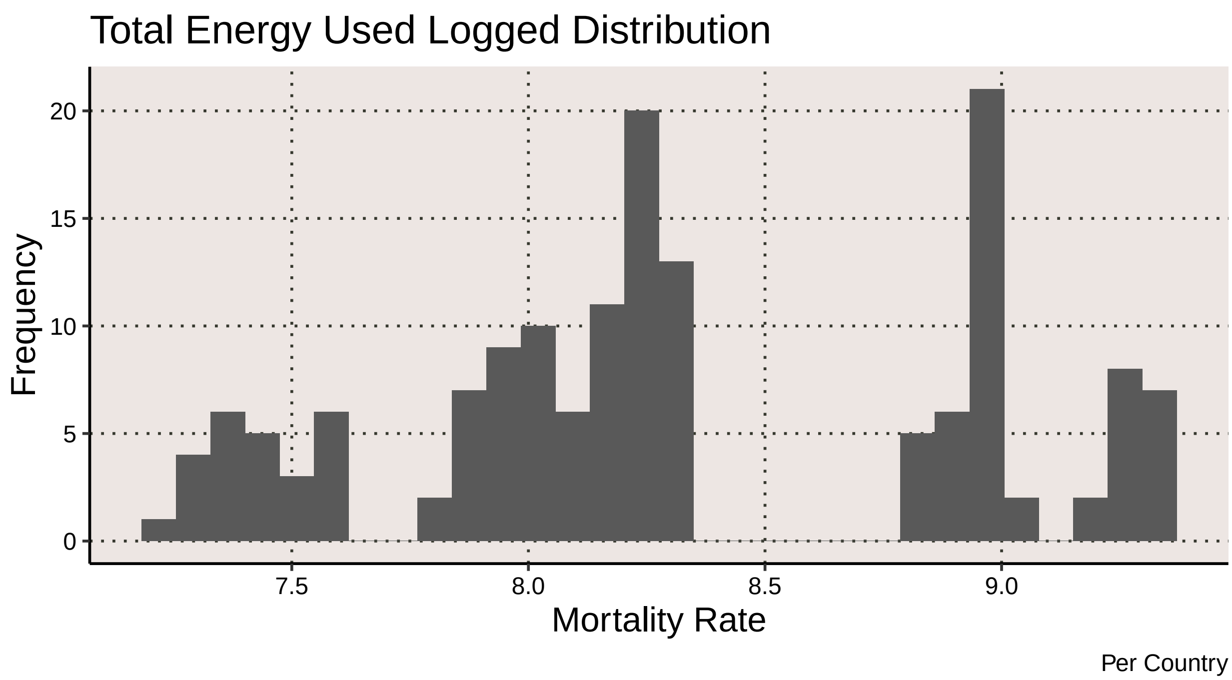

Figure 7: Total Energy Distribution Analysis

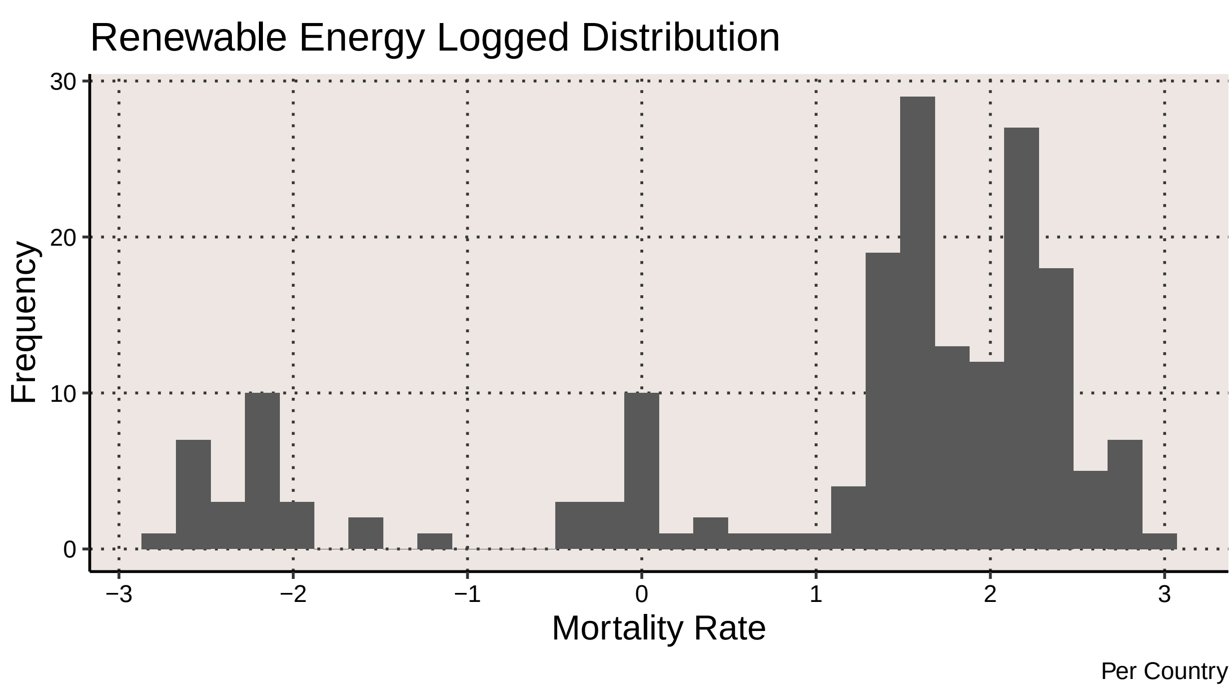

Figure 8: Renewable Energy Distribution Analysis

```
  xlab("Year") +
  ylab("Renewable Energy Used (% of Total Use")

renewable_consump_time
```

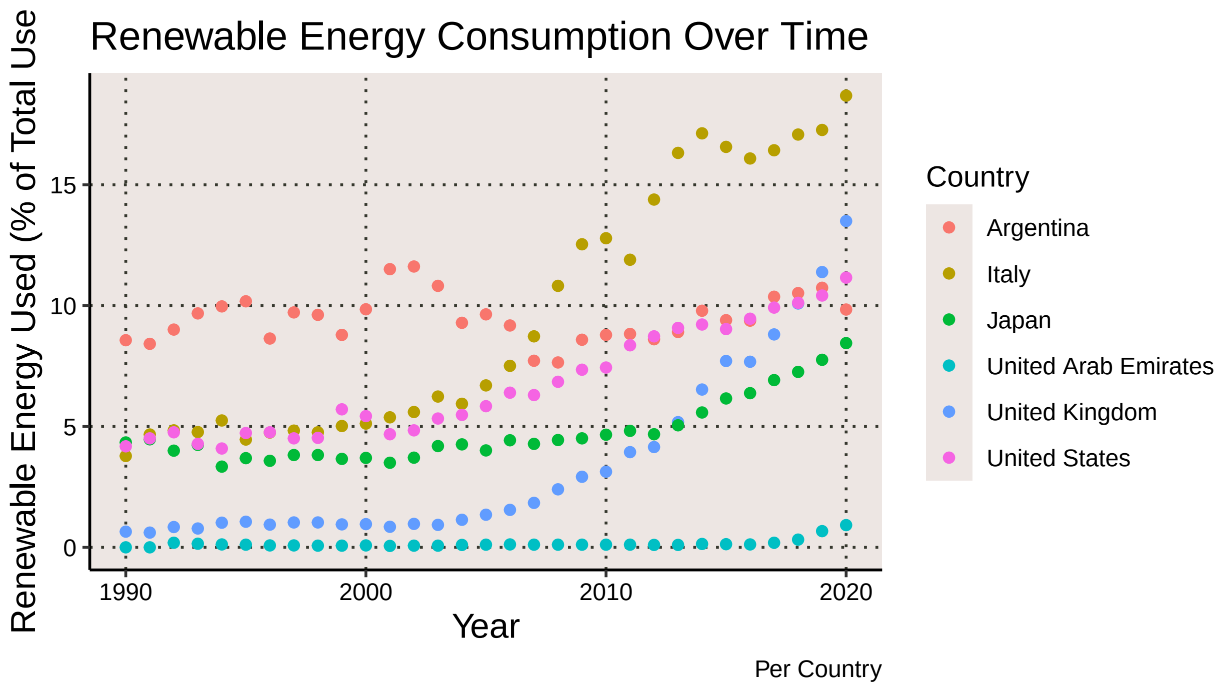

Figure 9: Renewable Energy Scatterplot

```
energy_consump_time <- ggplot(energyUse_mort, aes(x = date, y = Energy_Use, color = country)) +
  geom_point() +
  labs(title = "Energy Consumption as kg of Oil Equivalent per Capita",
       caption = "Per Country",
       color = "Country") +
  xlab("Year") +
  ylab("Energy Consumption")

energy_consump_time
```

```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
ghg_emissions_time <- ggplot(energyUse_mort, aes(x = date, y = Total_GHG_Emissions, color = country)) +
  geom_point() +
  labs(title = "Total GhG Emissions Over Time (kt of CO2 Equivalent)",
       caption = "Per Country",
       color = "Country") +
  xlab("Year") +
  ylab("GhG Emissions")
```

