Investigating the incidence and potential risk factors for hypertension-related mortalities in adults in the United States.

Shaun van den Hurk

Warning: package 'magick' was built under R version 4.4.3

**Authors**

* First Autor (ORCID: 0000-0000-1234-5678)
* Second Author
* Third Autor
* Last Author

**Author affiliations**

1. College of Veterinary Medicine, University of Georgia, Athens, GA, USA.

These authors contributed equally to this work.

Corresponding author: some@email.com

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

Lifestyle diseases such as cardiovascular disease make up the highest causes of death in developed nations such as the United States. This project focuses on hypertension-related cardiovascular disease and hopes to help provide a greater understandign of the dynamics and patterns that may exist with hypertension and cardiovascualr diseased related mortality in the United States. The project seeks to evaluate data from the CDC on the rates of hypertension-related cardiovascular disease mortality and the associated demographic factors for this collected data, including county, age group, race/ethnicity, and sex. The data represents information collected from 2000 to 2019. This will be used to see if there are any potential patterns or correlation between the rate of mortality and these other demographic factors within the data.

# 2. Introduction

## 2.1 General Background Information

Lifestyle diseases such as cardiovascular disease make up the highest causes of death in developed nations such as the United States. This data analysis hopes to evaluate the rates of hypertension-related cardiovascular disease mortality in adults in the United States and trends that might be associated with demographic and geographic factors.

## 2.2 Description of data and data source

The dataset was collected from the CDC website and the information was sourced from the National Vital Statistics System. The dataset contains rates and trends in the hypertension-related cardiovascular disease (CVD) death rates. The study period is from 2000 to 2019, with two main intervals displayed in the data 2000-2009, and 2010-2019. The data is further distributed by county, age group (36-64 years and 65 years and older), race/ethnicity, and sex.

## 2.3 Questions/Hypotheses to be addressed

* Hypertension-related cardiovascular disease mortalities are higher in select demographics (age groups, race/ethnicities and sex), and thus these demographics show correlated with mortalities rates.
* There are clear regional/geographic distributions associated with hypertension-related cardiovascular disease mortalities.

# 3. Methods

## 3.1 Data import and cleaning

This data was downloaded from the CDC online database at the following url: https://data.cdc.gov/Heart-Disease-Stroke-Prevention/Rates-and-Trends-in-Hypertension-related-Cardiovas/uc9k-vc2j/about\_data. A csv file was downloaded which was almost 300MB in size. The initial document/dataset contained 1103872 observations across 24 variables. There were some initial processing steps performed to evaluate and delete some of the unnecessary columns and to save this data to a .rds document to reduce the size. This was all outlined in the “processingcode” document, which will not run automatically because the Excel (csv) file will not be available due to size concerns.

The intial exploration and cleaning and removal of the redundant columns (columns with repeated single values without any additional information). Some of the other columns were renamed to have the name of the column reflect what these single value columns had contained instead. The columns were checked to see what they contained and to see how many unique entries there were and based upon this columns were deleted and others renamed.

Deleted columns were: DataSource, Topic, Class, StratificationCategory1/2/3, Data\_Value\_Footnote, Data\_Value\_Footnote\_Symbol, Geographic\_Level, TopicID

Renamed columns: Stratification1 column to age\_group, Stratification2 column to race\_ethnicity, Stratification3 column to sex, Data\_Value to mortalities.

After the initial cleaning the data was evaluated again and some further small cleaning and processing steps were performed (code is visible in the processingfile-v1.qmd file) to reduce the number of rows by removing NA values. The resulting dataframe contains 572572 observations and 14 variables.

Our summaries and the skimr function help to provide a good overview for further evaluation of the data. We can see that our dataset contains a combination of numeric and character data. The numeric data includes the actual incidence rates of the hypertension-related cardiovascular disease death rates, as well as the confidence intervals that were used, and geographic coordinates. Our character data includes the race, sex, age group, location data and the year.

This character data will be used as predictors of the mortality data as we break down these different variables and look for potential patterns and associations. The primary outcome of interest is the mortality rates but the important potential predictors are the location, year, race/ethnicity, and sex. AGe groups only provide two broad categories, namely 35 to 65 years old and 65 years and older, which is a bit limited but will allow for comparisons between these two groups.

## 3.2 Statistical analysis

# 4. Results

## 4.1 Exploratory/Descriptive analysis

A summary table was produced using the skimr function which is printed below.

