Data Wrangling (Data Preprocessing)

Practical Assessment 2

Kavinda Vihan Goonesekere

26-05-2023

Setup

```
# Load the necessary packages required to reproduce the report. For example:

library(kableExtra)
library(editrules)
library(magrittr)
library(stringr)
library(ggplot2)
library(ggpubr)
library(dplyr)
library(tidyr)
```

Student names, numbers and percentage of contributions

Group Information

Student name	Student number	Percentage of contribution
Kavinda Vihan Goonesekere	S3987368	100%

Executive Summary

The pre-processing performed in this report attempts to combine **mortality**, a dataset containing death rate estimates for various causes of death by US state, and **county_data**, a dataset containing socioeconomic variables for all the US counties. This requires the following pre-processing steps:

- Performing the relevant type conversions and converting categorical variables to factors/ordered factors
- Verifying the way time periods in **mortality** are defined and using it to justify filtering **mortality** to 12-month aggregate records only.
- Filtering **mortality** to only crude death rates (instead of age adjusted) within 2019
- Removing NULL values from both **mortality** and **county_data**
- Correctly applying summary functions to **county_data** and group by state to make data less granular
- Pivoting gender and state columns in **mortality** to single columns
- Removing unnecessary columns from **mortality** after pre-processing

- Pivoting gender and ethnicity columns in **county_data** to single columns
- Fix string columns in both **county_data** and **mortality**
- Creating percentage columns for employed and gender population variables in **county_data**
- Scaling percentage values from 0-100 to 0-1
- Joining **county_data** and **mortality** on their common columns
- Retaining only complete cases from the merged dataframe
- Checking for rule violations from a preset rule list
- Plotting boxplots to identify outliers and removing any, if necessary
- Transforming right-skewed population data to resemble a normal distribution by using the *ln()* transformation

Data

The data pre-processing conducted in this report attempts to combine socioeconomic factors such as income, poverty rates, and ethnicity with the death rate estimates for the 15 leading causes of death in the US by state. This should hopefully provide a clearer picture as to whether there is a correlation between the above factors and various causes of death across US states. The creation of this dataset requires the two datasets listed below:

- **NCHS - VSRR Quarterly provisional estimates for selected indicators of mortality**
  This dataset, originally sourced from Healthdata.gov, contains provisional estimates of death for the 15 leading causes of death in the United States (Centers for Disease Control and Prevention, 2016). In addition to these, estimates are given for deaths caused by drug overdoses, falls (for those aged 65 and above), HIV, homicide, and deaths related to firearms. Estimates are given from 2019 Quarter 1 till 2022 Quarter 3. The variables in this dataset are discussed below:

  Year and Quarter: Contains the year and quarter for which the estimate is valid (eg: “2019 Q1”)
  Time Period: The time period over which the estimate is valid
  Cause of Death: The cause of death for the given estimate
  Rate Type: One of two categories, “Crude” for which estimates are further broken down into age groups, and “Age adjusted” for which there is no additional breakdown by age
  Unit: Unit for estimates (all estimates are given as “Deaths per 100,000”)
  Overall Rate: Overall estimate for death rate
  Rate Sex Female: Death rate estimate for females
  Rate Sex Male: Death rate estimate for males
  Rate Age 1-4 → Rate Age 85 plus: 10 columns, each of which breaks down death rate estimates across 10 age ranges. Contains NULL if “Rate Type” column is “Age adjusted”
  Rate Alaska → Rate Wyoming: 51 columns, each of which breaks down death rate estimates across the 51 US states

- **ACS county data**:
  This dataset, sourced from the American Community Survey (ACS), contains county-level data on various demographics such as gender and ethnicity, along with information on income, occupation, unemployment, and poverty for the year 2017 (MuonNeutrino, 2019). Since the objective of the dataset is to correlate death rates with socioeconomic indicators, variables related to personal transportation methods and occupation types were dropped in favour of variables related to gender, ethnicity, poverty,

and unemployment rates. The variables selected from this dataset are discussed below:

CountyId: FIP code for US county
State: Name of US state for the specified county
County: Name of US county
TotalPop: Total population of county
Men: Total population of men in county
Women: Total population of women in county
White: Percentage of county population that is white
Black: Percentage of county population that is black
Native: Percentage of county population that is native american
Asian: Percentage of county population that is asian
Pacific: Percentage of county population that is pacific islander
Income: Average income for county
Poverty: Percentage of county population that is in poverty
ChildPoverty: Percentage of children in county experiencing poverty
Employed: Total population of county that is employed
Unemployment: Percentage of county population that is unemployed

```
# Import the data, provide your R codes here.

setwd("C:/Work/Master in Analytics/Semester 1/Data Wrangling MATH2349/Assessment 2")

mortality <- read.csv("indicators_of_mortality.csv", )
county_data <- read.csv("acs2017_county_data.csv")[ ,c('CountyId', 'State', 'County', 'TotalP
op', 'Men', 'Women', 'White', 'Black', 'Native', 'Asian', 'Pacific', 'Income', 'Poverty', 'Ch
ildPoverty', 'Employed', 'Unemployment')]

# glance at data
head(mortality)
```

	Year.and.Quarter <chr>	**Time.Period** <chr>
1	2019 Q1	12 months ending with quarter
2	2019 Q1	12 months ending with quarter
3	2019 Q1	12 months ending with quarter
4	2019 Q1	12 months ending with quarter
5	2019 Q1	12 months ending with quarter
6	2019 Q1	12 months ending with quarter

6 rows | 1-3 of 70 columns

```
head(county_data)
```

	CountyId	State	County	TotalPop	Men	Women	White	Black	Native
	<int>	<chr>	<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>
1	1001	Alabama	Autauga County	55036	26899	28137	75.4	18.9	0.3
2	1003	Alabama	Baldwin County	203360	99527	103833	83.1	9.5	0.8
3	1005	Alabama	Barbour County	26201	13976	12225	45.7	47.8	0.2
4	1007	Alabama	Bibb County	22580	12251	10329	74.6	22.0	0.4
5	1009	Alabama	Blount County	57667	28490	29177	87.4	1.5	0.3
6	1011	Alabama	Bullock County	10478	5616	4862	21.6	75.6	1.0

6 rows | 1-10 of 17 columns

Understand

Checking the structure of **mortality** shows that the categorical variables are read in as character columns and all the death rates are read as numeric. The categorical variables are identified and converted to factors in the subsequent step. The numeric type is suitable for death rates since they are all decimal values.