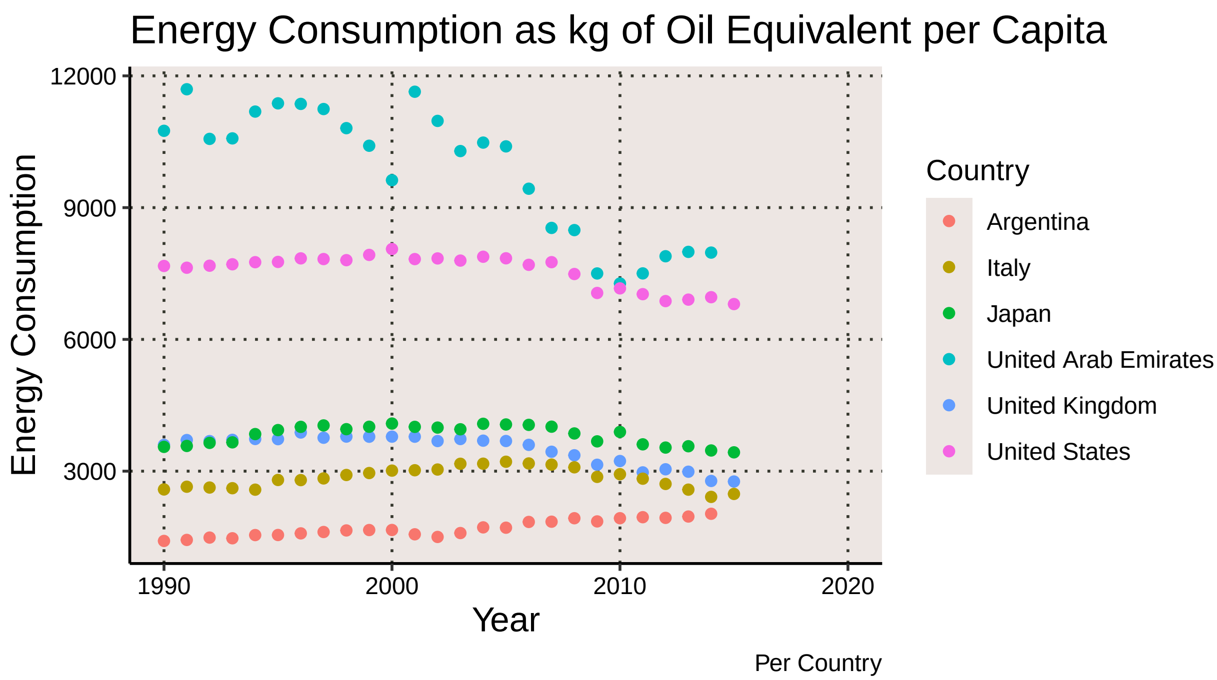

Figure 10: Total Energy Use Scatterplot

```
ghg_emissions_time
```

```
mortality_time <- ggplot(energyUse_mort, aes(x = date, y = Mortality_Rate, color = country)) +
  geom_point() +
  labs(title = "Mortality Rate (Under 5 per 1,000 Births) per Country Over Time",
       caption = "Per Country",
       color = "Country") +
  xlab("Year") +
  ylab("Mortality Rate")

mortality_time
```