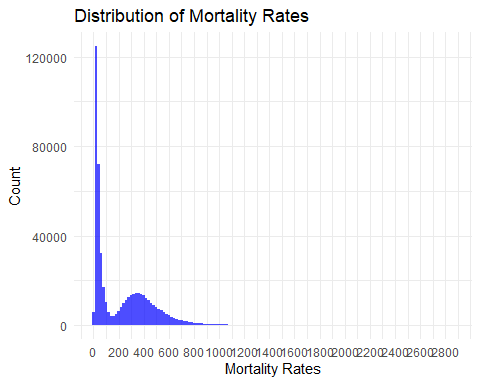
skim\_summary\_hypertens\_filtered <- readRDS(here::here("data","processed-data", "skim\_summary\_hypertens\_filtered.rds"))  
# Print the skim summary  
print(skim\_summary\_hypertens\_filtered)

skim\_type skim\_variable n\_missing complete\_rate character.min  
1 character LocationID 0 1.0000000 5  
2 character LocationAbbr 0 1.0000000 2  
3 character LocationDesc 0 1.0000000 3  
4 character Data\_Value\_Unit 0 1.0000000 1  
5 character Data\_Value\_Type 0 1.0000000 20  
6 character age\_group 0 1.0000000 14  
7 character race\_ethnicity 0 1.0000000 5  
8 character sex 0 1.0000000 4  
9 numeric Year 52052 0.9090909 NA  
10 numeric mortalities 0 1.0000000 NA  
11 numeric Confidence\_limit\_Low 0 1.0000000 NA  
12 numeric Confidence\_limit\_High 0 1.0000000 NA  
13 numeric X\_long 0 1.0000000 NA  
14 numeric Y\_lat 0 1.0000000 NA  
 character.max character.empty character.n\_unique character.whitespace  
1 5 0 3041 0  
2 2 0 51 0  
3 62 0 1776 0  
4 11 0 2 0  
5 48 0 2 0  
6 16 0 2 0  
7 33 0 6 0  
8 7 0 3 0  
9 NA NA NA NA  
10 NA NA NA NA  
11 NA NA NA NA  
12 NA NA NA NA  
13 NA NA NA NA  
14 NA NA NA NA  
 numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50 numeric.p75  
1 NA NA NA NA NA NA  
2 NA NA NA NA NA NA  
3 NA NA NA NA NA NA  
4 NA NA NA NA NA NA  
5 NA NA NA NA NA NA  
6 NA NA NA NA NA NA  
7 NA NA NA NA NA NA  
8 NA NA NA NA NA NA  
9 2009.5000 5.766287 2000.0000 2004.75000 2009.50000 2014.25000  
10 203.7201 223.834163 -94.8000 27.00000 75.80000 360.40000  
11 164.8418 189.516431 -95.7000 19.30000 56.60000 295.60000  
12 252.1425 266.580007 -93.7000 37.50000 105.70000 437.90000  
13 -90.9198 11.807409 -124.1563 -96.75245 -89.03454 -82.53677  
14 37.7873 4.813002 23.7515 34.13675 37.87732 41.26984  
 numeric.p100 numeric.hist  
1 NA <NA>  
2 NA <NA>  
3 NA <NA>  
4 NA <NA>  
5 NA <NA>  
6 NA <NA>  
7 NA <NA>  
8 NA <NA>  
9 2019.00000 ▇▇▇▇▇  
10 2857.90000 ▇▁▁▁▁  
11 2374.40000 ▇▁▁▁▁  
12 3414.00000 ▇▁▁▁▁  
13 -67.62915 ▂▂▆▇▂  
14 48.82564 ▁▅▇▇▃

Our summary data shows that the mean of the hypertension cardiovascular-diseases mortalities is around 203, with a standard deviation of 223.

A histogram was made of the mortalities columns since this is the main (only) column of interest with numerical data.

revised\_mortalities\_histogram <- readRDS(here::here("data","processed-data", "revised\_mortalities\_histogram.rds"))  
# Print the mortalities histogram  
plot(revised\_mortalities\_histogram)