Checking the structure of **county_data** shows that the **CountyId** is read as an integer while **State** and **County** are read in as character. All three are subsequently converted to factors since they represent categorical variables. **TotalPop**, **Men**, **Women**, **Income**, and **Employed** are read in as integers, which is a suitable format since these columns all contain whole numbers. Additionally, the remaining columns are read as numeric which is suitable once again, as they are all decimal values representing percentages.

```
# check structure
str(mortality)
```

```
## 'data.frame':    1320 obs. of  69 variables:
##  $ Year.and.Quarter         : chr  "2019 Q1" "2019 Q1" "2019 Q1" "2019 Q1" ...
##  $ Time.Period              : chr  "12 months ending with quarter" "12 months ending with
quarter" "12 months ending with quarter" "12 months ending with quarter" ...
##  $ Cause.of.Death           : chr  "All causes" "Alzheimer disease" "COVID-19" "Cancer"
...
##  $ Rate.Type                : chr  "Age-adjusted" "Age-adjusted" "Age-adjusted" "Age-adjus
ted" ...
##  $ Unit                     : chr  "Deaths per 100,000" "Deaths per 100,000" "Deaths per 1
00,000" "Deaths per 100,000" ...
##  $ Overall.Rate             : num  712.2 29.6 NA 148.1 11 ...
##  $ Rate.Sex.Female          : num  600.3 33.1 NA 127.9 7.7 ...
##  $ Rate.Sex.Male            : num  843.7 23.8 NA 175.4 14.7 ...
##  $ Rate.Age.1.4             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.5.14            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.15.24           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.25.34           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.35.44           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.45.54           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.55.64           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.65.74           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.75.84           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Age.85.plus         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Rate.Alaska              : num  711.3 27.5 NA 144.5 15.8 ...
##  $ Rate.Alabama             : num  899.7 44.2 NA 168 12.8 ...
##  $ Rate.Arkansas            : num  870.7 37.8 NA 168.3 12.7 ...
##  $ Rate.Arizona             : num  655.3 31.8 NA 131.5 13.4 ...
##  $ Rate.California          : num  597.1 36 NA 133.1 12.1 ...
##  $ Rate.Colorado            : num  644.6 29.7 NA 127.1 13.9 ...
##  $ Rate.Connecticut         : num  644.6 18.1 NA 133.8 8 ...
##  $ Rate.District.of.Columbia: num  702.4 13.3 NA 152.5 7.9 ...
##  $ Rate.Delaware            : num  741.4 30.8 NA 157 9 ...
##  $ Rate.Florida             : num  646.1 18.5 NA 140.9 11.4 ...
##  $ Rate.Georgia             : num  771.5 43.5 NA 151.1 10 ...
##  $ Rate.Hawaii              : num  574 20.2 NA 121.4 7.7 ...
##  $ Rate.Iowa                : num  717.6 28.6 NA 153.9 9.3 ...
##  $ Rate.Idaho               : num  716.7 32.7 NA 146.7 11.6 ...
##  $ Rate.Illinois            : num  703.1 24.7 NA 153 9.3 ...
##  $ Rate.Indiana             : num  818 31.2 NA 165.2 12.6 ...
##  $ Rate.Kansas              : num  756.5 22.5 NA 156.8 10.6 ...
##  $ Rate.Kentucky            : num  907.4 31.9 NA 182 13.4 ...
##  $ Rate.Louisiana           : num  856.8 40 NA 170.7 8.9 ...
##  $ Rate.Massachusetts       : num  663 19 NA 143 9.2 30.8 15.3 32 69.3 3.4 ...
##  $ Rate.Maryland            : num  704.6 14.6 NA 149.2 7 ...
##  $ Rate.Maine               : num  758.8 27.3 NA 163.8 9.9 ...
##  $ Rate.Michigan            : num  766.2 33.7 NA 158.2 10.9 ...
##  $ Rate.Minnesota           : num  641 33 NA 143 10 ...
##  $ Rate.Missouri            : num  801.6 32.1 NA 164.7 9.3 ...
##  $ Rate.Mississippi         : num  922.9 45.4 NA 181.2 12.2 ...
##  $ Rate.Montana             : num  731.4 23 NA 143.8 12.4 ...
##  $ Rate.North.Carolina      : num  761.3 36.9 NA 154.1 10.4 ...
##  $ Rate.North.Dakota        : num  680.6 32.3 NA 142.2 14.3 ...
##  $ Rate.Nebraska            : num  709.9 27.2 NA 147.9 11.5 ...
##  $ Rate.New.Hampshire       : num  710 25.2 NA 145.5 11.7 ...
##  $ Rate.New.Jersey          : num  664.3 21.6 NA 140.7 7.8 ...
##  $ Rate.New.Mexico          : num  746 21.8 NA 138.6 25.5 ...
##  $ Rate.Nevada              : num  734 22.5 NA 148.7 13.3 ...
##  $ Rate.New.York            : num  618 13.4 NA 135.8 6.9 ...
##  $ Rate.Ohio                : num  826.7 34 NA 165.2 10.6 ...
##  $ Rate.Oklahoma            : num  878 37.6 NA 177 15.6 ...
##  $ Rate.Oregon              : num  685.7 36.2 NA 147.3 13.2 ...
##  $ Rate.Pennsylvania        : num  750.3 20.6 NA 155.1 8.4 ...
##  $ Rate.Rhode.Island        : num  705.5 27.8 NA 152.4 13.2 ...
##  $ Rate.South.Carolina      : num  805.1 41.2 NA 156.3 12.1 ...
##  $ Rate.South.Dakota        : num  713.8 36.4 NA 146.6 19.4 ...
##  $ Rate.Tennessee           : num  873 43.2 NA 168.3 13.2 ...
##  $ Rate.Texas               : num  713.5 37.3 NA 141.5 13.7 ...
##  $ Rate.Utah                : num  688.6 40.1 NA 121.7 10.1 ...
##  $ Rate.Virginia            : num  700 26.6 NA 147.5 9.6 ...
##  $ Rate.Vermont             : num  696.2 36.8 NA 154.2 8 ...
##  $ Rate.Washington          : num  664.3 44.2 NA 145 11.7 ...
##  $ Rate.Wisconsin           : num  710 30.9 NA 149.7 10.3 ...
##  $ Rate.West.Virginia       : num  940 30.7 NA 176.4 15.8 ...
##  $ Rate.Wyoming             : num  743.2 36 NA 140.8 22.1 ...
```

```
str(county_data)
```

```
## 'data.frame':    3220 obs. of  16 variables:
##  $ CountyId    : int  1001 1003 1005 1007 1009 1011 1013 1015 1017 1019 ...
##  $ State       : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ County      : chr  "Autauga County" "Baldwin County" "Barbour County" "Bibb County" ...
##  $ TotalPop    : int  55036 203360 26201 22580 57667 10478 20126 115527 33895 25855 ...
##  $ Men         : int  26899 99527 13976 12251 28490 5616 9416 55593 16320 12862 ...
##  $ Women       : int  28137 103833 12225 10329 29177 4862 10710 59934 17575 12993 ...
##  $ White       : num  75.4 83.1 45.7 74.6 87.4 21.6 52.2 72.7 56.2 91.8 ...
##  $ Black       : num  18.9 9.5 47.8 22 1.5 75.6 44.7 20.4 39.3 5 ...
##  $ Native      : num  0.3 0.8 0.2 0.4 0.3 1 0.1 0.2 0.3 0.5 ...
##  $ Asian       : num  0.9 0.7 0.6 0 0.1 0.7 1.1 1 1 0.1 ...
##  $ Pacific     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Income      : int  55317 52562 33368 43404 47412 29655 36326 43686 37342 40041 ...
##  $ Poverty     : num  13.7 11.8 27.2 15.2 15.6 28.5 24.4 18.6 18.8 16.1 ...
##  $ ChildPoverty: num  20.1 16.1 44.9 26.6 25.4 50.4 34.8 26.6 29.1 20 ...
##  $ Employed    : int  24112 89527 8878 8171 21380 4290 7727 47392 14527 9879 ...
##  $ Unemployment: num  5.2 5.5 12.4 8.2 4.9 12.1 7.6 10.1 6.4 5.3 ...
```

```
# identifying categorical variables and converting to factor
mortality_factors <- c('Time.Period', 'Cause.of.Death', 'Rate.Type', 'Unit')
county_factors <- c('CountyId', 'State', 'County')
mortality[mortality_factors] <- lapply(mortality[mortality_factors], factor)
county_data[county_factors] <- lapply(county_data[county_factors], factor)

# create ordered factor from 'Year and Quarter'
mortality$Year.and.Quarter <- ordered(mortality$Year.and.Quarter, levels =c('2019 Q1', '2019
Q2', '2019 Q3', '2019 Q4', '2020 Q1', '2020 Q2', '2020 Q3', '2020 Q4', '2021 Q1', '2021 Q2',
'2021 Q3', '2021 Q4', '2022 Q1', '2022 Q2', '2022 Q3'))

# check factor conversions
lapply(mortality[mortality_factors], class)
```

```
## $Time.Period
## [1] "factor"
##
## $Cause.of.Death
## [1] "factor"
##
## $Rate.Type
## [1] "factor"
##
## $Unit
## [1] "factor"
```

```
lapply(county_data[county_factors], class)
```

```
## $CountyId
## [1] "factor"
##
## $State
## [1] "factor"
##
## $County
## [1] "factor"
```

```
# create year column
mortality %<>% mutate(., year = as.integer(substr(Year.and.Quarter, 1, 4)))
```