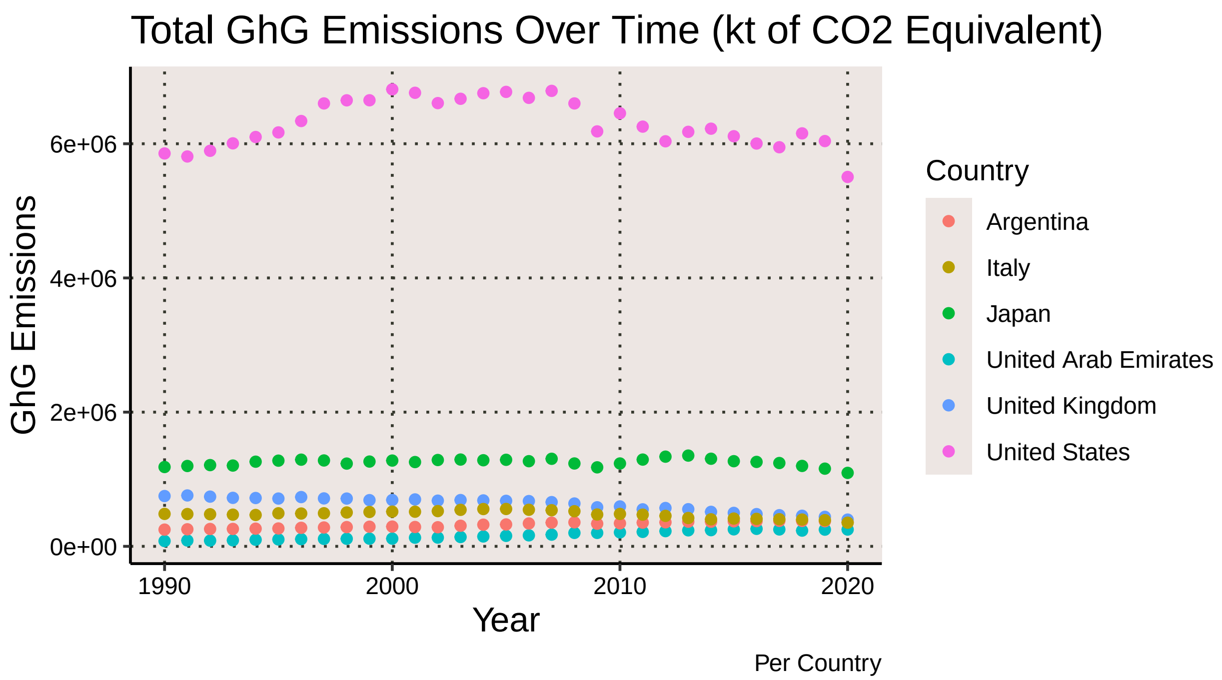

Figure 11: GHG Scatterplot

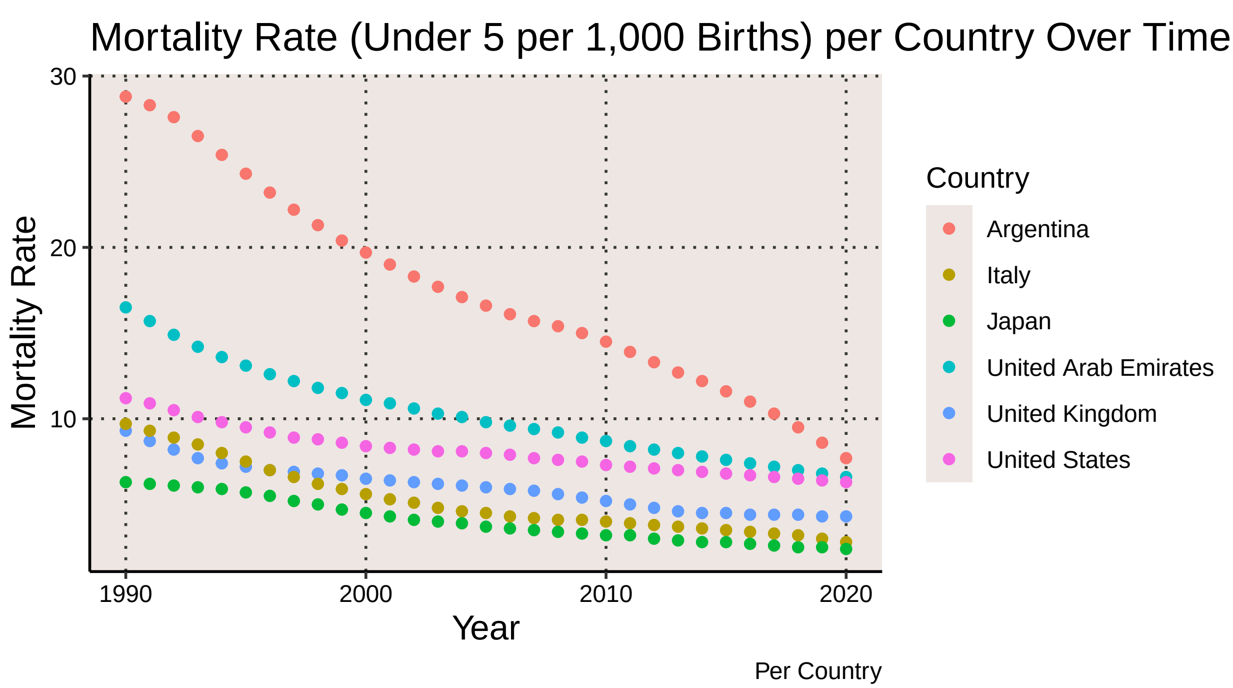

Figure 12: Mortality Scatterplot

Analysis

```r
#Use AIC function to review which independent variables can be used for multiple regression.
AIC <- lm(data = energyUse_mort, log(Mortality_Rate) ~ Renewable_Consump + Energy_Use + Total_GHG_Emiss
AIC
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ Renewable_Consump + Energy_Use +
##     Total_GHG_Emissions, data = energyUse_mort)
##
## Coefficients:
##         (Intercept)    Renewable_Consump           Energy_Use
##           1.670e+00            3.391e-02            6.009e-05
## Total_GHG_Emissions
##          -4.932e-08
```

```r
step(AIC)
```

```
## Start:  AIC=-186.93
## log(Mortality_Rate) ~ Renewable_Consump + Energy_Use + Total_GHG_Emissions
##
##                       Df Sum of Sq    RSS     AIC
## <none>                              43.432 -186.93
## - Total_GHG_Emissions  1    1.3188 44.751 -184.32
## - Renewable_Consump    1    1.4736 44.906 -183.79
## - Energy_Use           1    2.3502 45.783 -180.81
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ Renewable_Consump + Energy_Use +
##     Total_GHG_Emissions, data = energyUse_mort)
##
## Coefficients:
##         (Intercept)    Renewable_Consump           Energy_Use
##           1.670e+00            3.391e-02            6.009e-05
## Total_GHG_Emissions
##          -4.932e-08
```

```r
# Run multiple regressions
AICmodel <- lm(data = energyUse_mort,  log(Mortality_Rate) ~ Renewable_Consump + Energy_Use + Total_GHG_
summary(AICmodel)
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ Renewable_Consump + Energy_Use +
##     Total_GHG_Emissions, data = energyUse_mort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10766 -0.32088  0.01821  0.26459  1.32740
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.670e+00  1.528e-01  10.931   <2e-16 ***
## Renewable_Consump   3.391e-02  1.503e-02   2.256   0.0255 *
## Energy_Use          6.009e-05  2.109e-05   2.849   0.0050 **
## Total_GHG_Emissions -4.932e-08  2.311e-08  -2.134   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5381 on 150 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.05606,    Adjusted R-squared:  0.03718
## F-statistic: 2.969 on 3 and 150 DF,  p-value: 0.03383
```

```
AICmodel
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ Renewable_Consump + Energy_Use +
##     Total_GHG_Emissions, data = energyUse_mort)
##
## Coefficients:
##         (Intercept)    Renewable_Consump           Energy_Use
##           1.670e+00            3.391e-02            6.009e-05
## Total_GHG_Emissions
##          -4.932e-08
```