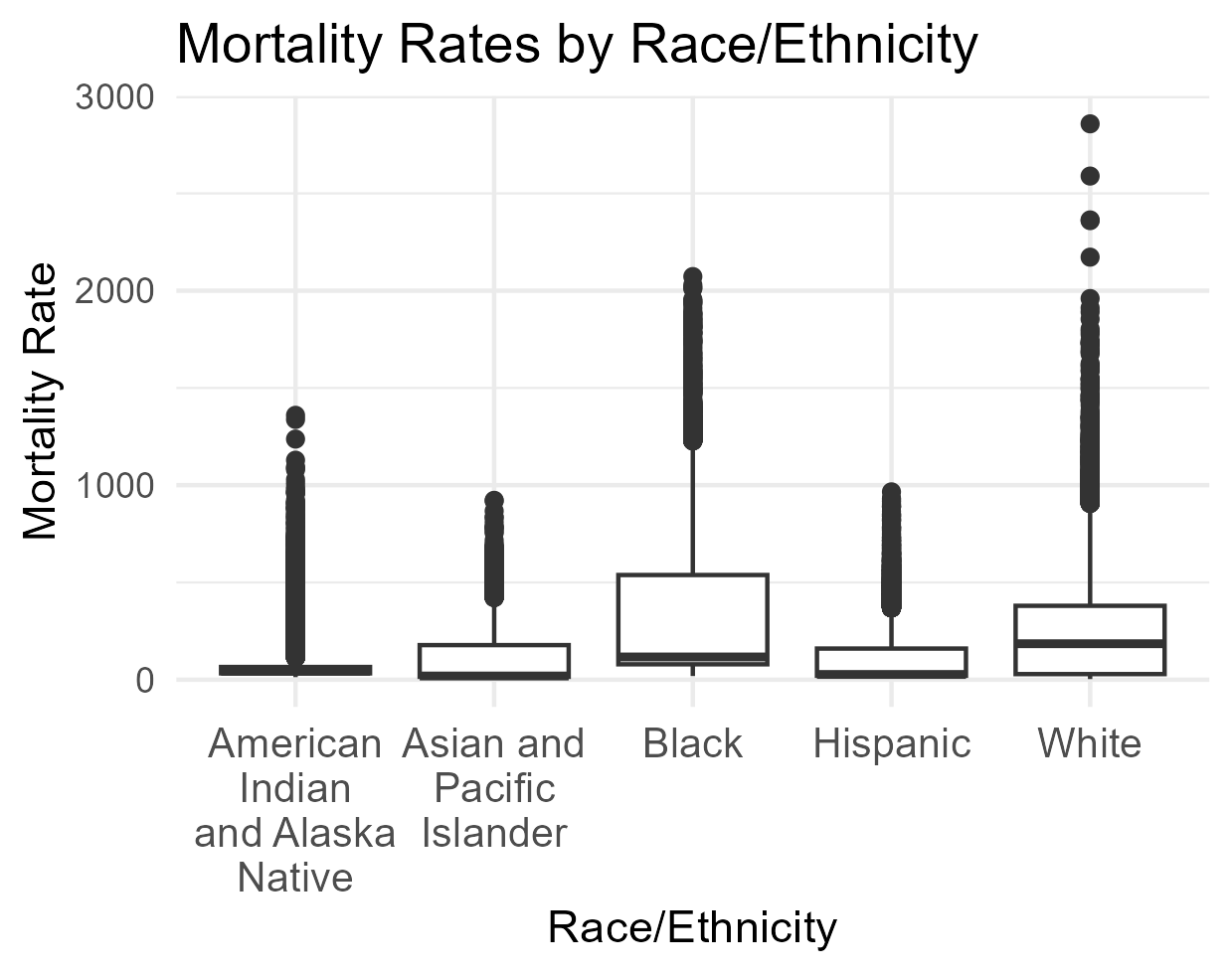


Our histogram depicts the hypertension-related cardiovasculat disease mortalities in the dataset, showing the number of mortalities and the frequency (count) of sprecific mortality values. The mortality counts are clustered towards the lower end of the scale, with a strong right-skew (long tail to the right/higher values). Peak mortalities are below 100, and the total values likely account for the second peak.

Boxplots were created to evaluate the mortalities by race and by sex.