Looking at the **Time Period** column, it is observed that there are two possible values: “3-month period” and “12 months ending with quarter”. This implies that a row where **Time Period** = “12 months ending with quarter” is simply an aggregate (mean) of the 4 quarters that came before where **Time Period** = “3-month period”. For example, the death rate of 2021 Q4 where **Time Period** = “12 months ending with quarter” is the mean of the death rates of 2021 Q1, 2021 Q2, 2021 Q3, and 2021 Q4 where **Time Period** = “3-month period”. If this is the case, it is possible to remove all instances of “3-month period” since this level of granularity is unnecessary for the final dataset. To confirm that this is true, the end-of-year crude death rates for **Cause of Death** = “All causes” are compared to their calculated equivalents as follows:

```
# confirming that the '12 months ending with quarter' time period is aggregated from '3-month period'

# only checking for crude death rates from all causes
mortality.filtered <- filter(mortality, mortality$Rate.Type == 'Crude' & mortality$Cause.of.Death == 'All causes')

# subsetting the original '12 months ending with quarter' data for comparison
twelve.months <- filter(mortality.filtered, mortality.filtered$Time.Period == '12 months ending with quarter' & str_detect(Year.and.Quarter, "Q4")) %>% select(., Year.and.Quarter, Overall.Rate)

# calculating mean of all '3-month period' records by year
calculated <- mortality.filtered %>%
              filter(., mortality.filtered$Time.Period == '3-month period' & year != 2022) %>%
              group_by(year) %>%
              summarise_at(vars(Overall.Rate), list(calculated = mean))

s <- data.frame(cbind(calculated, twelve.months$Overall.Rate))
colnames(s) <- c("**Year**", "**Calculated Rate**", "**Rate From Data**")

s %>% kbl(caption = "**Comparison of calculated vs original death rate**") %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

Comparison of calculated vs original death rate

Year	Calculated Rate	Rate From Data
2019	869.800	869.7
2020	1026.775	1027.0
2021	1043.625	1043.8

From the above, we can confirm that instances of **Time Period** = “12 months ending with quarter” are actually aggregates of the previous 4 quarters where **Time Period** = “3-month period”. Therefore, instances of **Time Period** = “3-month period” are removed to simplify the dataset.

Later on in the “Tidy & Manipulate Data I” section, we drop the age range columns, which removes the need for age adjusted values in the dataset. To account for this, records where **Rate Type** is “Age adjusted” are removed. In addition, the dataset is further filtered to retain only records from 2019 Q4 since not all the quarters are necessary for the final dataset to be analyzed.

```
mortality %<>% filter(., mortality$Time.Period != '3-month period')
mortality %<>% filter(., mortality$Rate.Type != 'Age-adjusted')
mortality %<>% filter(., mortality$Year.and.Quarter == '2019 Q4')
```

The summaries for mortality and county data provide a picture of the variables by producing summary statistics for each column (the output for *summary(mortality)* is trimmed since the dataframe is quite large and the output takes up too much space).

Checking NULL counts for **mortality** shows that the columns which denote death rates by age to be the ones with the most NULLs. This is to be expected since these columns are meant to be NULL when **Rate Type** is “Age adjusted”. The other NULLs seen in the columns with death rates for the 51 states implies that data is unavailable for certain causes of death within certain periods in certain states. **Overall Rate** is observed to have a single NULL value which can be removed.

Checking NULL counts for **county_data** shows that the only **ChildPoverty** has a single NULL field. This record is also filtered out.

```
# check summaries (trimmed output)
output <- capture.output(summary(mortality))
output[1:20]
```

```
##  [1] " Year.and.Quarter                        Time.Period"
##  [2] " 2019 Q4:22       12 months ending with quarter:22  "
##  [3] " 2019 Q1: 0       3-month period               : 0  "
##  [4] " 2019 Q2: 0                                         "
##  [5] " 2019 Q3: 0                                         "
##  [6] " 2020 Q1: 0                                         "
##  [7] " 2020 Q2: 0                                         "
##  [8] " (Other): 0                                         "
##  [9] "                             Cause.of.Death        Rate.Type "
## [10] " All causes                         : 1     Age-adjusted: 0  "
## [11] " Alzheimer disease                  : 1     Crude       :22  "
## [12] " Cancer                             : 1                      "
## [13] " Chronic liver disease and cirrhosis: 1                      "
## [14] " Chronic lower respiratory diseases : 1                      "
## [15] " COVID-19                           : 1                      "
## [16] " (Other)                            :16                      "
## [17] "                 Unit     Overall.Rate    Rate.Sex.Female  Rate.Sex.Male    "
## [18] " Deaths per 100,000:22   Min.   :  1.50   Min.   :  0.80   Min.   :  2.30   "
## [19] "                         1st Qu.: 11.70   1st Qu.:  8.20   1st Qu.: 13.40   "
## [20] "                         Median : 15.70   Median : 14.90   Median : 23.00   "
```

```
summary(county_data)
```

```
##     CountyId         State                      County       TotalPop
##  1001   :   1   Texas   : 254   Washington County:  30   Min.   :      74
##  1003   :   1   Georgia : 159   Jefferson County :  25   1st Qu.:   11214
##  1005   :   1   Virginia: 133   Franklin County  :  24   Median :   25848
##  1007   :   1   Kentucky: 120   Jackson County   :  23   Mean   :  100768
##  1009   :   1   Missouri: 115   Lincoln County   :  23   3rd Qu.:   66608
##  1011   :   1   Kansas  : 105   Madison County   :  19   Max.   :10105722
##  (Other):3214   (Other) :2334   (Other)          :3076
##       Men              Women             White           Black
##  Min.   :     39   Min.   :     35   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:   5646   1st Qu.:   5554   1st Qu.: 63.50   1st Qu.: 0.600
##  Median :  12879   Median :  12994   Median : 83.60   Median : 2.000
##  Mean   :  49588   Mean   :  51180   Mean   : 74.92   Mean   : 8.682
##  3rd Qu.:  33017   3rd Qu.:  33594   3rd Qu.: 92.80   3rd Qu.: 9.500
##  Max.   :4979641   Max.   :5126081   Max.   :100.00   Max.   :86.900
##
##      Native           Asian          Pacific             Income
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.00000   Min.   : 11680
##  1st Qu.: 0.100   1st Qu.: 0.200   1st Qu.: 0.00000   1st Qu.: 40622
##  Median : 0.300   Median : 0.600   Median : 0.00000   Median : 47637
##  Mean   : 1.768   Mean   : 1.289   Mean   : 0.08342   Mean   : 48995
##  3rd Qu.: 0.600   3rd Qu.: 1.200   3rd Qu.: 0.10000   3rd Qu.: 55476
##  Max.   :90.300   Max.   :41.800   Max.   :33.70000   Max.   :129588
##
##     Poverty       ChildPoverty      Employed        Unemployment
##  Min.   : 2.40   Min.   : 0.00   Min.   :     39   Min.   : 0.000
##  1st Qu.:11.47   1st Qu.:14.90   1st Qu.:   4573   1st Qu.: 4.475
##  Median :15.40   Median :21.50   Median :  10612   Median : 6.100
##  Mean   :16.78   Mean   :23.04   Mean   :  47093   Mean   : 6.666
##  3rd Qu.:19.80   3rd Qu.:28.60   3rd Qu.:  28747   3rd Qu.: 8.000
##  Max.   :65.20   Max.   :83.60   Max.   :4805817   Max.   :40.900
##                  NA's   :1
```

