Project 1

Tatum Bowen

Due: 03/10/2024

An Analysis of Mortality and its Causes for the Female and Male Populations in Multiple Regions Throughout the U.S.

Abstract

This a study about the data set USMortality looking into the relationship between male and female mortality rates for different causes of death along with the relationship between the different regions' mortality rates. The primary objective was to use different modes of data visualization in order to arrive at a conclusion regarding the analysis of the data set; this was accomplished as there were four different methods of plotting and data mapping used throughout this study. Furthermore, two conclusions were reached: 1) overall, men have a higher mortality rate than women regardless of the cause, 2) mortality rates remain consistent across every region of the US, with little to no fluctuation.

Introduction

The motivations behind the study and analysis of the data set USMortality were primarily focused on the number and relationship of the variables. This data set provided an excellent opportunity to look into various methods of displaying and comparing data, which was the main goal I wanted to accomplish. Furthermore, this data set contained information I found interesting and easy to engage with, making the analysis of the plots more exciting. The objectives of the study, beyond being able to experiment with new plotting methods, were to compare causes of mortality and mortality rates to a third variable. One element of the analysis focuses on the effects of being male or female; I used a point graph along with a linear regression model to visualize this comparison. The second part of my analysis works to compare the regions and this is done use a heatmap and a faceted group of plots. The information to learn how to do the plots came from RPubs.

Data

The data set I used in my study was USMortality. It has six variables (region, status, sex, cause, rate, and standard error) that are recorded for 400 observations. The variables that are the focus of this study are region, sex, cause, and rate. The cause refers to the cause of death, such as suicide, cancer, and other large contributors to mortality; the rate refers to the death rate per 100,000 people in a population; the region has ten subsections representing ten regions in the U.S.; and sex has two levels - male and female. There are some limitations to the data regarding the number of observations; since it is a small sample size, it is more difficult to observe relationships, it also introduces more uncertainty in the data because there may not be an accurate representation of the population in the sample size. The summary of the data set can be seen here: Rural:20 , Urban:20 , NA, NA, NA, NA, NA, Female:20 , Male :20 , NA, NA, NA, NA, NA, Alzheimers : 4 , Cancer : 4 , Cerebrovascular diseases: 4 , Diabetes : 4 , Flu and pneumonia : 4 , Heart

disease : 4 , (Other) :16 , Min. : 5.30 , 1st Qu.: 18.98 , Median : 30.05 , Mean : 58.73 , 3rd Qu.: 52.90 , Max. :242.70 , NA, Min. :0.00 , 1st Qu.:0.10 , Median :0.20 , Mean :0.18 , 3rd Qu.:0.20 , Max. :0.60 , NA.

Analysis and Results (Male vs. Female)

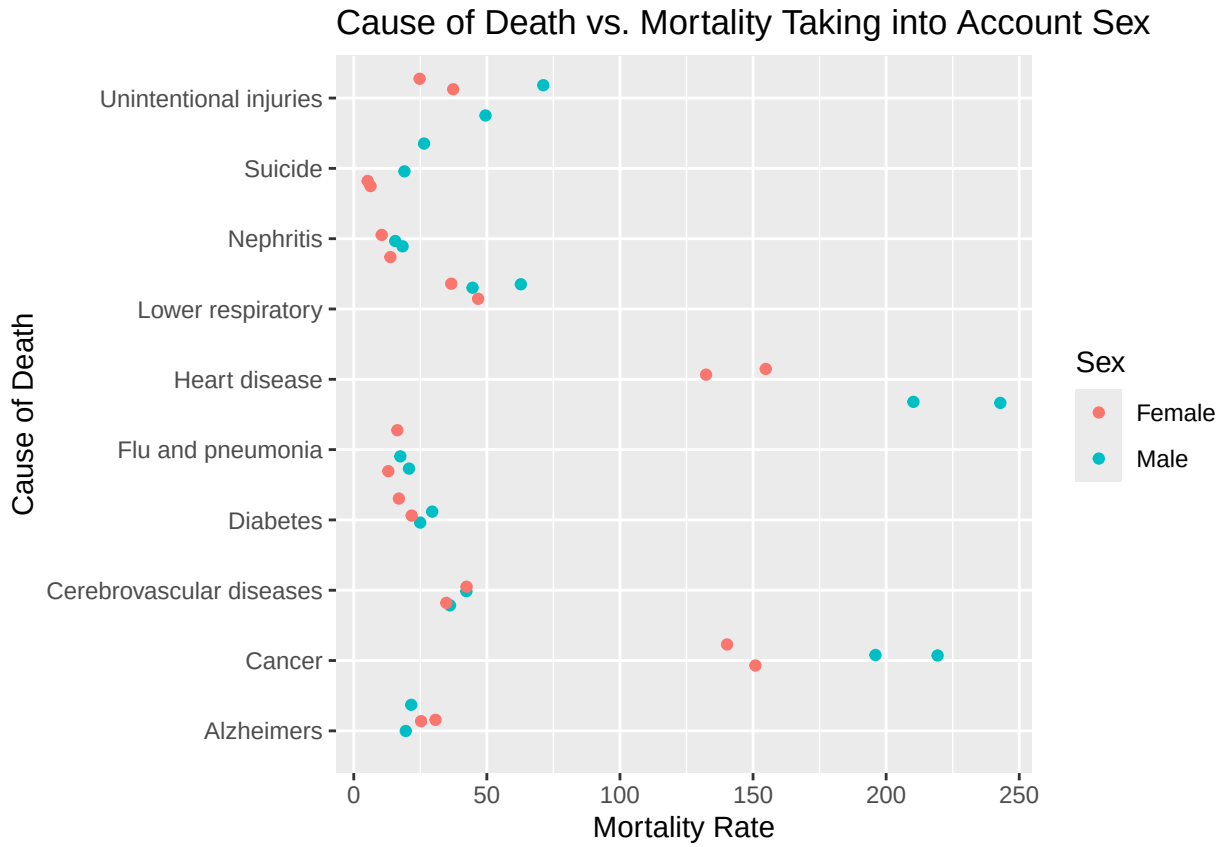


Figure 1: A comaprison of the rate or mortality for particular causes between males and females in the US.

```
## [[1]]

## `geom_smooth()` using formula = 'y ~ x'


##
## [[2]]

## `geom_smooth()` using formula = 'y ~ x'


##
## [[3]]

## `geom_smooth()` using formula = 'y ~ x'
```