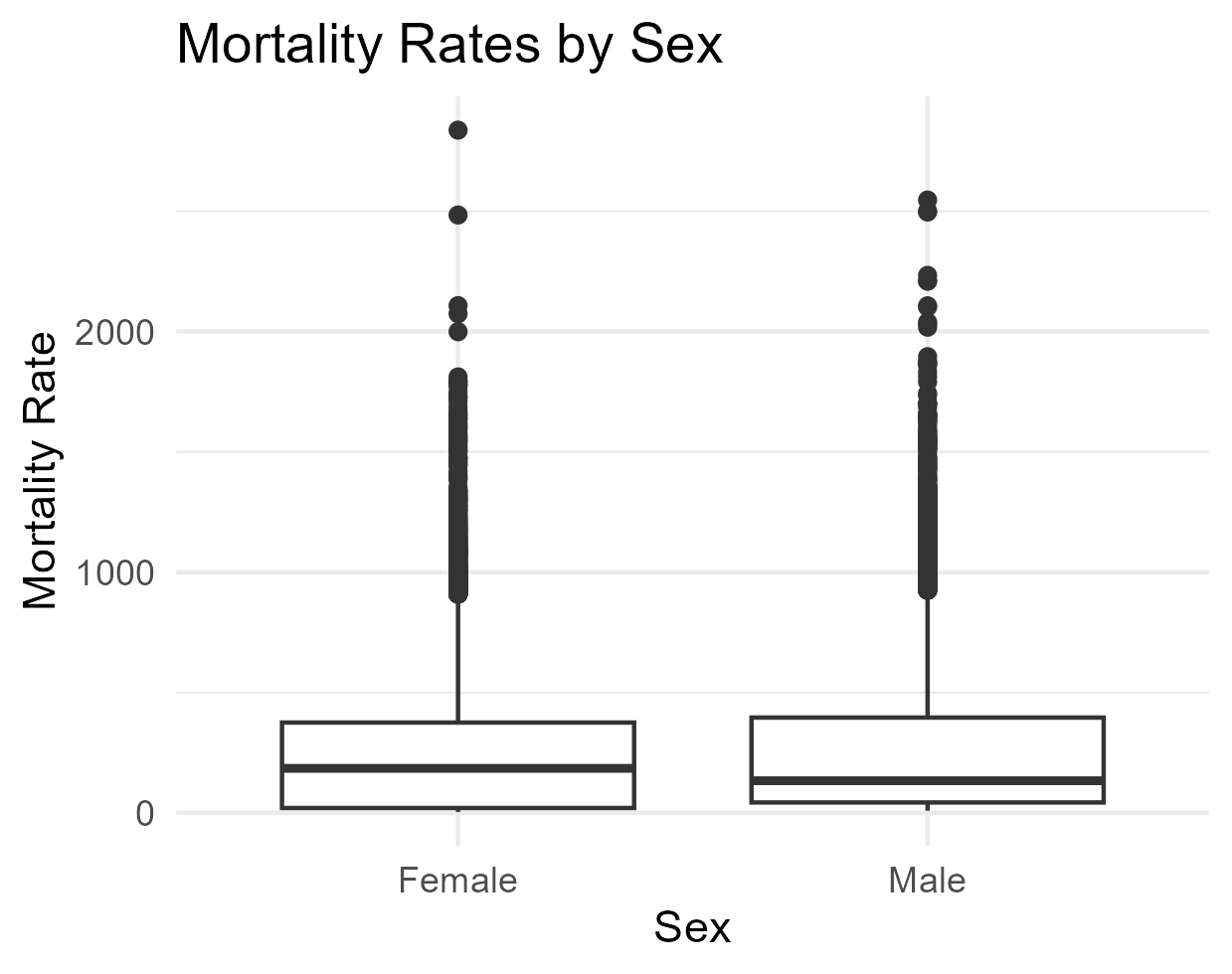
mortality\_race\_boxplot <- image\_read(here::here("products", "manuscript", "supplement", "mortality\_by\_race.png"))  
print(mortality\_race\_boxplot)

# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 1239 980 sRGB TRUE 63722 118x118



mortality\_sex\_boxplot <- image\_read(here::here("products", "manuscript", "supplement", "mortality\_by\_sex.png"))  
print(mortality\_sex\_boxplot)

# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 1239 980 sRGB TRUE 37253 118x118



## 4.2 Key Findings and Next Steps

* The distribution of mortality rates is right-skewed, suggesting potential transformations before modeling.
* Mortality rates vary markedly by race and sex, justifying stratified analysis.
* A proof-of-concept temporal trend was created for a single county (Calhoun, GA) using Location ID, ensuring unique selection.
* Temporal trend indicate a general increase the selected county, which should be modeled.
* All processed data has been saved into a single RDS file (final\_processed\_data.rds) for use in statistical analysis.

## 4.3 Basic statistical analysis

Spatial plots of mortalities per geographic regions were generated for each race and compared on one plot. Two generalized linear models were evaluated, the first examined how time, race/ethnicity, and sex (plot below) influence mortality rates, and the second tested whether the effect of time varied by race/ethnicity.