```
# check NULL counts
colSums(is.na(mortality))
```

```
##          Year.and.Quarter               Time.Period            Cause.of.Death 
##                         0                         0                         0 
##                 Rate.Type                      Unit              Overall.Rate 
##                         0                         0                         1 
##           Rate.Sex.Female             Rate.Sex.Male              Rate.Age.1.4 
##                         1                         1                        11 
##             Rate.Age.5.14            Rate.Age.15.24            Rate.Age.25.34 
##                         9                         5                         4 
##            Rate.Age.35.44            Rate.Age.45.54            Rate.Age.55.64 
##                         4                         2                         2 
##            Rate.Age.65.74            Rate.Age.75.84          Rate.Age.85.plus 
##                         1                         1                         1 
##               Rate.Alaska              Rate.Alabama             Rate.Arkansas 
##                         2                         1                         1 
##              Rate.Arizona           Rate.California             Rate.Colorado 
##                         1                         1                         1 
##          Rate.Connecticut Rate.District.of.Columbia             Rate.Delaware 
##                         1                         1                         2 
##              Rate.Florida              Rate.Georgia               Rate.Hawaii 
##                         1                         1                         2 
##                 Rate.Iowa                Rate.Idaho             Rate.Illinois 
##                         2                         2                         1 
##              Rate.Indiana               Rate.Kansas             Rate.Kentucky 
##                         1                         1                         1 
##            Rate.Louisiana        Rate.Massachusetts             Rate.Maryland 
##                         1                         1                         1 
##                Rate.Maine             Rate.Michigan            Rate.Minnesota 
##                         2                         1                         1 
##             Rate.Missouri          Rate.Mississippi              Rate.Montana 
##                         1                         1                         2 
##       Rate.North.Carolina         Rate.North.Dakota             Rate.Nebraska 
##                         1                         2                         2 
##        Rate.New.Hampshire           Rate.New.Jersey           Rate.New.Mexico 
##                         2                         1                         2 
##               Rate.Nevada             Rate.New.York                 Rate.Ohio 
##                         1                         1                         1 
##             Rate.Oklahoma               Rate.Oregon         Rate.Pennsylvania 
##                         1                         1                         1 
##         Rate.Rhode.Island       Rate.South.Carolina         Rate.South.Dakota 
##                         2                         1                         2 
##            Rate.Tennessee                Rate.Texas                 Rate.Utah 
##                         1                         1                         2 
##             Rate.Virginia              Rate.Vermont           Rate.Washington 
##                         1                         4                         1 
##            Rate.Wisconsin        Rate.West.Virginia              Rate.Wyoming 
##                         1                         2                         2 
##                      year 
##                         0
```

```
colSums(is.na(county_data))
```

```
##     CountyId        State       County     TotalPop          Men        Women 
##            0            0            0            0            0            0 
##        White        Black       Native        Asian      Pacific       Income 
##            0            0            0            0            0            0 
##      Poverty ChildPoverty     Employed Unemployment 
##            0            1            0            0
```

```
mortality %<>% filter(., !is.na(.$Overall.Rate))
county_data %<>% filter(., !is.na(.$ChildPoverty))
```

Pre-processing Prior to Join

An issue to address prior to joining the two datasets is the fact that **county_data** contains statistics at the county-level while **mortality** contains data at the state-level. Combining these datasets as-is on the state column will produce misleading county-level statistics for death rates. As a result, **county_data** must be aggregated to the state-level before combining the two datasets.

The aggregations performed depends on the column being aggregated. Since **TotalPop**, **Men**, **Women**, and **Employed** describe totals, they must be summed when grouping by state. In contrast, **White**, **Black**, **Native**, **Asian**, **Pacific**, **Poverty**, **ChildPoverty**, and **Unemployment** represent percentages and should therefore be averaged when grouping by state. Similarly, **Income** represents a mean for a particular county and therefore must be averaged when finding the mean income by state. The aggregations discussed above are performed as follows:

```
# selectively aggregating specific columns with specific functions when grouping by state
county_data <- county_data[,-1] %>%
  group_by(State) %>%
  summarise(across(.cols = c(TotalPop, Men, Women, Employed), .fns = sum),
            across(.cols = c(White, Black, Native, Asian, Pacific, Income, Poverty, ChildPove
rty, Unemployment), .fns = mean))
```

Tidy & Manipulate Data I

The **mortality** dataset does not conform to tidy data principles as three variables are spread out over multiple columns instead of having their own distinct column. These three variables are as follows:

- **Rate Sex Female & Rate Sex Male**: These can be combined into a single variable called **gender**
- **Rate Age 1-4 → Rate Age 85 plus**: These can be combined into a single variable called **age.range**
- **Rate Alaska → Rate Wyoming**: These can be combined into a single variable called **state**

Since age information contains many NULLs, these columns are not pivoted as this would result in many rows with NULL values. The remaining columns are made to comply to tidy principles by using the *pivot_longer()* function as follows:

```
# pivot longer on gender columns
mortality %<>%
  pivot_longer(
    cols = c(7:8),
    names_to = 'gender',
    values_to = 'gender.rate'
  )

# removing unnecessary age range columns
mortality <- mortality[,-7:-16]

# pivot longer on state columns
mortality %<>%
  pivot_longer(
    cols = c(7:57),
    names_to = 'State',
    values_to = 'state.rate'
  )

# cleaning up 'gender' and 'age.range' using str_replace_all
mortality$gender %<>% str_replace_all(., c('Rate.Sex.Female' = 'Female', 'Rate.Sex.Male' = 'Male'))

# fix 'state' column
mortality$State %<>% substring(., 6)
mortality$State %<>% gsub('\\.', ' ', .)
```

Once the gender and state columns are all combined into a single column called **gender** and **State** and the age range columns are removed, the **Overall Rate** column loses its meaning as it is defined as the combined rate over states, genders, and ages. Therefore, this column is removed from **mortality**, along with other unnecessary columns with redundant information.

```
# removing 'Overall Rate' and other unnecessary columns
mortality <- mortality[,-c(2, 4, 6, 7)]
```

Similarly, the **county_data** dataset also does not conform to tidy data principles as two variables are spread out over multiple columns instead of having their own distinct column. These two variables are as follows:

- **Men and Women**: These can be combined into a single variable called **gender**
- **Hispanic → Pacific**: These 6 columns can be combined into a single variable called **ethnicity**

The above columns are made to comply to tidy principles by using the *pivot_longer()* function as follows:

```
# pivot longer on gender columns
county_data %<>%
  pivot_longer(
    cols = c(3:4),
    names_to = 'gender',
    values_to = 'gender.pop'
  )

# pivot longer on ethnicity columns
county_data %<>%
  pivot_longer(
    cols = c(4:8),
    names_to = 'ethnicity',
    values_to = 'ethnicity.pct'
  )

# change 'gender' column to be the same as the 'gender' column of 'mortality'
county_data$gender %<>% str_replace_all(., c('Men' = 'Male', 'Women' = 'Female'))
```

Tidy & Manipulate Data II

In **county_data**, most variables are expressed as a percentage of the population. The variables that are not expressed as a percentage of the total are **Employed** and **gender.pop**. New columns can be created to express these values as a percentage of total population by dividing by the **TotalPop** column.

```
# Creating percentage columns for 'Employed' and 'gender.pop'
county_data %<>%
  mutate(., employed.pct = Employed/TotalPop, gender.pct = gender.pop/TotalPop)
```

In addition, the existing columns denoting percentages are divided by 100 so that they range between 0 and 1. This would make any future calculations easier to perform.

```
# Dividing percentage columns by 100 to range between 0 and 1
county_data %<>%
  mutate(
    across(c(5:7, 11),
           .fns = ~./100))
```

Joining mortality and county_data

At this point, the **mortality** and **county_data** are in a suitable condition to be combined into a single dataset. Since both datasets contain two common columns (**State** and **gender**), the merge is performed on both columns. It is important to note that the final dataset is still not fully compliant with tidy data principles since each observation does not have a single row. However, due to the structure of this dataset, no further action can be taken without removing information from the dataset.

```
merged <- merge(mortality, county_data, by = c('State','gender'))
merged
```

State	gender	Year.and.Quarter	Cause.of.Death
<chr>	<chr>	<ord>	<fct>
Alabama	Female	2019 Q4	All causes
Alabama	Female	2019 Q4	All causes
Alabama	Female	2019 Q4	All causes
Alabama	Female	2019 Q4	All causes
Alabama	Female	2019 Q4	All causes
Alabama	Female	2019 Q4	Homicide
Alabama	Female	2019 Q4	Homicide
Alabama	Female	2019 Q4	Homicide
Alabama	Female	2019 Q4	Homicide
Alabama	Female	2019 Q4	Homicide

1-10 of 10,000 rows | 1-4 of 18 columns Previous **1** 2 3 4 5 6 ... 1000 Next

Scan I

Checking NULL counts again after the datasets are combined reveals that only **state.rate** contains NULL values, which are artifacts of the original **mortality** dataset and cannot be avoided. These are removed using the *complete.cases()* function to subset the dataframe. A rule set is defined for **merged** and loaded from a text file to check for violations. Zero violations are observed in this case. Checking summary statistics for **merged** doesn't reveal any obvious inconsistencies (for instance, all percentages are between 0 and 1).

The structure of **rules.txt** is given below:

```
# numerical rules
gender.rate >= 0
gender.rate <= 100000
state.rate >= 0
state.rate <= 100000
Employed <= TotalPop
gender.pop <= TotalPop
Poverty >= 0
Poverty <= 1
ChildPoverty >= 0
ChildPoverty <= 1
Unemployment >= 0
```

Unemployment <= 1
ethnicity.pct >= 0
ethnicity.pct <= 1
employed.pct >= 0
employed.pct <= 1
gender.pct >= 0
gender.pct <= 1

# categorical rules
gender %in% c(‘Male’,‘Female’)

```
# check NULLs
colSums(is.na(merged))
```

```
##           State          gender Year.and.Quarter   Cause.of.Death
##               0               0                0                0
##            Unit     gender.rate       state.rate         TotalPop
##               0               0              190                0
##        Employed          Income          Poverty      ChildPoverty
##               0               0                0                0
##    Unemployment      gender.pop        ethnicity    ethnicity.pct
##               0               0                0                0
##    employed.pct      gender.pct
##               0               0
```

```
# retain only complete cases
merged <- merged[complete.cases(merged), ]

# load rules file and check for violations
Rules <- editfile("rules.txt", type = "all")
summary(violatedEdits(Rules, merged))
```

```
## No violations detected, 0 checks evaluated to NA
```

```
## NULL
```

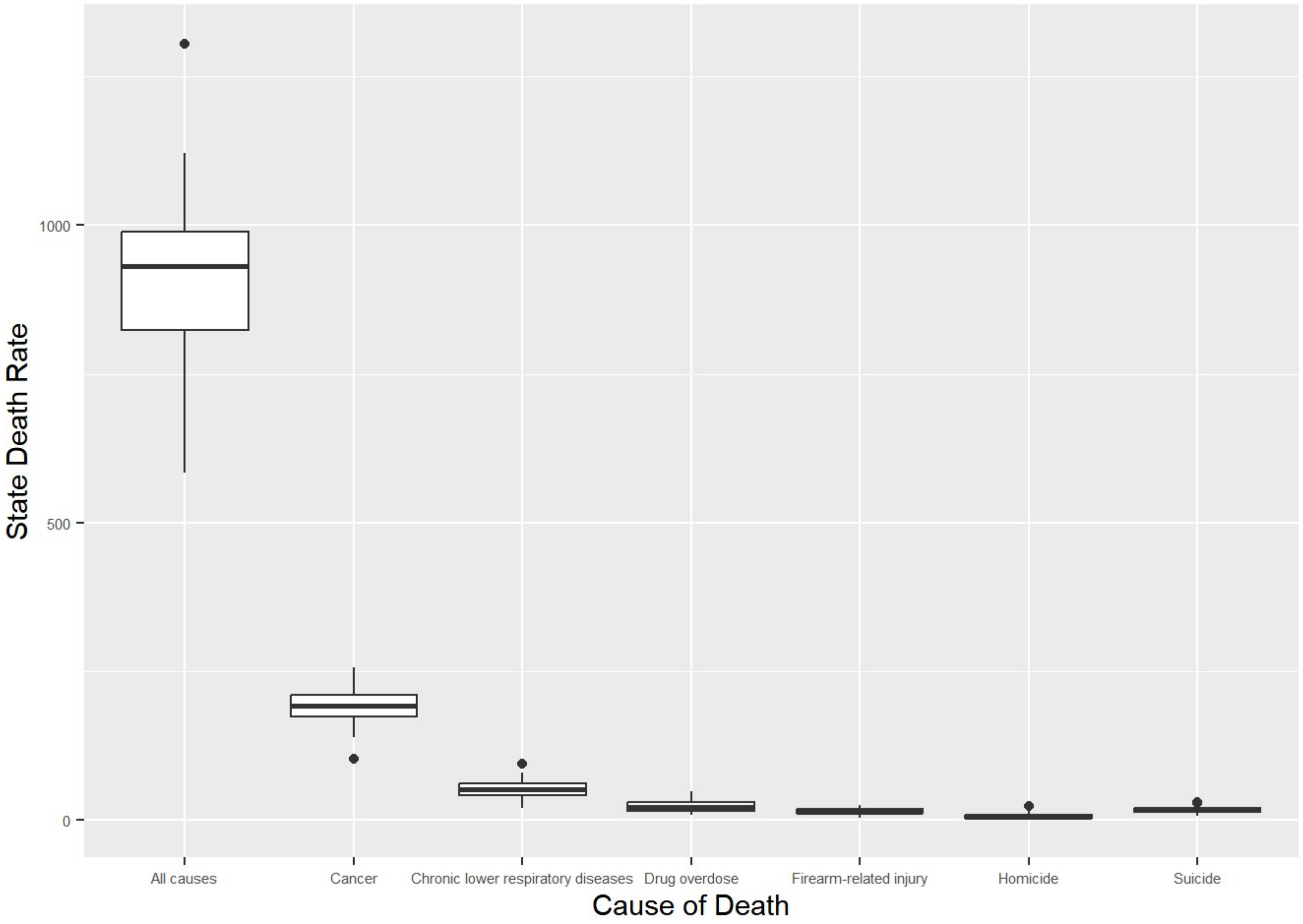
```
summary(merged)
```