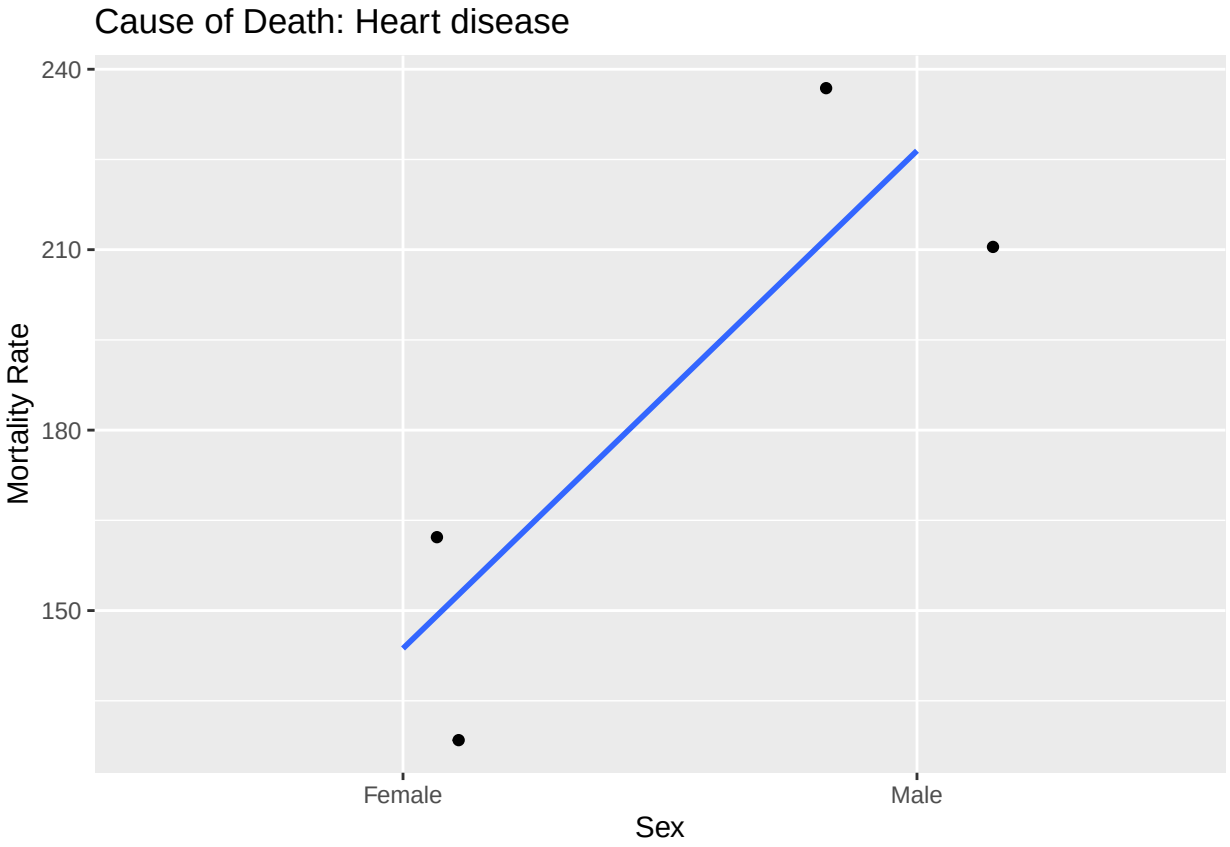


Figure 2: A regression model to better compare the difference in mortality for different causes between males and females.

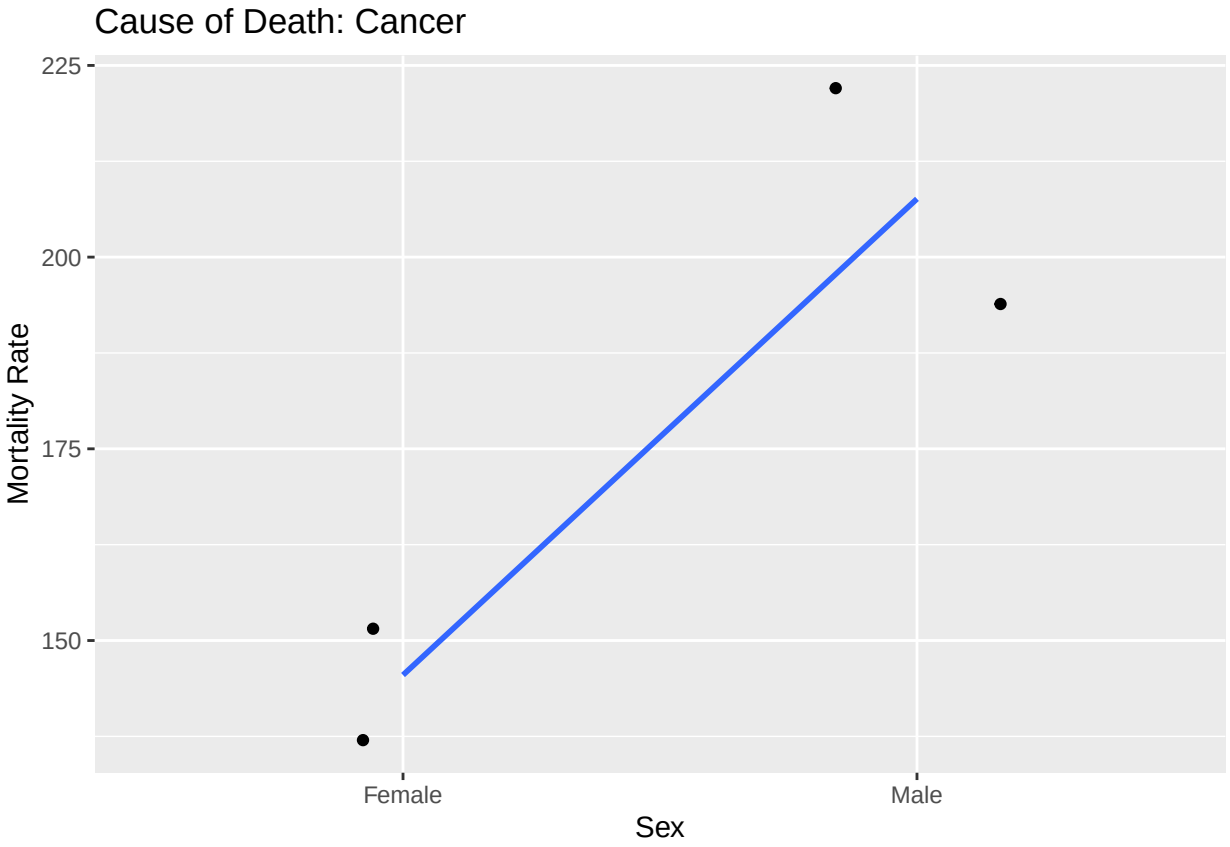


Figure 3: A regression model to better compare the difference in mortality for different causes between males and females.