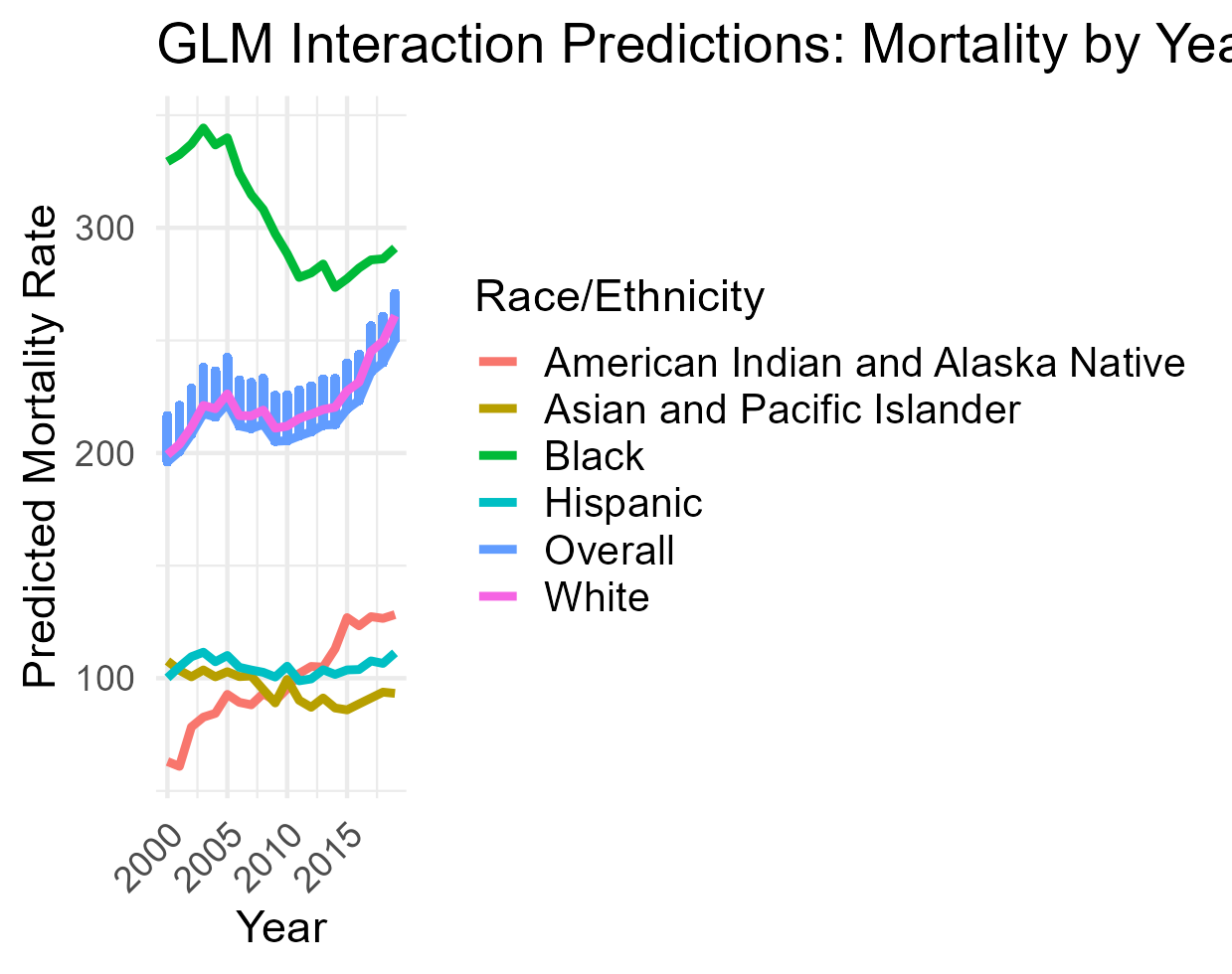
Overall, black individuals had the highest predcted mortality rates from the generalized linear models, followed by white and native american populations. Black individuals also had the highest hypertension-related mortalities on the boxplot follwoed by white individuals. Females appeared to have lower mortality rates than males across all races.

Race, sex and time all seem to be predictors of mortality. We see differences in trends over time for different racial groups.

Our spatial clustering has revealed geographic clustering of high-mortality regions, with high densities in the Eastern regions of the country. There also appears to be a higher density of mortalities from people in the race/ethnic group Black in certain regions. Asian and Pacific Islander groups had relatively lower mortality rates and geographic clustering.

#Print the glm plot  
glm\_interaction\_plot <- image\_read(here::here("products", "manuscript", "supplement", "glm\_interaction\_plot.png"))  
print(glm\_interaction\_plot)

# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 1239 980 sRGB TRUE 93902 118x118



#Print the combined spatial plot  
combined\_spatial\_plot <- image\_read(here::here("products", "manuscript", "supplement", "combined\_spatial\_plot.png"))  
print(combined\_spatial\_plot)

# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 4200 3000 sRGB FALSE 855811 118x118



## 4.4 Full analysis

\_

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References