

# ¶ Stat 405 Final

Meera Borle, Isis Burgos, Naomi Consiglio, Carson Foster, Aidan Gerber

December 10th, 2022

## Introduction

Our main dataset is a mortality dataset from the National Bureau of Economic Research. There are new versions of it each year and we are using the 2019 version. Each row is one death based on information filed in death certificates in the 50 states plus Washington DC. Columns are factors regarding the status of persons involved in deaths that can be gleaned from public death certificates. The dataset is focused on the different causes of death. Notable information the dataset contains is education, sex, age classification, day of month, place of death, weekday, manner of death, cause of death, and different risk factors that the deceased had.

## Secondary Dataset

For our secondary dataset, we are using the Behavioral Risk Factor Surveillance System Survey. This survey includes different free text survey questions from across the United States and territories with responses broken out by subgroup. There is also information on sample size, percent affirmative response, and confidence interval bounds. We will combine the secondary dataset by matching up subgroups between the death dataset and the risk factor dataset and trying to use aggregate statistics to analyze how risk factors can be matched with causes of death.

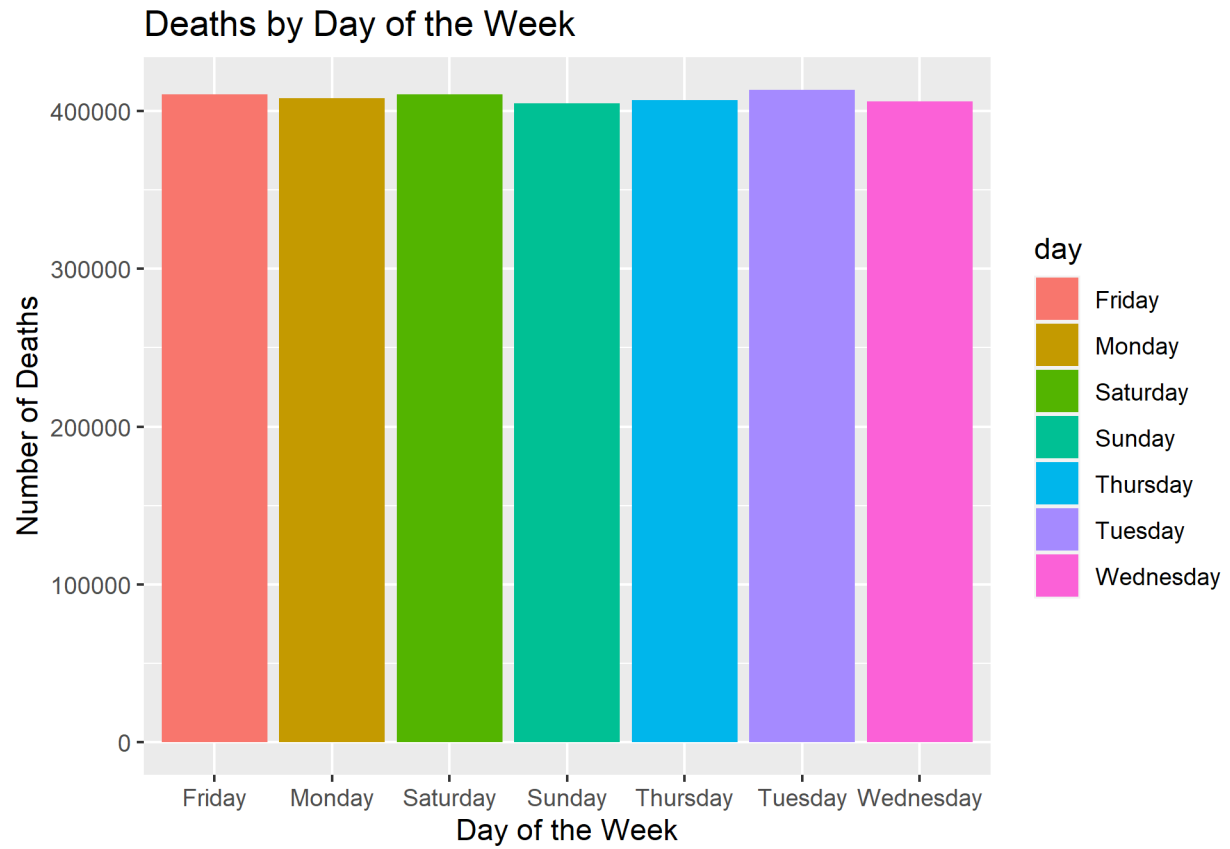
## Questions

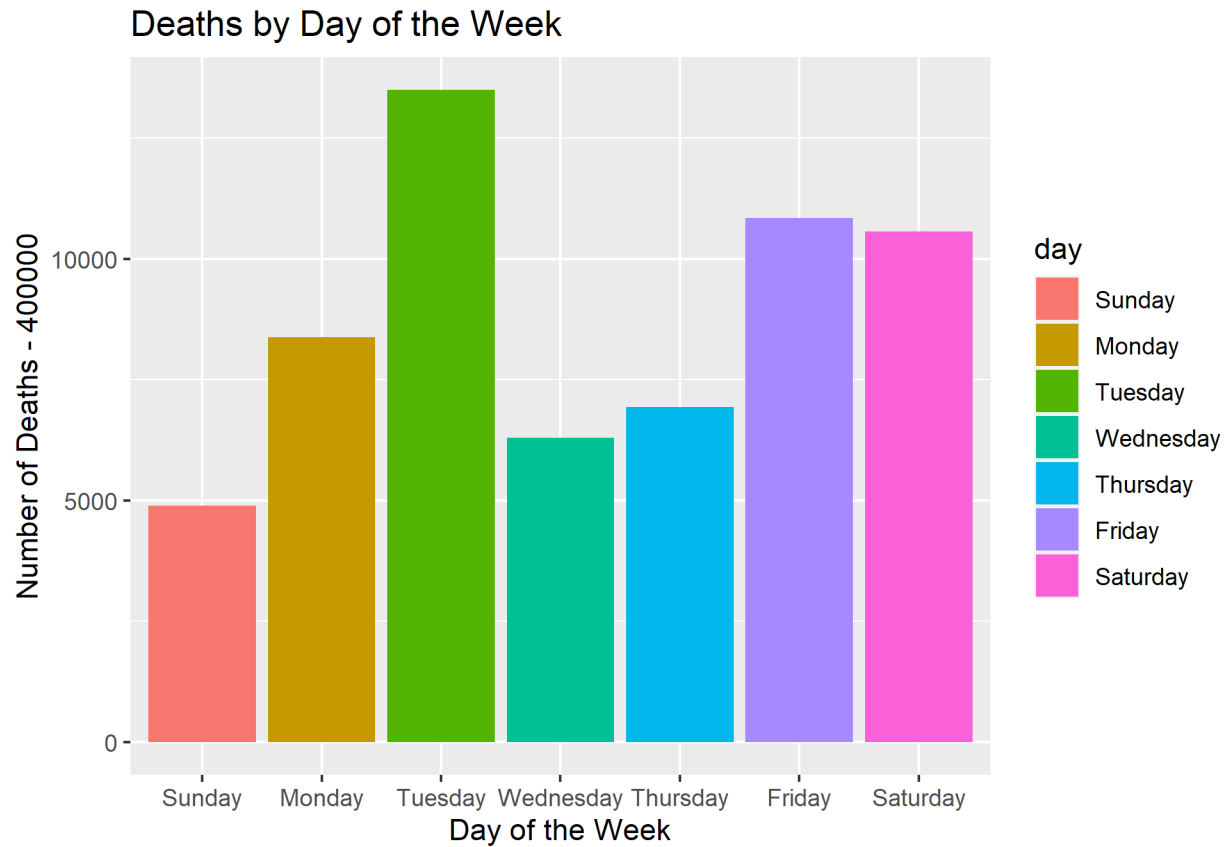
We are interested in what different factors are correlated with higher death rates based on cause of death, age groups, and weekdays. Depending on the different causes, different policies can be recommended. Moreover, we can break down the different causes of death by demographics such as race, age, and sex to determine where resources should specifically be directed.