```
##      State              gender          Year.and.Quarter
##  Length:10520       Length:10520       2019 Q4:10520    
##  Class :character   Class :character   2019 Q1:    0    
##  Mode  :character   Mode  :character   2019 Q2:    0    
##                                        2019 Q3:    0    
##                                        2020 Q1:    0    
##                                        2020 Q2:    0    
##                                        (Other):    0    
##                               Cause.of.Death                 Unit      
##  All causes                         : 510   Deaths per 100,000:10520  
##  Alzheimer disease                  : 510                             
##  Cancer                             : 510                             
##  Chronic liver disease and cirrhosis: 510                             
##  Chronic lower respiratory diseases : 510                             
##  Diabetes                           : 510                             
##  (Other)                            :7460                             
##   gender.rate       state.rate         TotalPop           Employed       
##  Min.   :  0.80   Min.   :   0.40   Min.   :  583200   Min.   :  293633  
##  1st Qu.: 11.70   1st Qu.:  11.10   1st Qu.: 1836843   1st Qu.:  748658  
##  Median : 21.10   Median :  20.00   Median : 4424376   Median : 1938150  
##  Mean   : 80.81   Mean   :  84.57   Mean   : 6383646   Mean   : 2994515  
##  3rd Qu.: 50.10   3rd Qu.:  52.33   3rd Qu.: 7169967   3rd Qu.: 3525672  
##  Max.   :911.70   Max.   :1305.90   Max.   :38982847   Max.   :17993915  
##                                                                          
##      Income         Poverty        ChildPoverty     Unemployment    
##  Min.   :37019   Min.   :0.0920   Min.   :0.1178   Min.   :0.02728  
##  1st Qu.:45817   1st Qu.:0.1186   1st Qu.:0.1561   1st Qu.:0.05446  
##  Median :51928   Median :0.1437   Median :0.1958   Median :0.06121  
##  Mean   :53618   Mean   :0.1507   Mean   :0.2069   Mean   :0.06423  
##  3rd Qu.:59209   3rd Qu.:0.1740   3rd Qu.:0.2550   3rd Qu.:0.07673  
##  Max.   :77649   Max.   :0.2494   Max.   :0.3513   Max.   :0.10007  
##                                                                     
##    gender.pop        ethnicity         ethnicity.pct       employed.pct   
##  Min.   :  284899   Length:10520       Min.   :0.000000   Min.   :0.4067  
##  1st Qu.:  907621   Class :character   1st Qu.:0.002872   1st Qu.:0.4517  
##  Median : 2245351   Mode  :character   Median :0.012410   Median :0.4727  
##  Mean   : 3191823                      Mean   :0.176903   Mean   :0.4758  
##  3rd Qu.: 3580888                      3rd Qu.:0.161320   3rd Qu.:0.5035  
##  Max.   :19616268                      Max.   :0.944309   Max.   :0.5372  
##                                                                           
##    gender.pct    
##  Min.   :0.4745  
##  1st Qu.:0.4917  
##  Median :0.5000  
##  Mean   :0.5000  
##  3rd Qu.:0.5083  
##  Max.   :0.5255  
## 
```

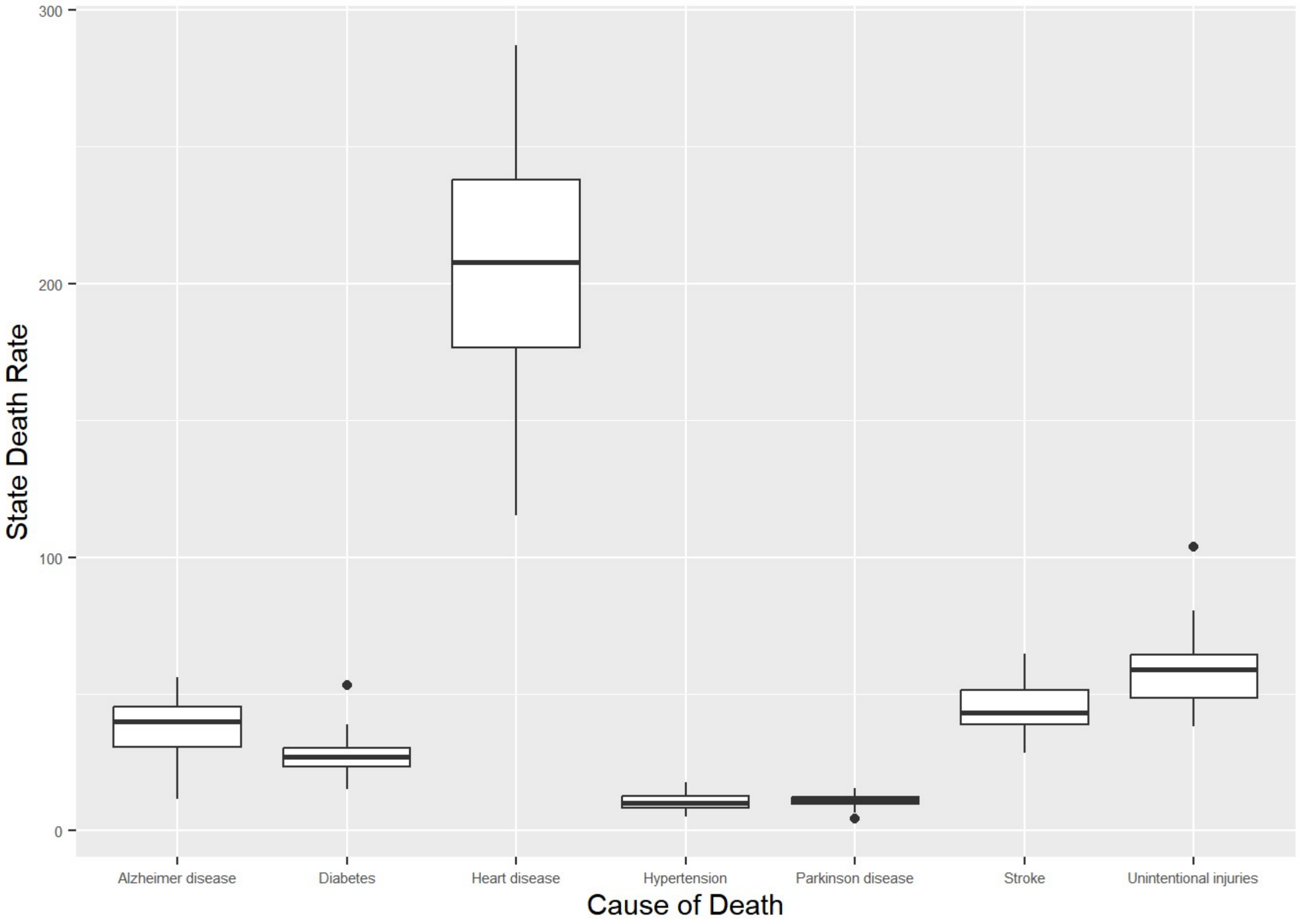
Scan II

Boxplots are plotted for **gender**, **Cause of Death**, **Total Population**, **Poverty**, **Child Poverty**, **Unemployment**, **ethnicity.pct**, **employed.pct**, and **gender.pct** to view potential outliers.

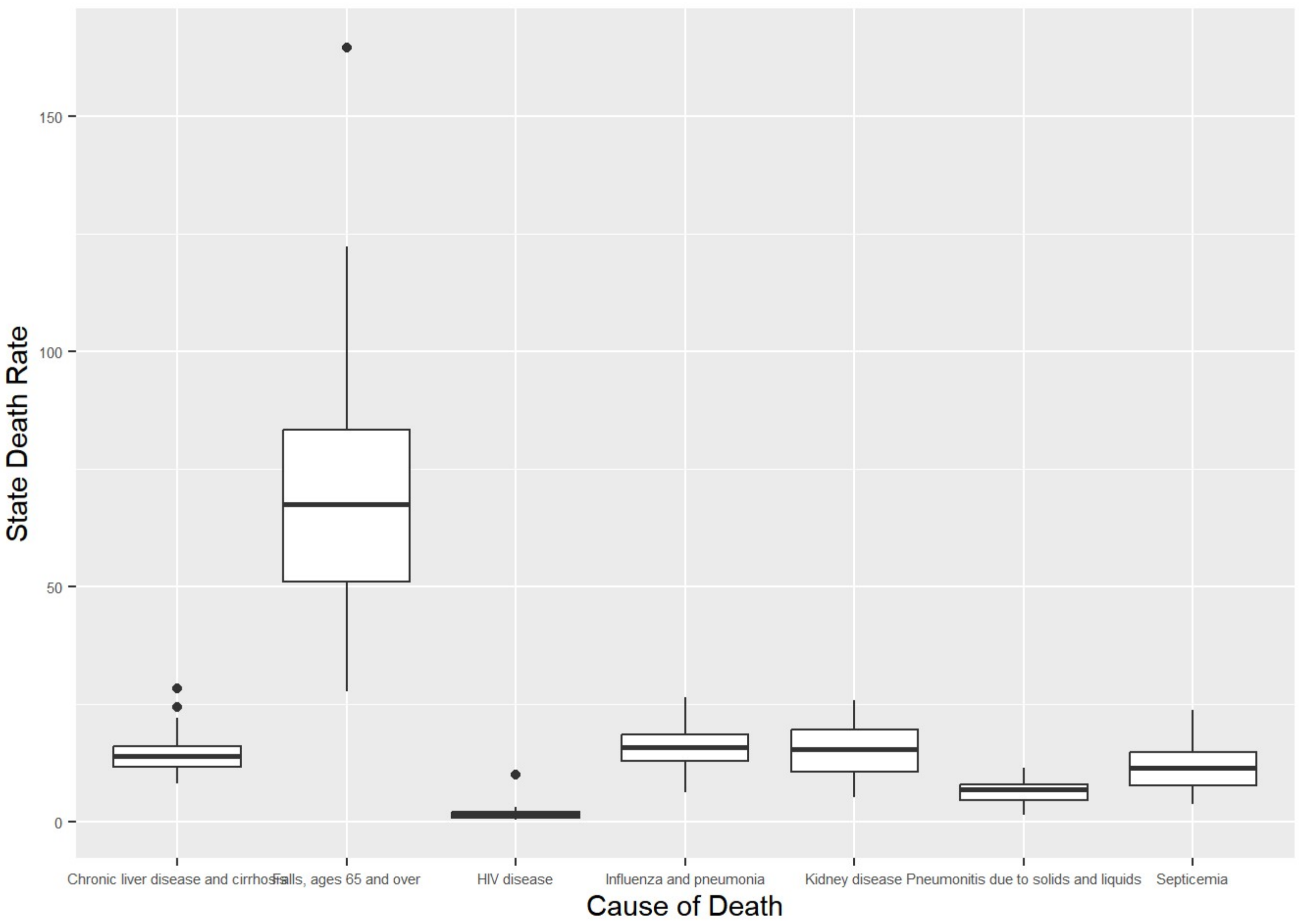
```
options(scipen=5)