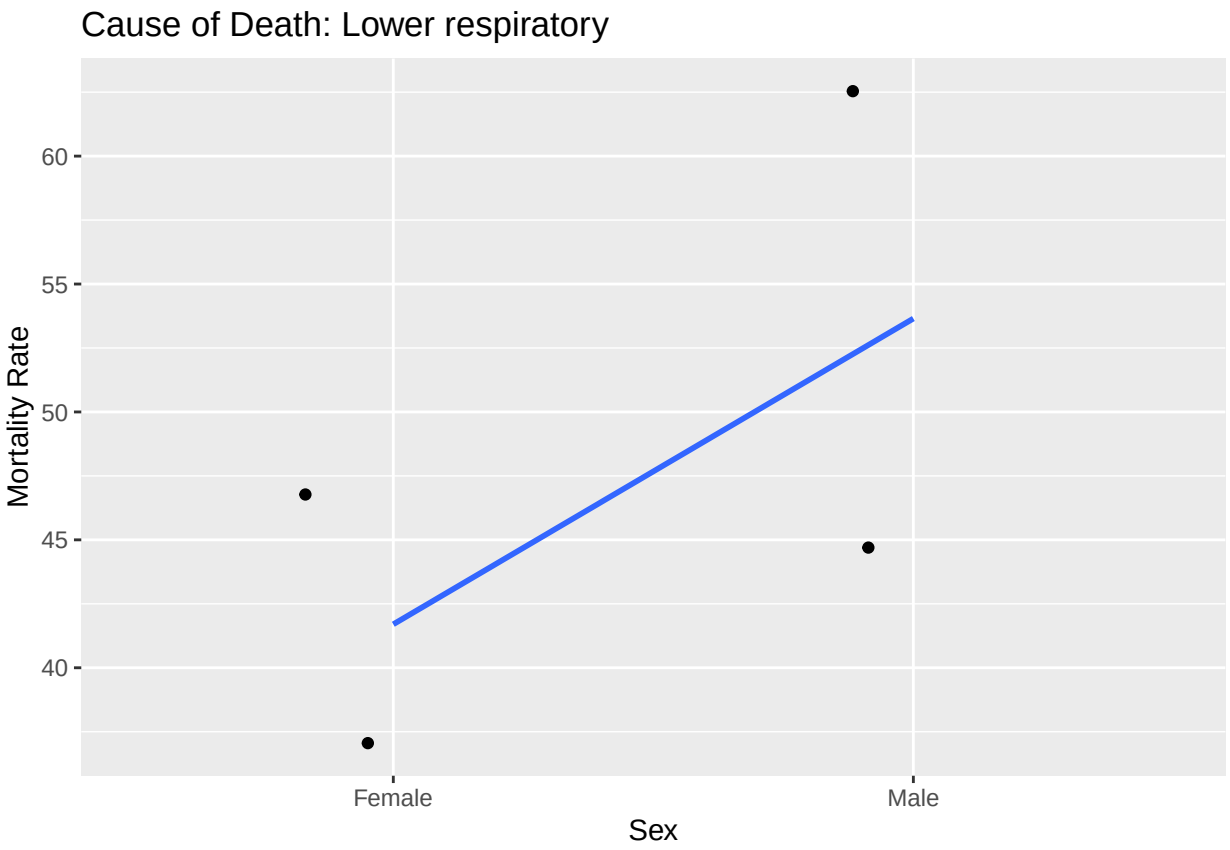


Figure 4: A regression model to better compare the difference in mortality for different causes between males and females.

```
##
## [[4]]

## `geom_smooth()` using formula = 'y ~ x'
```

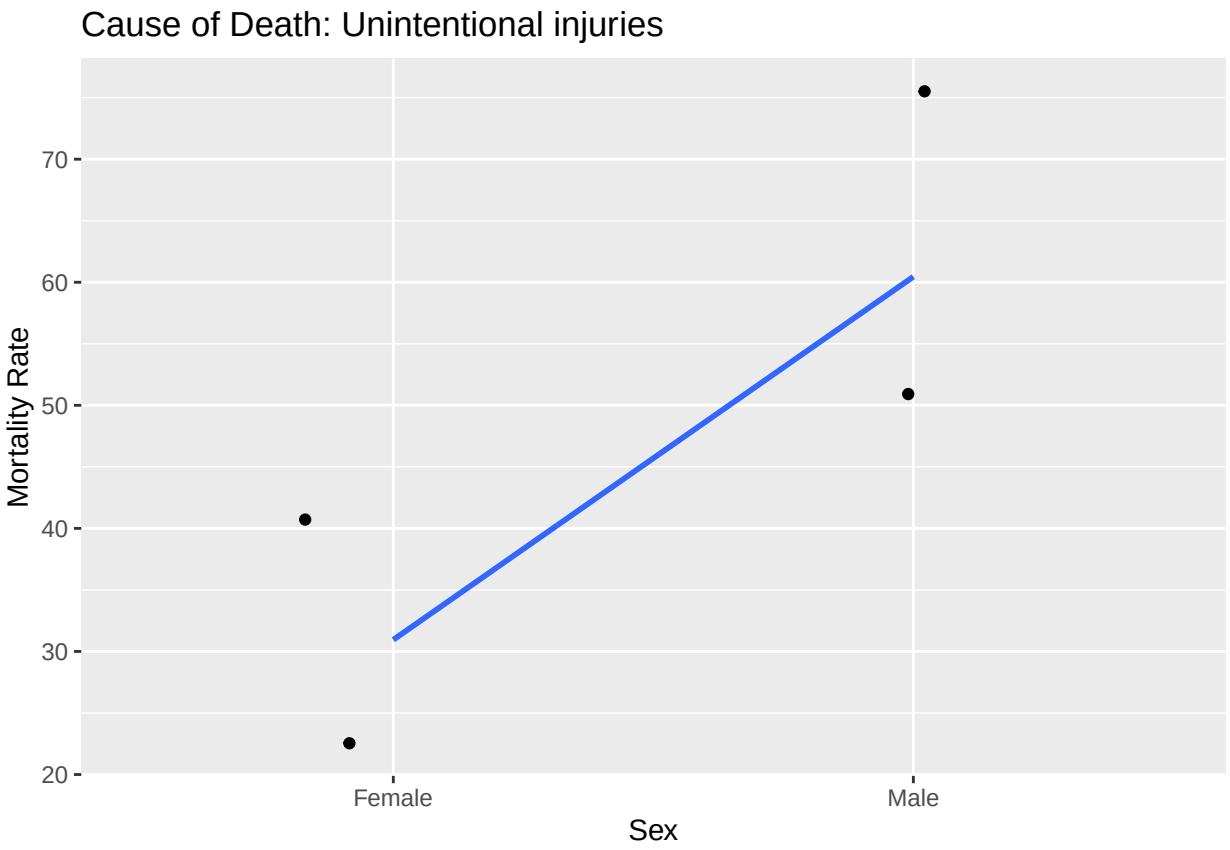


Figure 5: A regression model to better compare the difference in mortality for different causes between males and females.

```
##
## [[5]]

## `geom_smooth()` using formula = 'y ~ x'


##
## [[6]]

## `geom_smooth()` using formula = 'y ~ x'


##
## [[7]]

## `geom_smooth()` using formula = 'y ~ x'
```

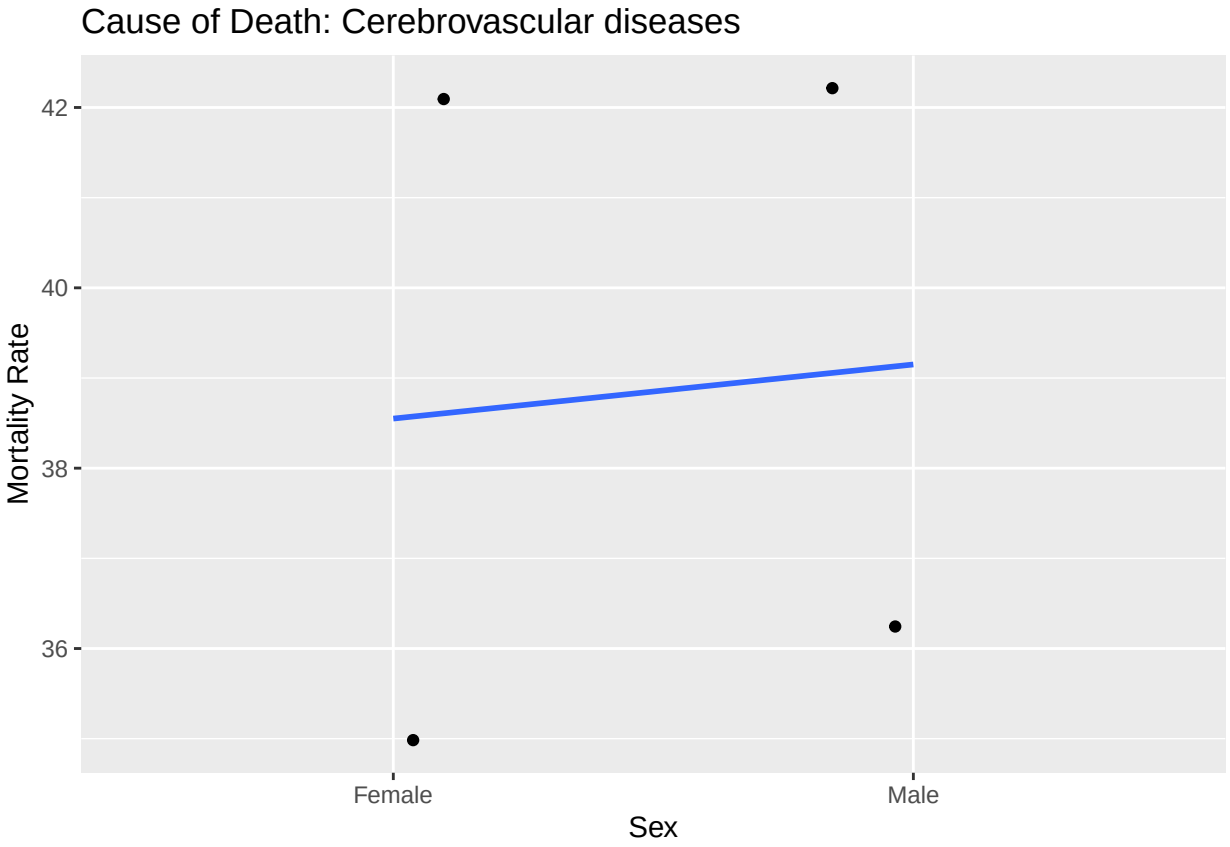


Figure 6: A regression model to better compare the difference in mortality for different causes between males and females.

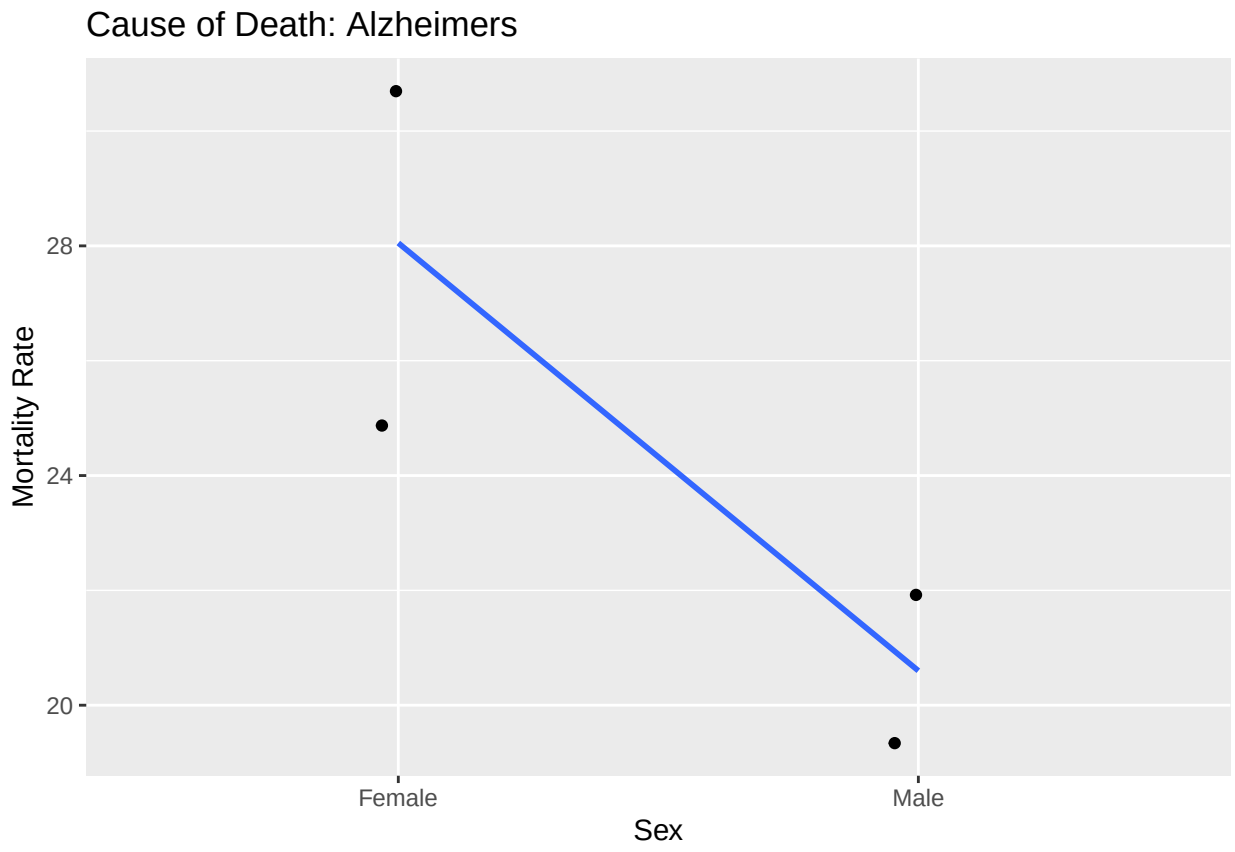


Figure 7: A regression model to better compare the difference in mortality for different causes between males and females.

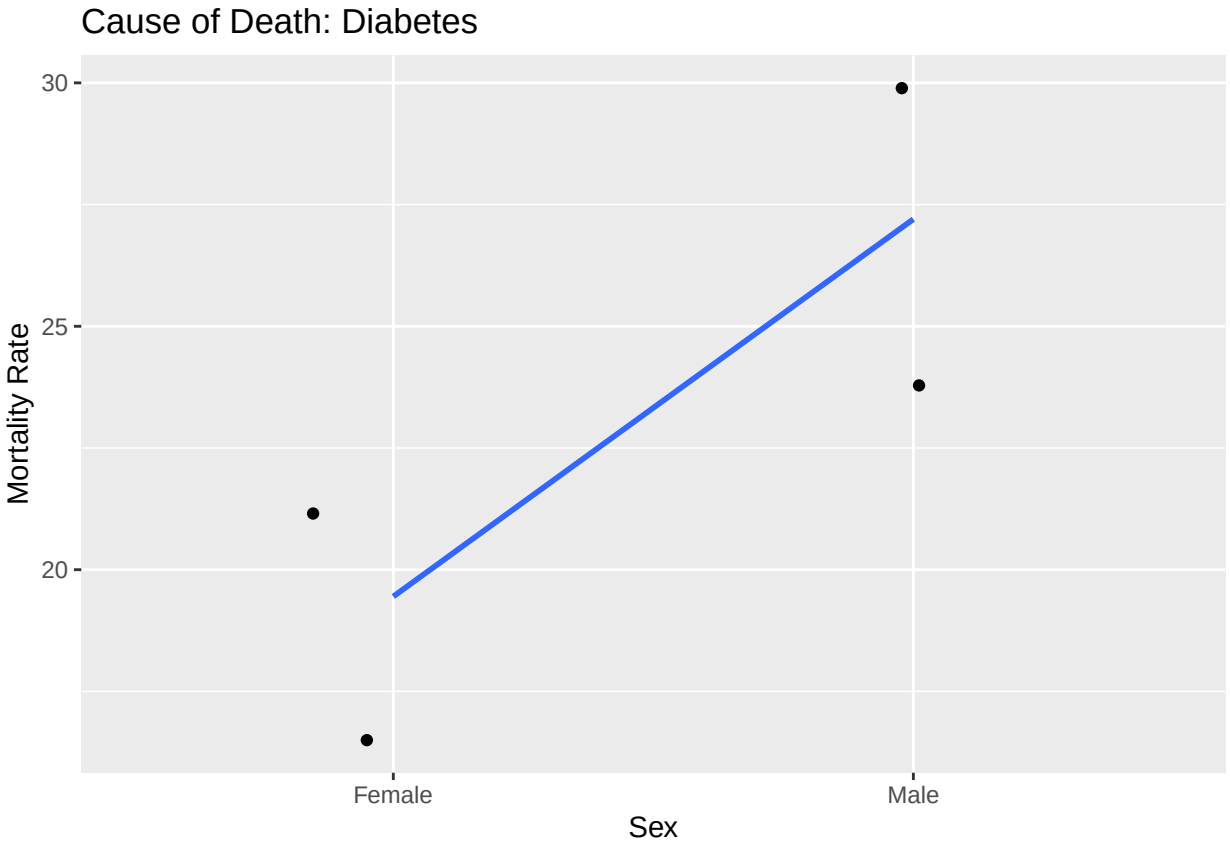


Figure 8: A regression model to better compare the difference in mortality for different causes between males and females.

```
##
## [[8]]

## `geom_smooth()` using formula = 'y ~ x'
```

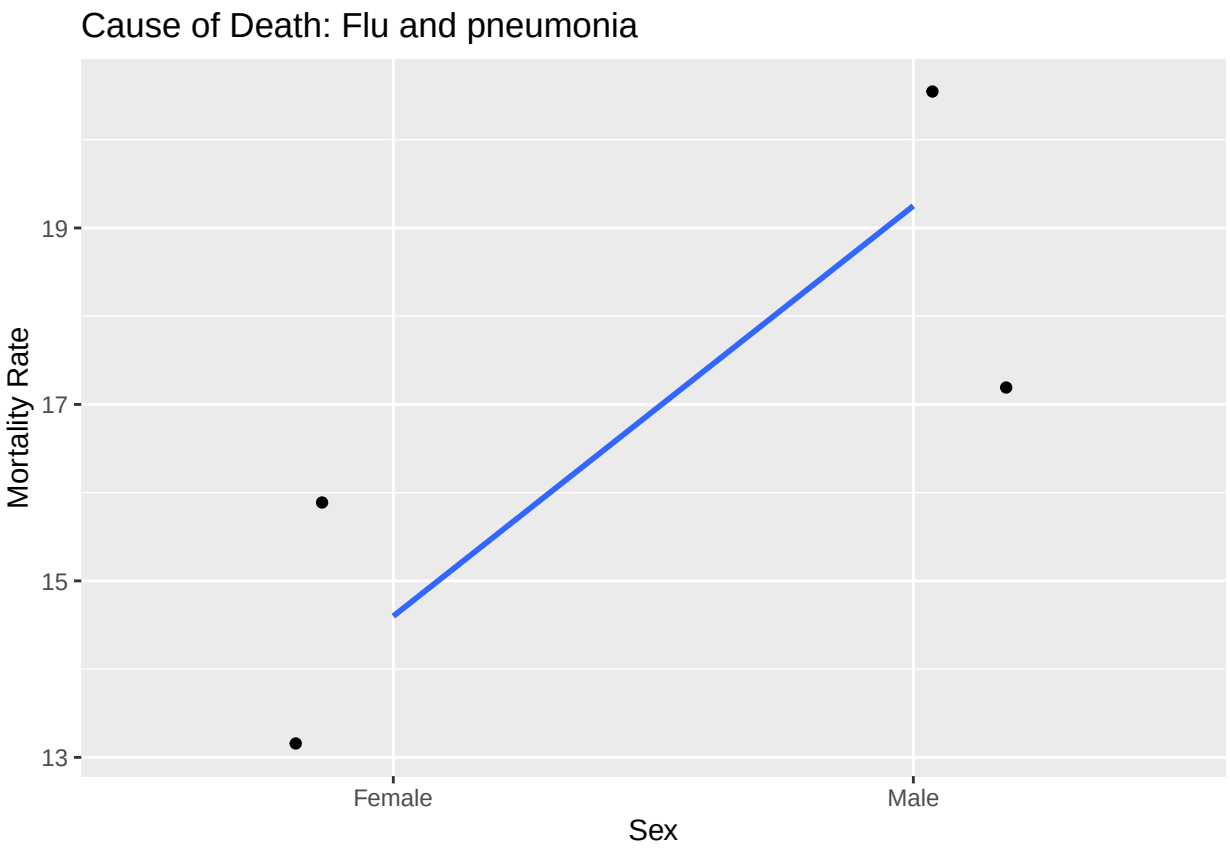


Figure 9: A regression model to better compare the difference in mortality for different causes between males and females.

```
##
## [[9]]

## `geom_smooth()` using formula = 'y ~ x'


##
## [[10]]

## `geom_smooth()` using formula = 'y ~ x'
```

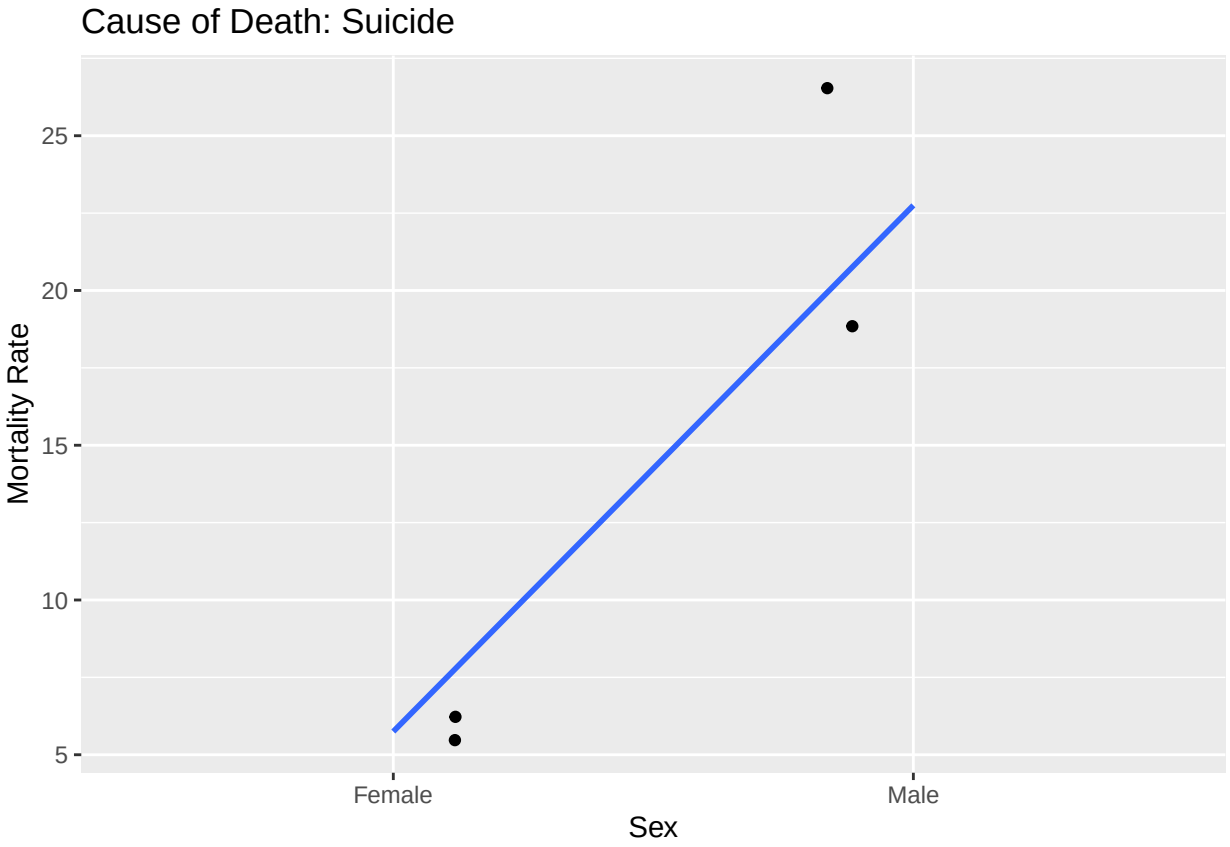
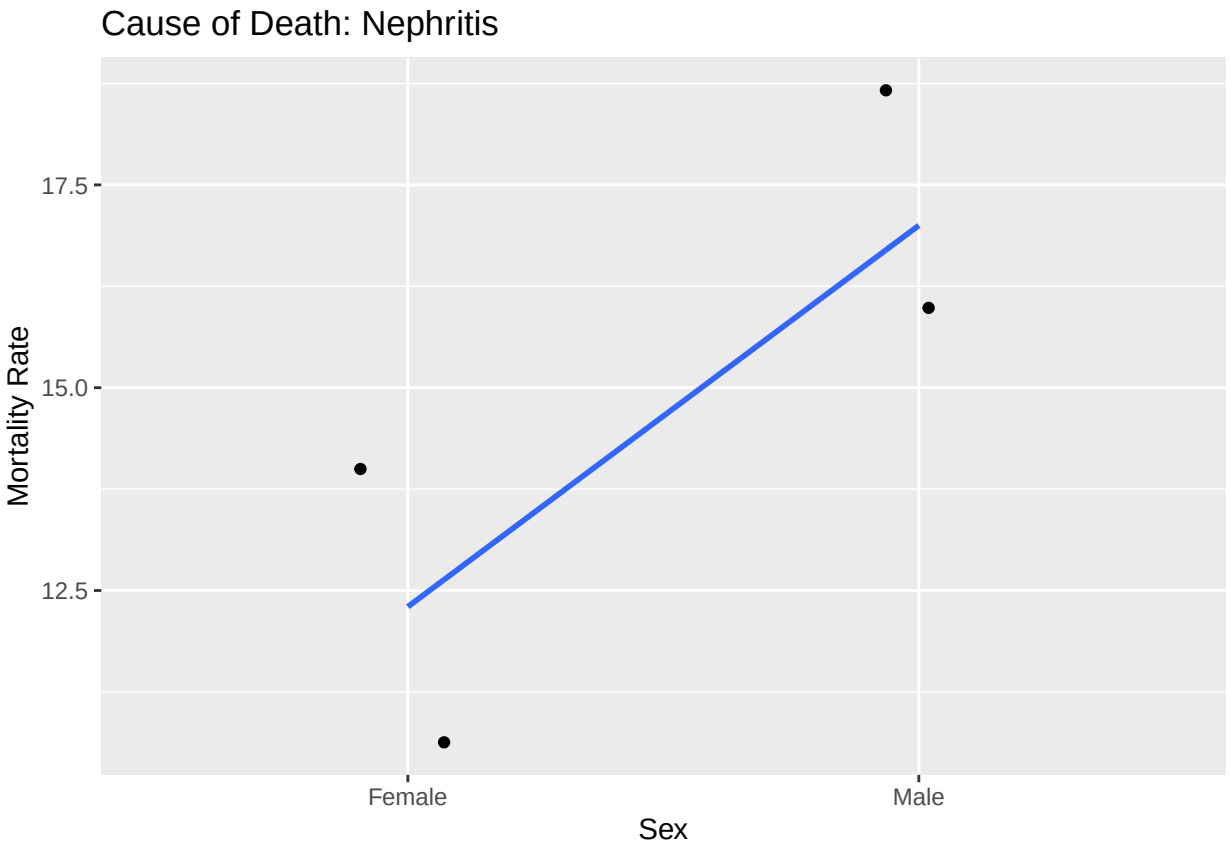


Figure 10: A regression model to better compare the difference in mortality for different causes between males and females.



The first element of analysis is comparing the difference in mortality rates for males and females with regards to different causes. This is displayed in a couple different ways, each will be discussed separately in order to land on a conclusion for the relationships between the variables. In Figure @ref{fig:male_female_comparison}, there is a display of three variables in one plot. Causes is on the y-axis and Mortality Rate is on the x-axis, while the male or female element is displayed by different colored points. In this visualization, there are no extremely obvious differences between the mortality rates of men and women, but it can be observed that the men do have remarkably higher mortality rates when it comes to heart disease and cancer. It seems the most obvious data to be observed from Figure @ref{fig:male_female_comparison} is that cancer and heart disease have a larger mortality rate than many other causes, but this was not the objective of this analysis. To further try and reach the objective of this analysis, there is a second visualization used; this is a group of plots that demonstrate a linear regression. A linear regression with this data may seem a bit trivial, but it does a better job of showing the differences in mortality rate between men and women than Figure @ref{fig:male_female_comparison} did. The relationship for the linear regression can be observed: c(`(Intercept)` = 48.055, SexMale = 21.355), c(`1` = 140.79, `2` = 173.29, `3` = 84.445, `4` = 106.845, `53` = 126.49, `54` = 149.89, `55` = 92.145, `56` = 102.745, `105` = -24.91, `106` = -6.61000000000001, `107` = -11.555, `108` = -1.15499999999999, `157` = -19.81, `158` = 1.88999999999999, `159` = -23.355, `160` = -10.855, `209` = -33.31, `210` = -27.21, `211` = -13.155, `212` = -5.85499999999999, `261` = -50.01, `262` = -47.61, `263` = -22.555, `264` = -17.455, `313` = -44.51, `314` = -39.91, `315` = -30.955, `316` = -26.255, `365` = -51.71, `366` = -48.61, `367` = -35.155, `368` = -31.755, `417` = -50.21, `418` = -43.11, `419` = -42.755, `420` = -41.855, `469` = -53.71, `470` = -51.11, `471` = -37.355, `472` = -34.155), c(`(Intercept)` = -371.456945351679, SexMale = -67.5304394328957, 89.5837011977736, 111.983701197774, 88.7606573607953, 112.160657360795, 97.2837011977736, 107.883701197774, -62.6393426392047, -44.3393426392047, -6.41629880222644, 3.98370119777356, -57.5393426392047, -35.8393426392047, -18.2162988022264, -5.71629880222644, -71.0393426392047, -64.9393426392047, -8.01629880222644, -0.716298802226437, -87.7393426392047, -85.3393426392047, -17.4162988022264, -12.3162988022264, -82.2393426392047, -77.6393426392047, -25.8162988022264, -21.1162988022264, -89.4393426392047, -86.3393426392047, -30.0162988022264, -26.6162988022264, -87.9393426392047, -80.8393426392047, -37.6162988022264, -

36.7162988022264, -91.4393426392047, -88.8393426392047, -32.2162988022264, -29.0162988022264), 2, c(1 = 69.4099999999997, 2 = 69.4100000000001, 3 = 48.055, 4 = 48.055, 53 = 69.41, 54 = 69.41, 55 = 48.055, 56 = 48.055, 105 = 69.41, 106 = 69.41, 107 = 48.055, 108 = 48.055, 157 = 69.41, 158 = 69.41, 159 = 48.055, 160 = 48.055, 209 = 69.41, 210 = 69.41, 211 = 48.055, 212 = 48.055, 261 = 69.41, 262 = 69.41, 263 = 48.055, 264 = 48.055, 313 = 69.41, 314 = 69.41, 315 = 48.055, 316 = 48.055, 365 = 69.41, 366 = 69.41, 367 = 48.055, 368 = 48.055, 417 = 69.41, 418 = 69.41, 419 = 48.055, 420 = 48.055, 469 = 69.41, 470 = 69.41, 471 = 48.055, 472 = 48.055), 0:1, list(qr = c(-6.32455532033676, 0.158113883008419, -3.16227766016838, -3.16227766016838, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024, 0.136527059495814, 0.136527059495814, -0.179700706521024, -0.179700706521024), qraux = c(1.15811388300842, 1.13652705949581), pivot = 1:2, tol = 1e-07, rank = 2), 38, list(Sex = "contr.treatment"), list(Sex = c("Female", "Male")), lm(formula = Rate ~ Sex, data = cause_data), Rate ~ Sex, list(Rate = c(210.2, 242.7, 132.5, 154.9, 195.9, 219.3, 140.2, 150.8, 44.5, 62.8, 36.5, 46.9, 49.6, 71.3, 24.7, 37.2, 36.1, 42.2, 34.9, 42.2, 19.4, 21.8, 25.5, 30.6, 24.9, 29.5, 17.1, 21.8, 17.7, 20.8, 12.9, 16.3, 19.2, 26.3, 5.3, 6.2, 15.7, 18.3, 10.7, 13.9), Sex = c(2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1)). In Figure @ref{fig:regression_models}, there is a plot for each cause that compares the mortality rate of that cause between men and women. The linear regression line can be observed in each plot as well; if I wanted to further analyze the relationship between mortality and sex, I could calculate the slope of the linear regression line for each cause. But for this analysis, the visual of the line served the purpose of demonstrating that men, overall, have a higher mortality rate than women regardless of the cause. It is obvious that the cause must come into play in some ways, as the difference between men and women is not consistent for every cause. It is also not always true, in the case of Alzheimers, women have a higher mortality rate than men. Overall, Figure @ref{fig:regression_models} allows for a rather definitive result, given the data and its potential limitations, that men have a higher mortality rate than women regardless of cause of death.

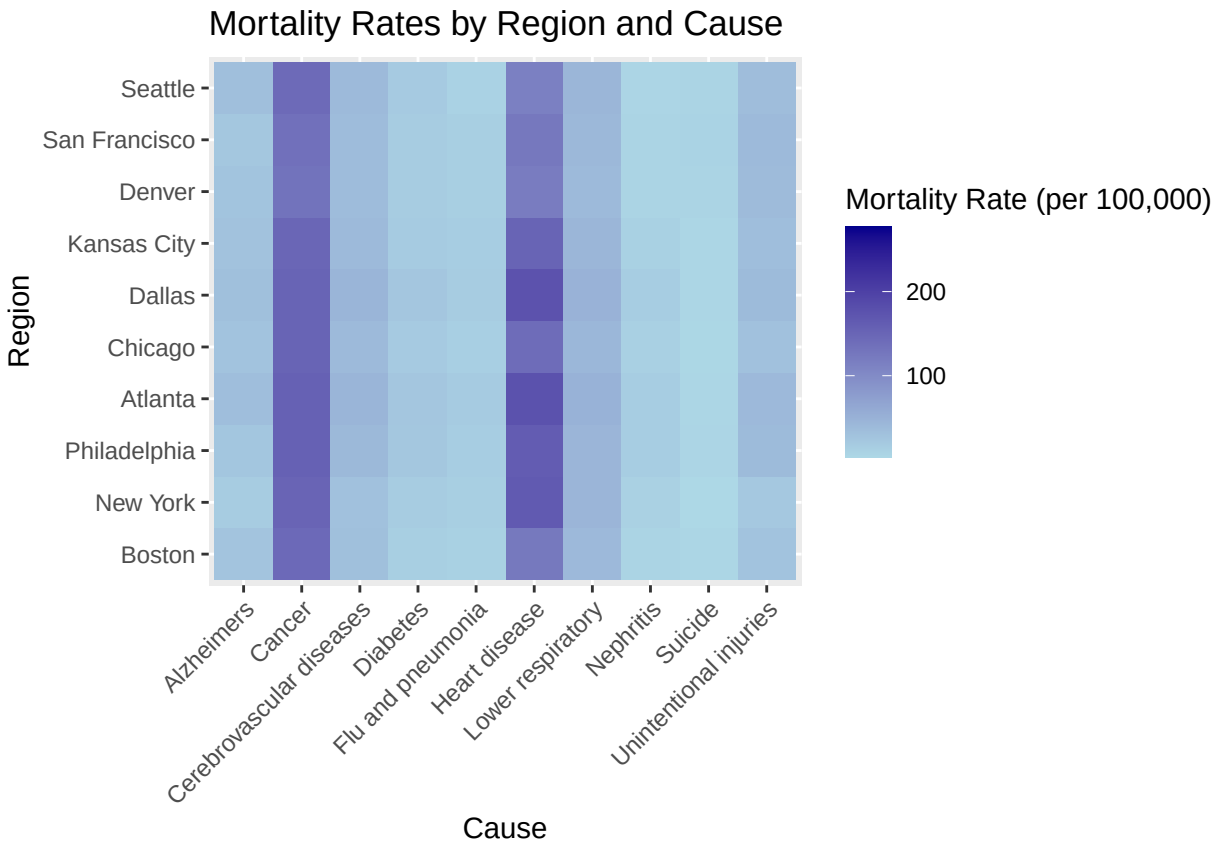
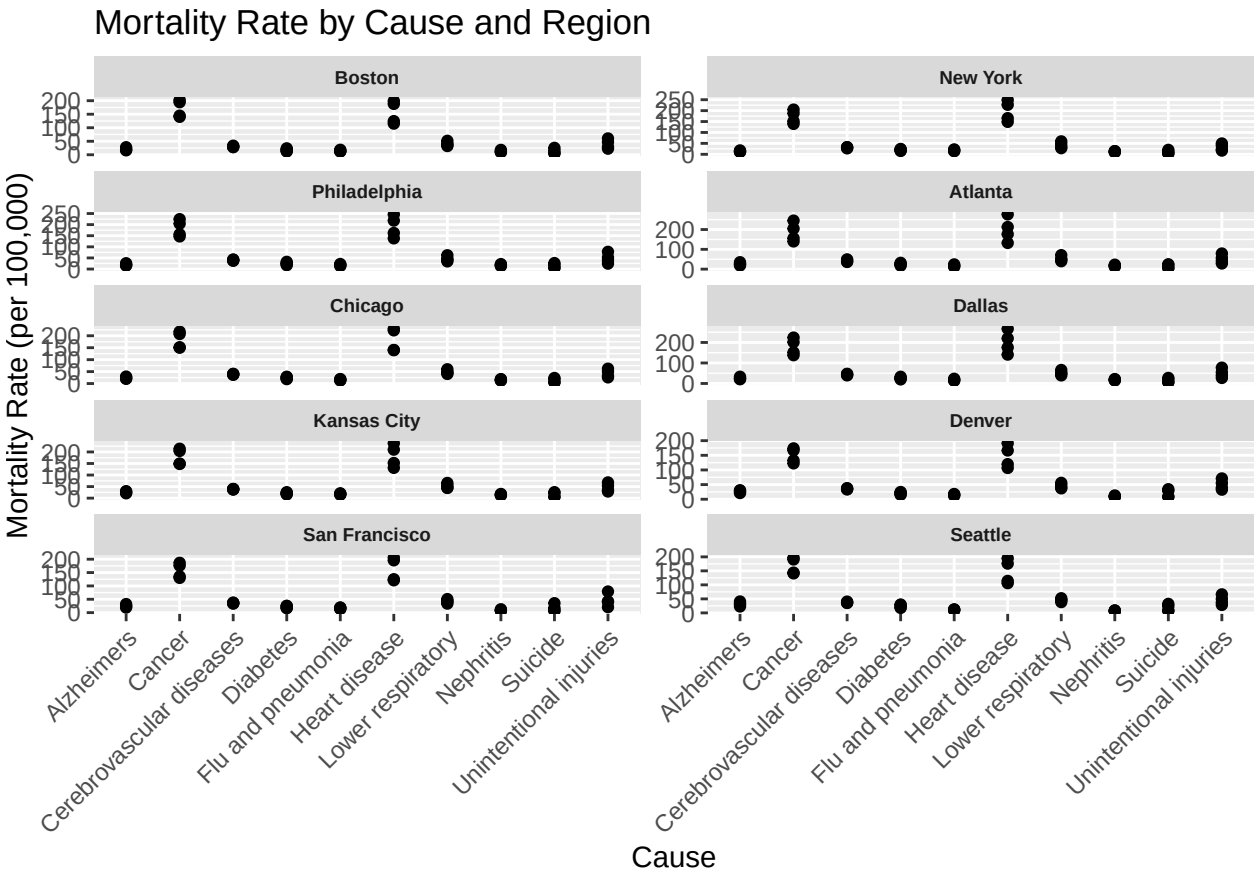


Figure 11: A heatmap examining the mortality rates of particular causes from a regional standpoint.

Analysis and Results (Regionally)



The second part of the analysis was to look at how mortality rates per cause changes regionally; this, similarly to the first part of analysis, is shown with two different methods. Firstly, Figure @ref{fig:regional_heatmap} uses a heat map to show the mortality rate for each cause regionally. This method of displaying data is not especially helpful in providing insight to the relationship between mortality rates and regions, but like Figure @ref{fig:male_female_comparison}, it highlights how high the mortality rates are for cancer and heart disease. The second visualization, Figure @ref{fig:faceted_regions}, more clearly demonstrates the mortality rate for each cause per region. Although both plots demonstrate the same relationship, the faceted plots show the relationship much better. Unfortunately, neither plot allows for a definite relationship between the mortality rates of causes in each of the different regions. Each region has a very similar mortality rate for every single one of the causes. One conclusion that can be made from this observation is that region has no role in the mortality rate; if a large enough sample is collected from any region, then the mortality rate should remain consistent, or nearly so, across every region.

Conclusion

My primary objective in this study was to be able to learn different ways to present data visually, which was very much accomplished; I was able to learn about heatmaps and faceted plots and how to create multiple plots based on a linear regression made with the variables from the dataset. The objectives for analyzing the dataset were also completed. Using Figure @ref{fig:male_female_comparison} and Figure @ref{fig:regression_models}, it can be determined that men having a higher mortality rate than women, overall, but not in every case. This could also be further investigated by looking into the ages of mortality for the males and females and the population of men and women at different ages. For example, if men do not live as long as women, that might be the reason the mortality rate for Alzheimer is higher for women. Using Figure @ref{fig:regional_heatmap} and Figure @ref{fig:faceted_regions}, it can be concluded that mortality

rate for the different causes is consistent across the different regions. This could be further analyzed by splitting the regions by status, a varible that stated whether the person was rural or urban, and then looking at the mortality rates for the rural and urban populations in each region.

References

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2022. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. https://www.rstudio.com/.

Sarkar, Deepayan, and Felix Andrews. 2022. *Lattice: Trellis Graphics for r*. https://cran.r-project.org/package=lattice.

University of Wisconsin-Madison. 2023. *Data Visulization in r with Ggplot2*. Social Science Computing Cooperative. https://sscc.wisc.edu/sscc/pubs/dvr/index.html.

Xu Liu. n.d. *Advanced Graphics*. RPubs by RStudio. https://rpubs.com/xliusufe/ch7.

Yihui Xie and Christophe Dervieux and Emily Riederer. 2024. *R Markdown Cookbook*. CRC Press. https://bookdown.org/yihui/rmarkdown-cookbook/bibliography.html.