Question 1: H0 = There is no observable effect of energy consumption, renewable energy consumption, and total ghg emissions on mortality rates (H0 = 0).

Question 2: Ha = There is observable effect of energy consumption, renewable energy consumption, and total ghg emissions on mortality rates (H0 != 0).

Summary and Conclusions

Although there is enough evidence to reject the null hypothesis, the regression coefficients seem to oddly indicate that increases in both renewable energy consumption and total energy use seems to correlate to increased mortality rates, while a decrease in total greenhouse gas emissions seem to correlate with a decrease in mortality rates for every increase in total greenhouse gas emissions.

All of these regression coefficients are statistically significant, with renewable consumption and total ghg emissions significant at the 5 percent level, and energy use significant at the 1 percent level.

However, there are a few things to point out. Quantitatively, the R-Squared is only 0.056, which means that only about 5.6% of variance in the model is explaned by this model, indicating drastic underfit of data. Additional data will be needed to raise this R-Squared to an acceptable level without overfitting the data.

Qualitatively, there are quite a few exogenous factors missing from the model that may help explain or even drastically change these results. Perhaps increases in energy consumption and energy use may correlate to higher mortality rates simply because of larger populations. Perhaps there are other factors within individual countries' economies that may explain higher mortality rates such as crime rates, poverty rates, etc. Negative correlation between mortality rates and ghg emissions may potentially be explained better by combination of factors including total population, total gdp, etc. Countries with higher greenhouse gas emissions may emit so much more than less developed countries that the data may be skewed. Industrial countries that pollute much more than smaller, less-developed countries tend to have more advanced infrastructure to support advanced medical facilities, etc. Thus, mortality rates cannot be explained soley by these three factors alone.

github repository link: https://github.com/ItsTheKGV/Spring24-ENV872-Final/tree/main

References

World Bank Data Mortality Rate: https://data.worldbank.org/indicator/SH.DYN.MORT Renewable Energy Consumption: https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS?view=chart Energy Use: https://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE?view=chart Total Greenhouse Gas Emissions: https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE?view=chart