# subsetting dataset to make boxplots easier to see
merged1 <- filter(merged, Cause.of.Death %in% c('All causes','Homicide','Firearm-related inju
ry','Drug overdose','Chronic lower respiratory diseases','Suicide','Cancer'))
merged2 <- filter(merged, Cause.of.Death %in% c('Alzheimer disease', 'Diabetes', 'Stroke', 'P
arkinson disease', 'Heart disease', 'Unintentional injuries','Hypertension'))
merged3 <- filter(merged, Cause.of.Death %in% c('Chronic liver disease and cirrhosis','Influe
nza and pneumonia','Septicemia','Kidney disease','Pneumonitis due to solids and liquids','Fal
ls, ages 65 and over','HIV disease'))

# boxplots to view outliers for causes of death
ggplot(merged1, aes(x=Cause.of.Death, y=state.rate)) +
  geom_boxplot() +
  theme(axis.text = element_text(size = 5.5))  +
  xlab("Cause of Death") +
  ylab("State Death Rate")
```



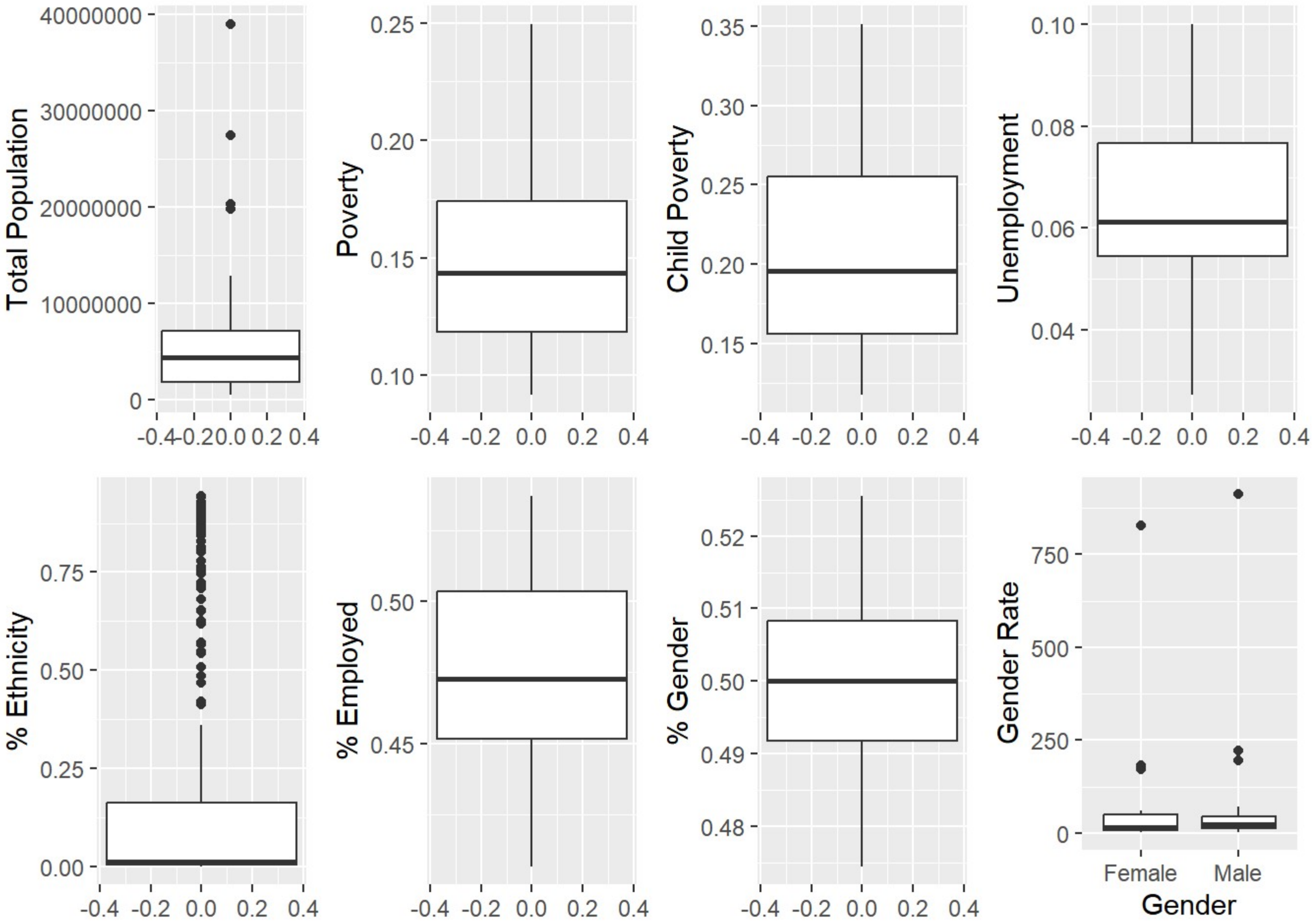
```
ggplot(merged2, aes(x=Cause.of.Death, y=state.rate)) +
  geom_boxplot() +
  theme(axis.text = element_text(size = 5.5))  +
  xlab("Cause of Death") +
  ylab("State Death Rate")
```



```
ggplot(merged3, aes(x=Cause.of.Death, y=state.rate)) +
  geom_boxplot() +
  theme(axis.text = element_text(size = 5.5))  +
  xlab("Cause of Death") +
  ylab("State Death Rate")
```



```
bx1 <- ggplot(merged, aes(y=merged$TotalPop)) + geom_boxplot() +  ylab("Total Population")
bx2 <- ggplot(merged, aes(y=merged$Poverty)) + geom_boxplot() +  ylab("Poverty")
bx3 <- ggplot(merged, aes(y=merged$ChildPoverty)) + geom_boxplot() +  ylab("Child Poverty")
bx4 <- ggplot(merged, aes(y=merged$Unemployment)) + geom_boxplot() +  ylab("Unemployment")
bx5 <- ggplot(merged, aes(y=merged$ethnicity.pct)) + geom_boxplot() +  ylab("% Ethnicity")
bx6 <- ggplot(merged, aes(y=merged$employed.pct)) + geom_boxplot() +  ylab("% Employed")
bx7 <- ggplot(merged, aes(y=merged$gender.pct)) + geom_boxplot() +  ylab("% Gender")
bx8 <- ggplot(merged, aes(y=merged$gender.rate, x=merged$gender)) + geom_boxplot() +  ylab("G
ender Rate") + xlab("Gender")

ggarrange(bx1, bx2, bx3, bx4, bx5, bx6, bx7, bx8,
          ncol = 4, nrow = 2)
```



If it was beneficial to remove the outliers from **merged**, it is possible to follow the method shown below. However, for this case, removing outliers would produce an inaccurate dataset (for instance, removing outliers from **TotalPop** would remove records corresponding to the most populous states). Therefore, the R code below is provided as a demonstration.