## Exploration

### Deaths by Weekday

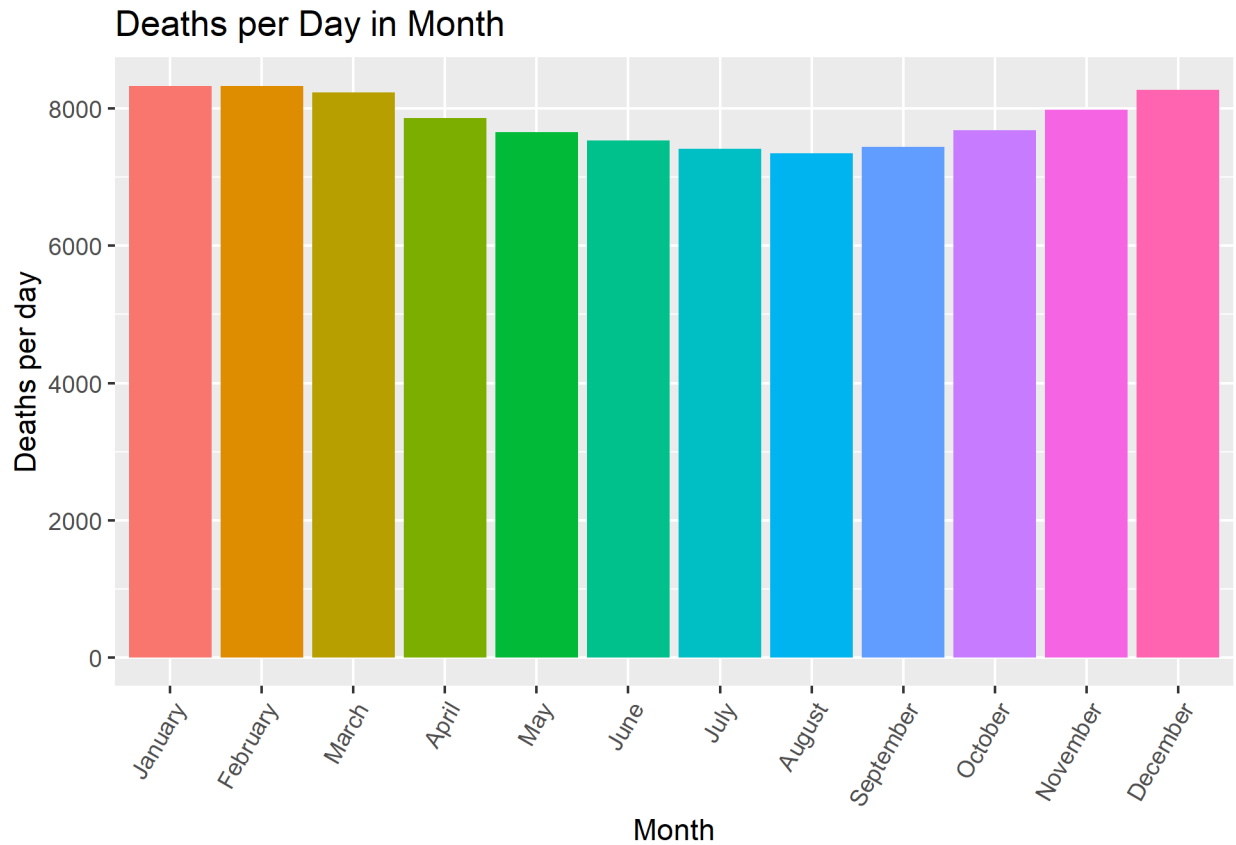
First, we plotted weekday of death versus death counts. There were the most deaths on Tuesday. However, days have an average of 7839.79 deaths and 2019 had an extra Tuesday so adjusting for that, the most deaths were on Fridays.





## Deaths by Month

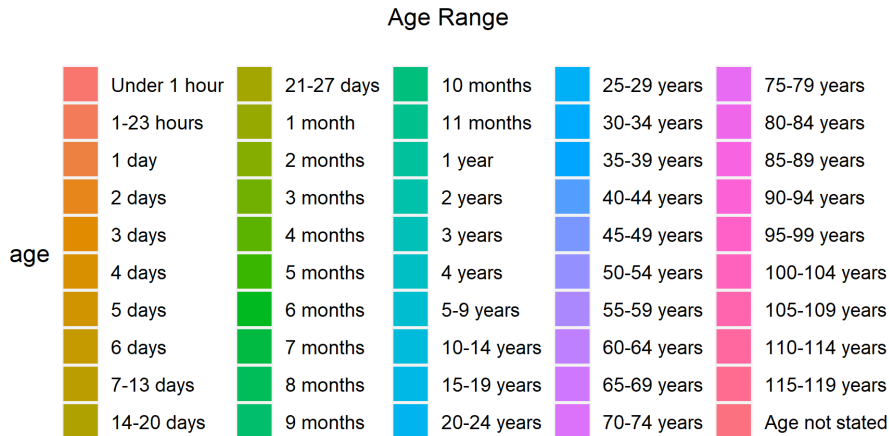
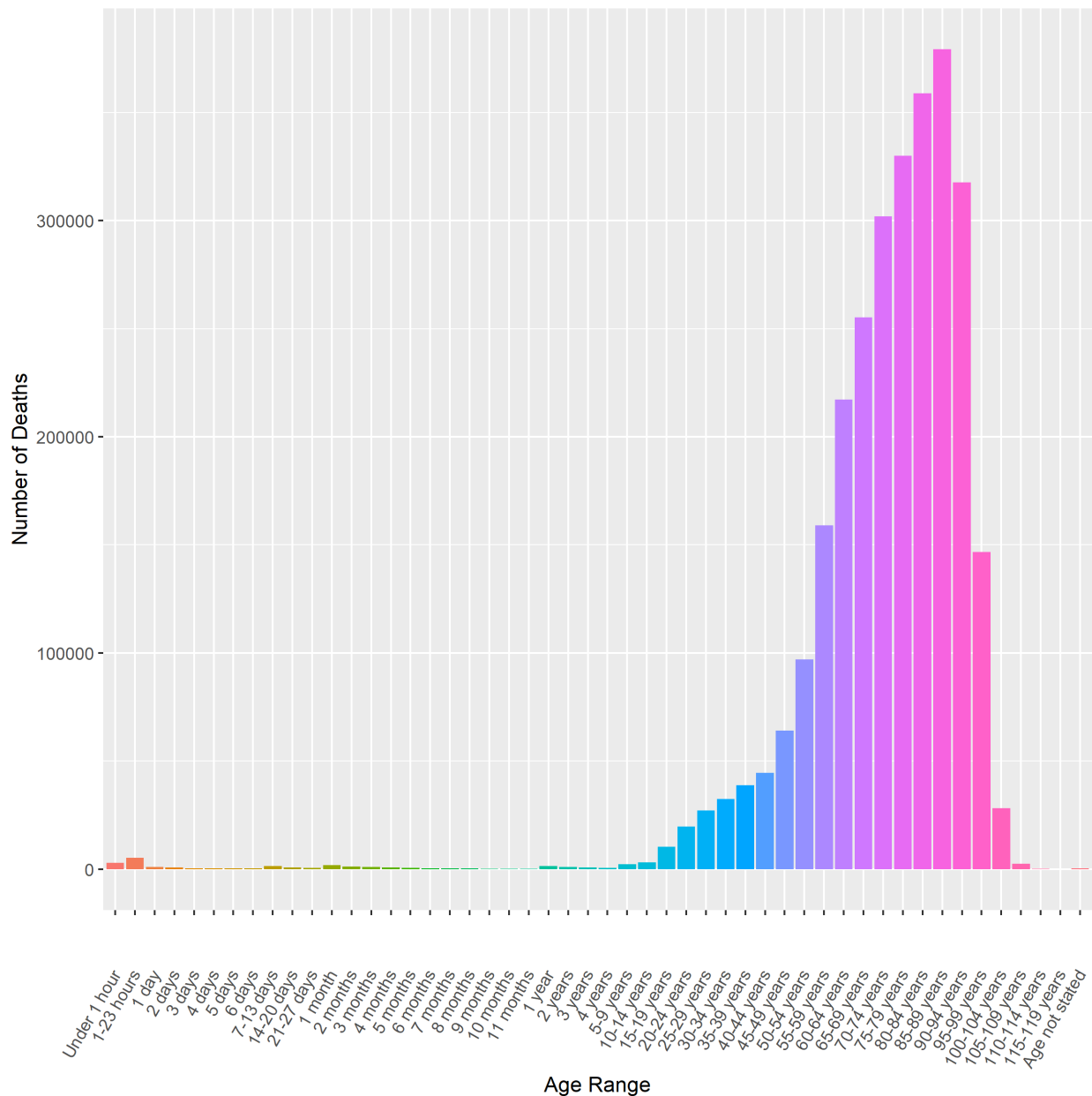
The most deaths occur in the coldest and darkest months of the year which are February, January, December, and March. Summer months have lower deaths by around 10-11%.



## Deaths by Age

Next, we plotted age versus death counts. Deaths were most prevalent among older age groups such as those between 70 and 84, although deaths start increasing more quickly at age 60. There is also a spike in those less than 1 day old. However, those greater than 1 day old do not frequently die.

Deaths by Age



## Deaths by Manner

Here, we plotted the manner of death versus age and counted how many people of a certain aged died based on a certain manner of death. A few key finding of this analysis shows that the majority of people die from natural causes, especially those aged 60+ and less than 1 day old, and accidental causes, spanning across all age groups. What this plot may help to inform us about is the behavior and activities that people in a general age group may commonly engage in that may have lead to their manner of passing. By being observant of the manners of death based on age group, preventative methods can be used to decrease the number of accidental related deaths if we are able to determine commonly engaged activities for age groups. Using this plot will help us answer the cause of death among the different age groups, and further promote research in what actual activities people are participating in that lead to their manner of death.

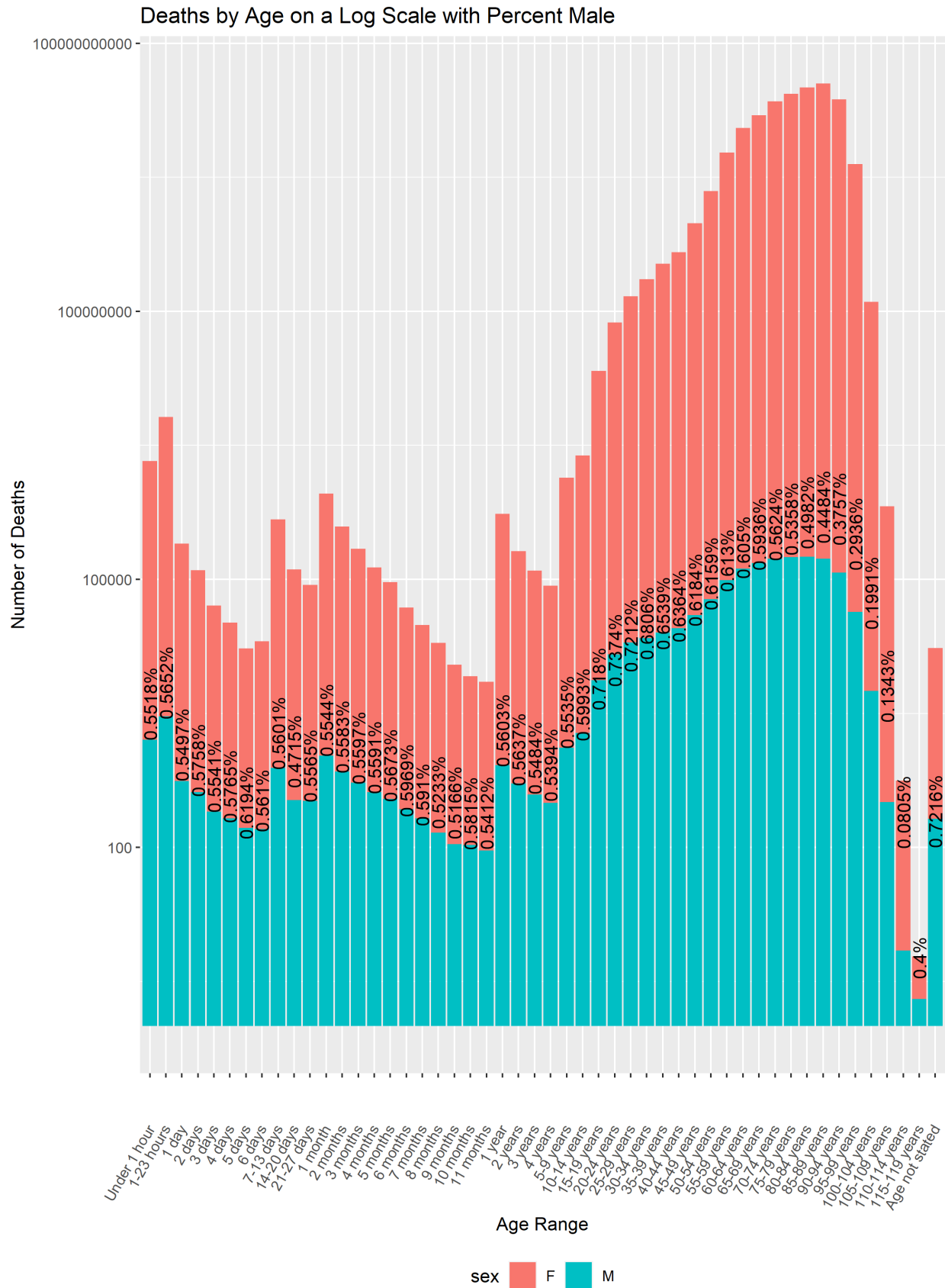
Figure 1 consists of two heatmaps showing the spatial distribution of the number of COVID-19 cases in the Veneto region. The top heatmap shows the number of cases per 100,000 inhabitants, and the bottom heatmap shows the number of cases per 100,000 inhabitants. Both heatmaps use a color scale from light blue (low) to dark blue (high). The top heatmap shows a high concentration of cases in the northern part of the region, while the bottom heatmap shows a more uniform distribution.

Under 1 hour  
1-2 hours  
3-5 days  
6-10 days  
11-20 days  
21-30 days  
1-3 months  
3-6 months  
6-12 months  
12-18 months  
18-24 months  
24-36 months  
36-48 months  
48-60 months  
60-72 months  
72-84 months  
84-96 months  
96-108 months  
108-120 months  
120-132 months  
132-144 months  
144-156 months  
156-168 months  
168-180 months  
180-192 months  
192-204 months  
204-216 months  
216-228 months  
228-240 months  
240-252 months  
252-264 months  
264-276 months  
276-288 months  
288-300 months  
300-312 months  
312-324 months  
324-336 months  
336-348 months  
348-360 months  
360-372 months  
372-384 months  
384-396 months  
396-408 months  
408-420 months  
420-432 months  
432-444 months  
444-456 months  
456-468 months  
468-480 months  
480-492 months  
492-504 months  
504-516 months  
516-528 months  
528-540 months  
540-552 months  
552-564 months  
564-576 months  
576-588 months  
588-600 months  
600-612 months  
612-624 months  
624-636 months  
636-648 months  
648-660 months  
660-672 months  
672-684 months  
684-696 months  
696-708 months  
708-720 months  
720-732 months  
732-744 months  
744-756 months  
756-768 months  
768-780 months  
780-792 months  
792-804 months  
804-816 months  
816-828 months  
828-840 months  
840-852 months  
852-864 months  
864-876 months  
876-888 months  
888-900 months  
900-912 months  
912-924 months  
924-936 months  
936-948 months  
948-960 months  
960-972 months  
972-984 months  
984-996 months  
1000 months  
Age not stated



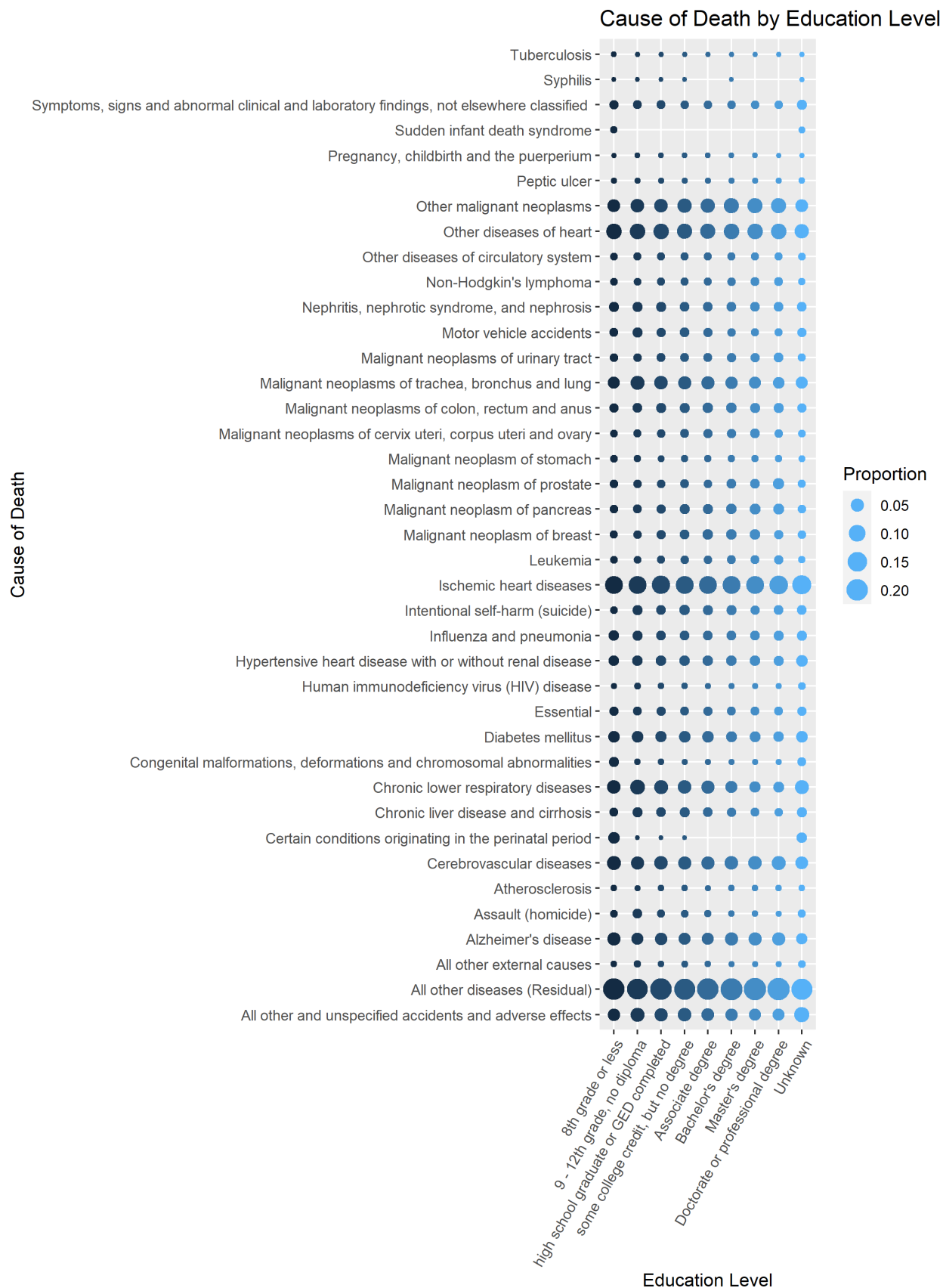
## Deaths by Age and Gender

In this plot, we plotted the number of deaths versus age ranges while demonstrating how many men compared to women passed away in each age category. In each of the bars, the red fill represents the amount of women who passed away in that particular age range while the blue accounts for the amount of men. The percentage seen in each bar represents the proportion of men in a given age range that passed away. This analysis shows that the majority of people under the age of 80 who pass away tend to be men, as nearly every bar from ages 0-80 shows the proportion of male deaths to be above 50%. This proportion of male deaths goes down after 80 years of age, and is likely because women who are of an older age tend to live to a complete life expectancy. What this plot may help to inform us about is the differences in male and females lives and life expectancies. Further research into differences in lifestyle choices for men versus women as a whole may help better explain why women tend to live longer than men. Furthermore, this plot accompanied with a plot on cause of death by gender, may assist in determining what kind of, potentially more risky, behaviors men may partake in during their lifetimes that lead to an earlier death than women.



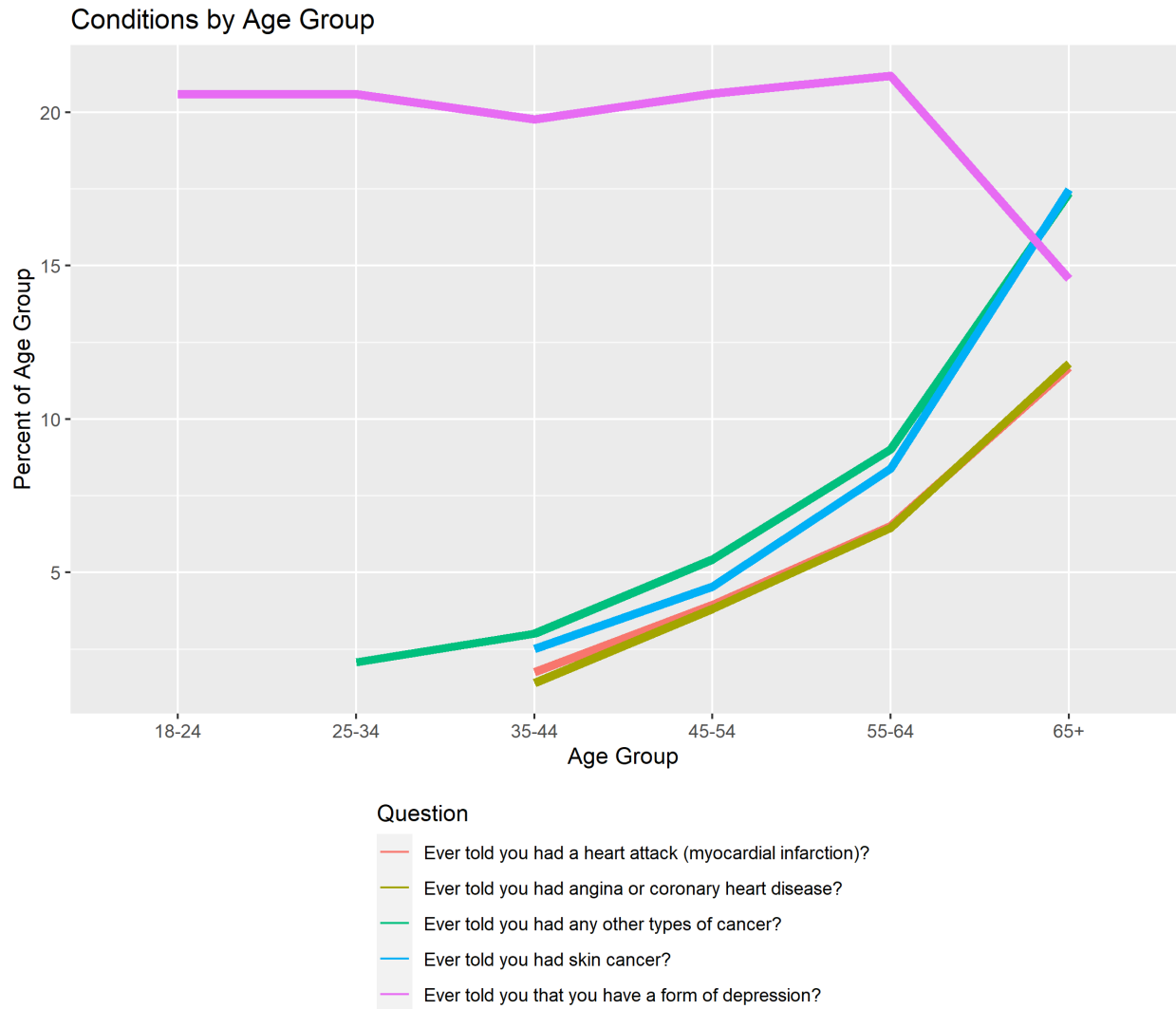
## **Cause of Death by Education**

For most causes of death, level of education does not have an impact on what proportion of people have that cause of death. The largest difference belongs to “Certain conditions originating in the perinatal period” with high occurrences in those with 8th grade or less education and those with unknown education and nearly no occurrences in all others. Another large proportion difference is in “Congenital malformations” where 8th grade or less has a much higher mortality proportion than other education levels. For causes of death that are not highly tied to conditions at birth, “Syphilis” and “Assault (homicide)” have the highest differing proportions. “Syphilis” has a quite small sample size but unknown education has the highest mortality proportion. For “Assault (homicide)”, 9 - 12th grade, no diploma has the highest mortality proportion.



## Free Text Analysis: Selected Health Conditions by Age Group

Using regular expressions, we polled the BRFSS data set for questions related to heart conditions, cancer, and depression, while grouping by the age of respondents. Furthermore, we restricted entries to those where participants responded positively, indicating that they did have those conditions. Interestingly, all age groups except 65+ had a high incidence of depression, hovering around the 20% mark. This dips significantly to 15% for the 65+ age group, perhaps because mental health was more stigmatized during their lives and psychological diagnoses were less readily available. In addition, the orange and olive green lines (for heart attack and coronary heart disease, respectively) have a significant degree of overlap, which makes sense, given the conditions. The green line (corresponding to non-skin cancer), is slightly higher than the blue line (corresponding to skin cancer) for all age groups. Furthermore, both cancer have positive slopes, indicating that as your age increases, you become more likely to be diagnosed with cancer.



## Killer

This plot demonstrates the most common manners of death among people in different cross sections of age and marriage. Head scale is determined by natural causes. Neck scale is determined by pending investigation. Left arm scale is determined by accident. Right arm scale is determined by homicide. Left leg scale is determined by suicide. Right leg scale is determined by could not determine.

Killer Plot Draft 1