```
# how outliers may be isolated using boxplot()
bx <- boxplot(merged$TotalPop, plot=FALSE)
merged %>% filter(., merged$TotalPop %in% bx$out)
```

State	gender	Year.and.Quarter	Cause.of.Death	▶
<chr>	<chr>	<ord>	<fct>	
California	Female	2019 Q4	Cancer	
California	Female	2019 Q4	Cancer	
California	Female	2019 Q4	Cancer	
California	Female	2019 Q4	Cancer	
California	Female	2019 Q4	Cancer	
California	Female	2019 Q4	Suicide	
California	Female	2019 Q4	Suicide	

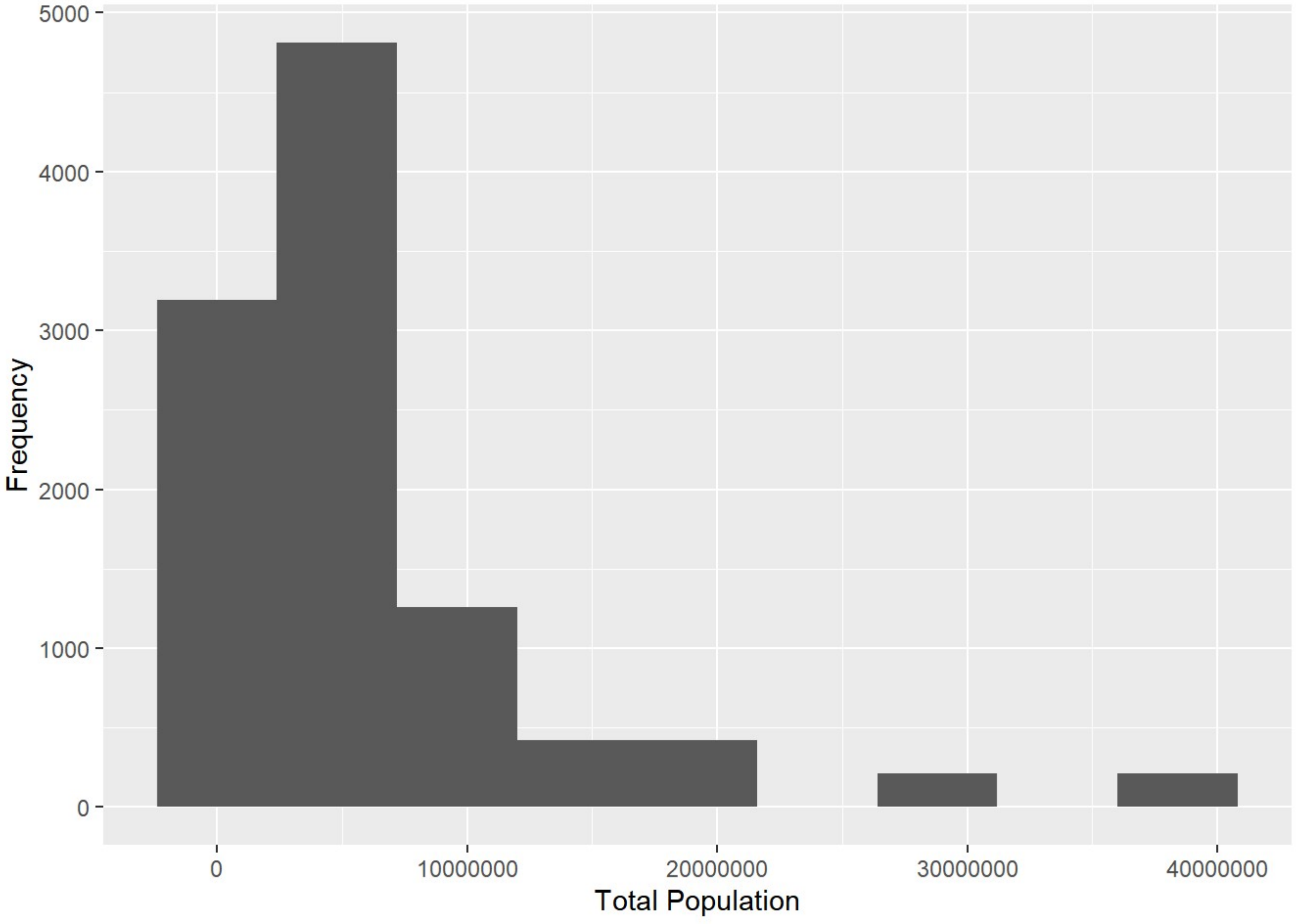
State <chr>	gender <chr>	Year.and.Quarter <ord>	Cause.of.Death <fct>
California	Female	2019 Q4	Suicide
California	Female	2019 Q4	Suicide
California	Female	2019 Q4	Suicide

1-10 of 840 rows | 1-4 of 18 columns Previous **1** 2 3 4 5 6 ... 84 Next

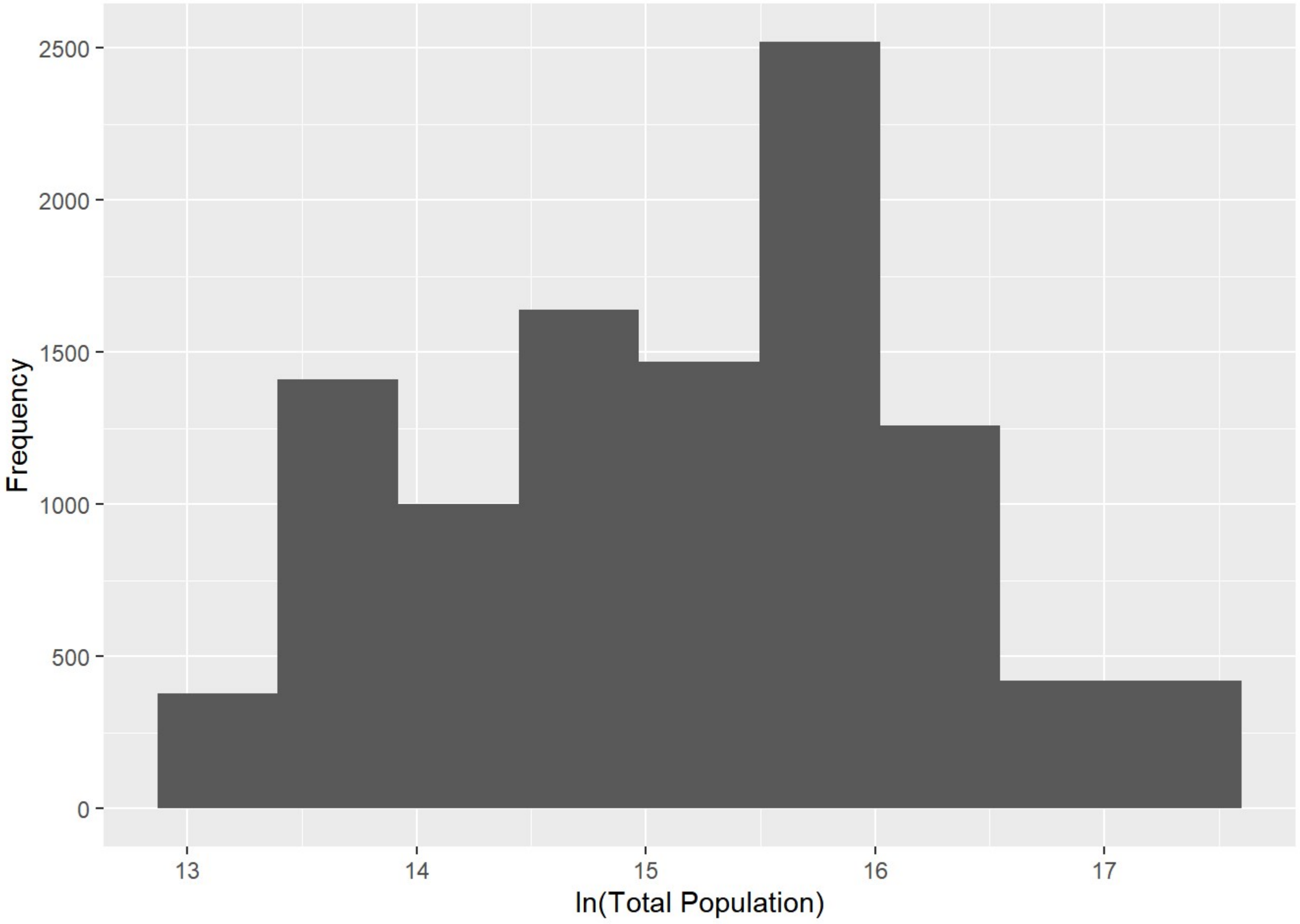
Transform

Plotting the histogram for **TotalPop** demonstrates that the variable is heavily right-skewed. To produce a more normal distribution, the *ln transform* is applied and the result is seen to have a distribution that is more normal than prior to transformation.

```
ggplot(merged, aes(x=TotalPop)) + geom_histogram(bins = 9) +  xlab("Total Population") +  yla
b("Frequency")
```



```
merged %<>% mutate(., ln_TotalPop = log(merged$TotalPop))
ggplot(merged, aes(x=ln_TotalPop)) + geom_histogram(bins = 9) +  xlab("ln(Total Population)")
+  ylab("Frequency")
```



References

1. Centers for Disease Control and Prevention (2016) NCHS - VSRR Quarterly provisional estimates for selected indicators of mortality, Data.gov website, accessed 23 May 2023. https://catalog.data.gov/dataset/nchs-vsrr-quarterly-provisional-estimates-for-selected-indicators-of-mortality (https://catalog.data.gov/dataset/nchs-vsrr-quarterly-provisional-estimates-for-selected-indicators-of-mortality)

2. MuonNeutrino (2019) US Census Demographic Data, Kaggle website, accessed 23 May 2023. https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv (https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv)