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

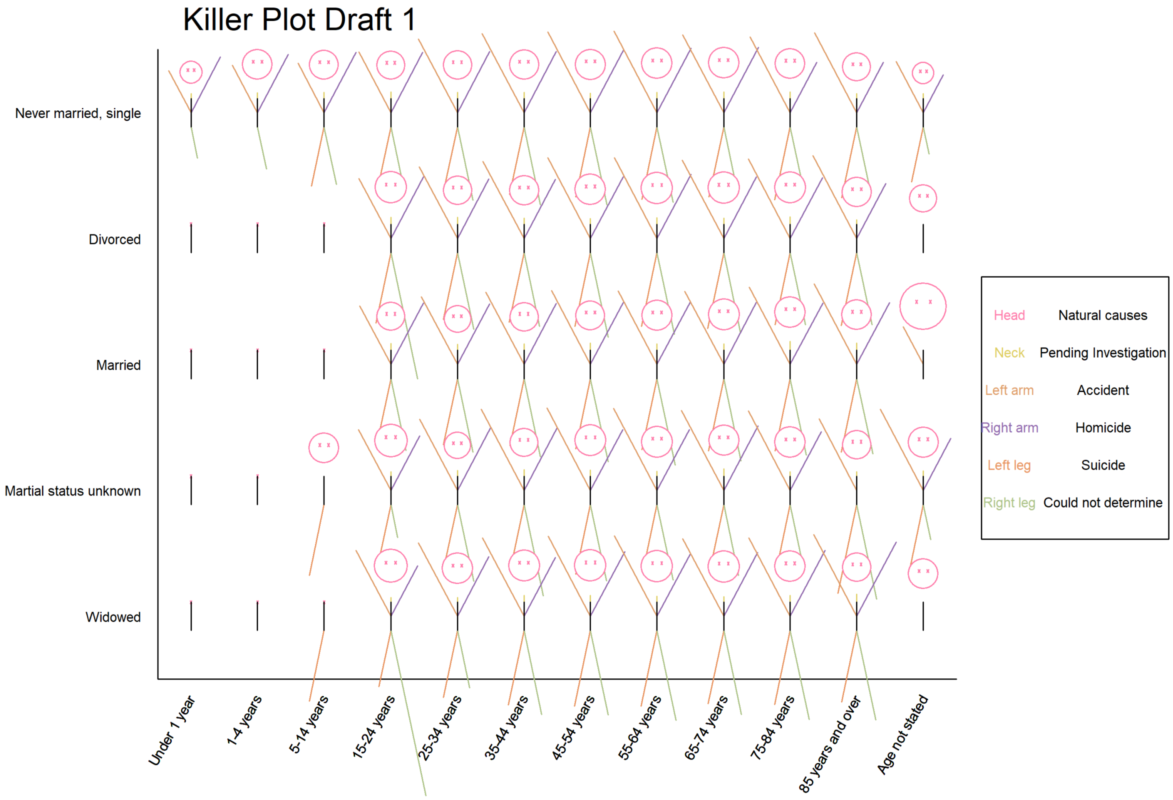
For our secondary dataset, we are using the Behavioral Risk Factor Surveillance System Survey. This survey includes different free text survey questions from across the United States and territories with responses broken out by subgroup. There is also information on sample size, percent affirmative response, 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.

Killer

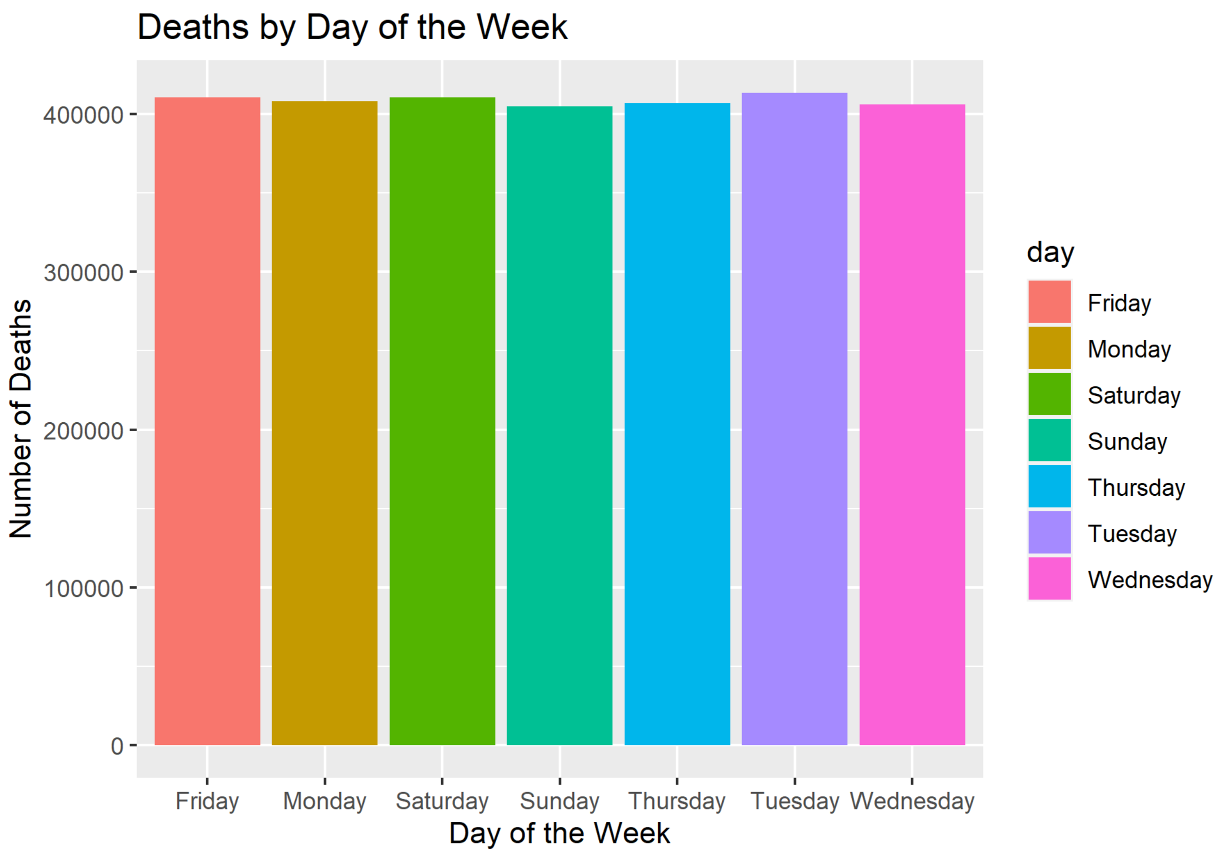
This plot demonstrates the most common manners of death among people in different cross sections of age and marriage. Head scale is determined by natural causes. Neck scale is determined by pending investigation. Left arm scale is determined by accident. Right arm scale is determined by homicide. Left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

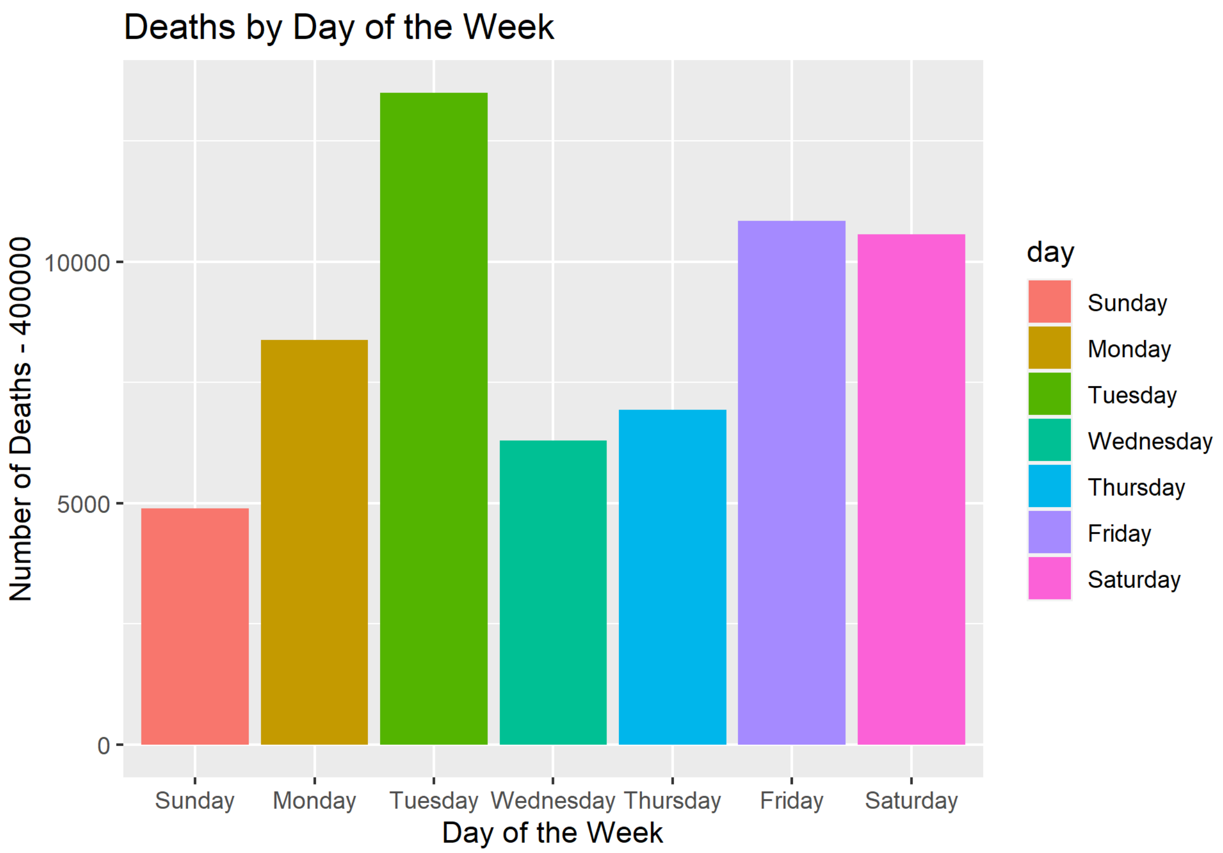


Exploration

Deaths by Weekday

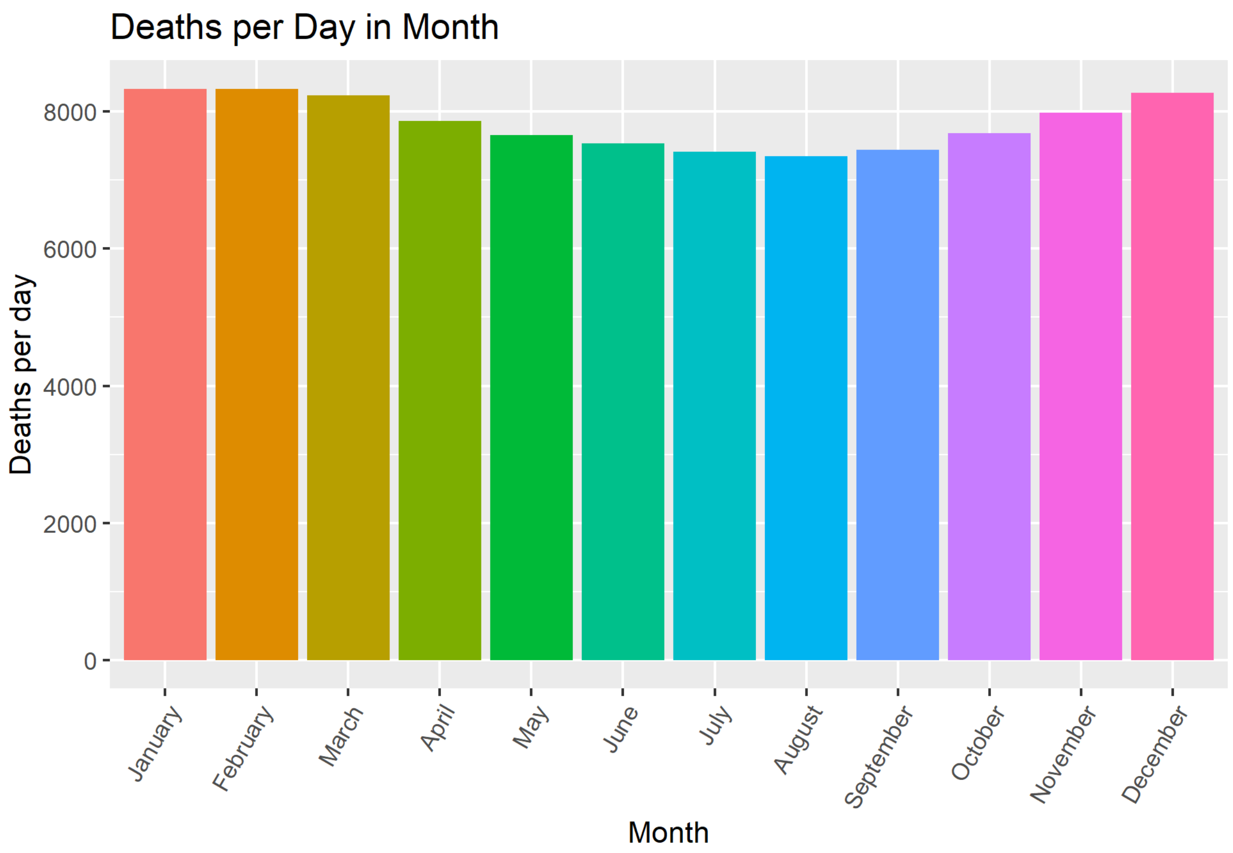
First, we plotted weekday of death versus death counts. There were the most deaths on Tuesday. However, days have an average of 7839.79 deaths and 2019 had an extra Tuesday so adjusting for that, the most deaths were on Fridays.


Deaths by Day of the Week
Number of Deaths
0
100000
200000
300000
400000
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day of the Week
day
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday



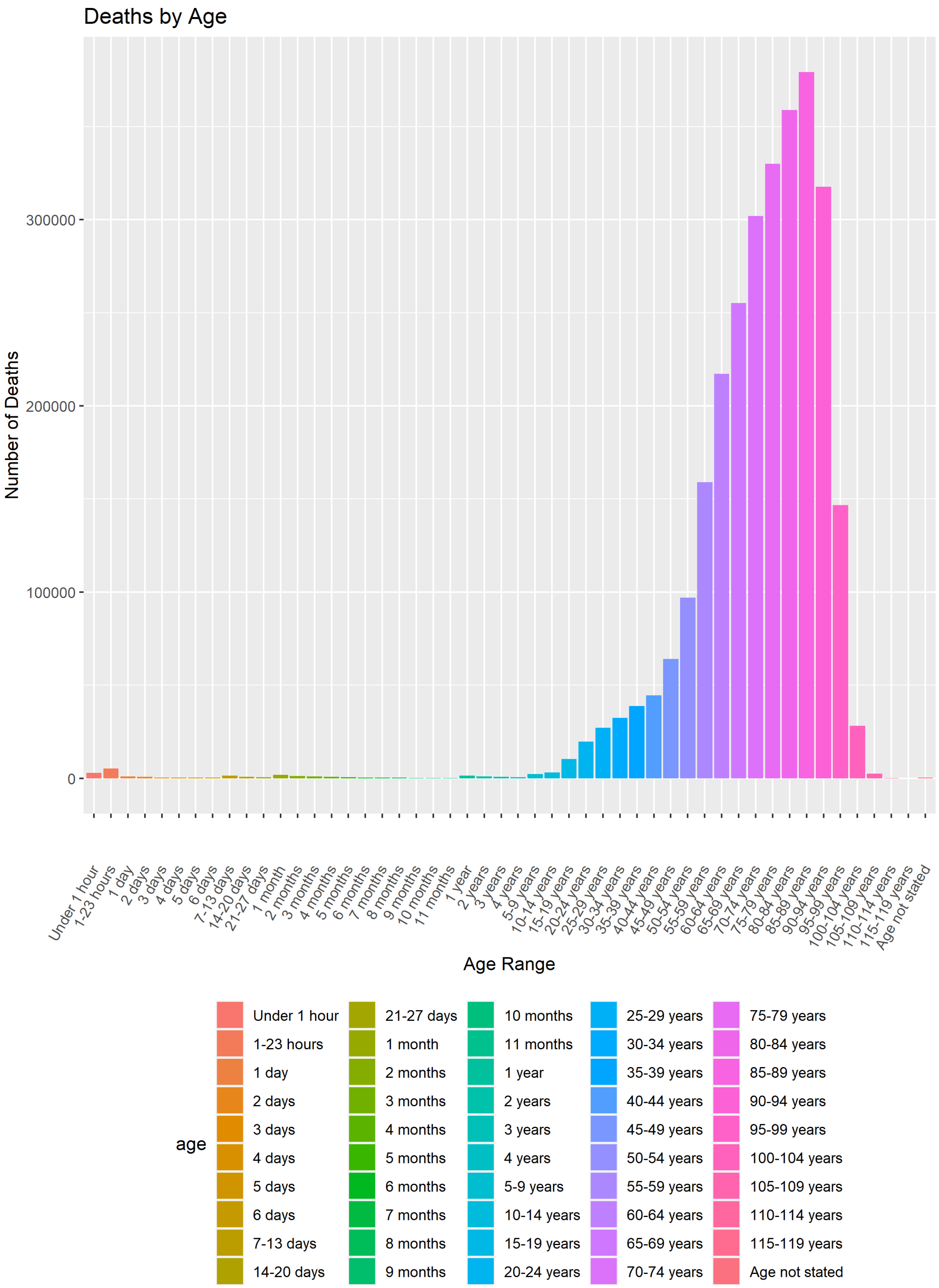
Deaths by Month

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%.



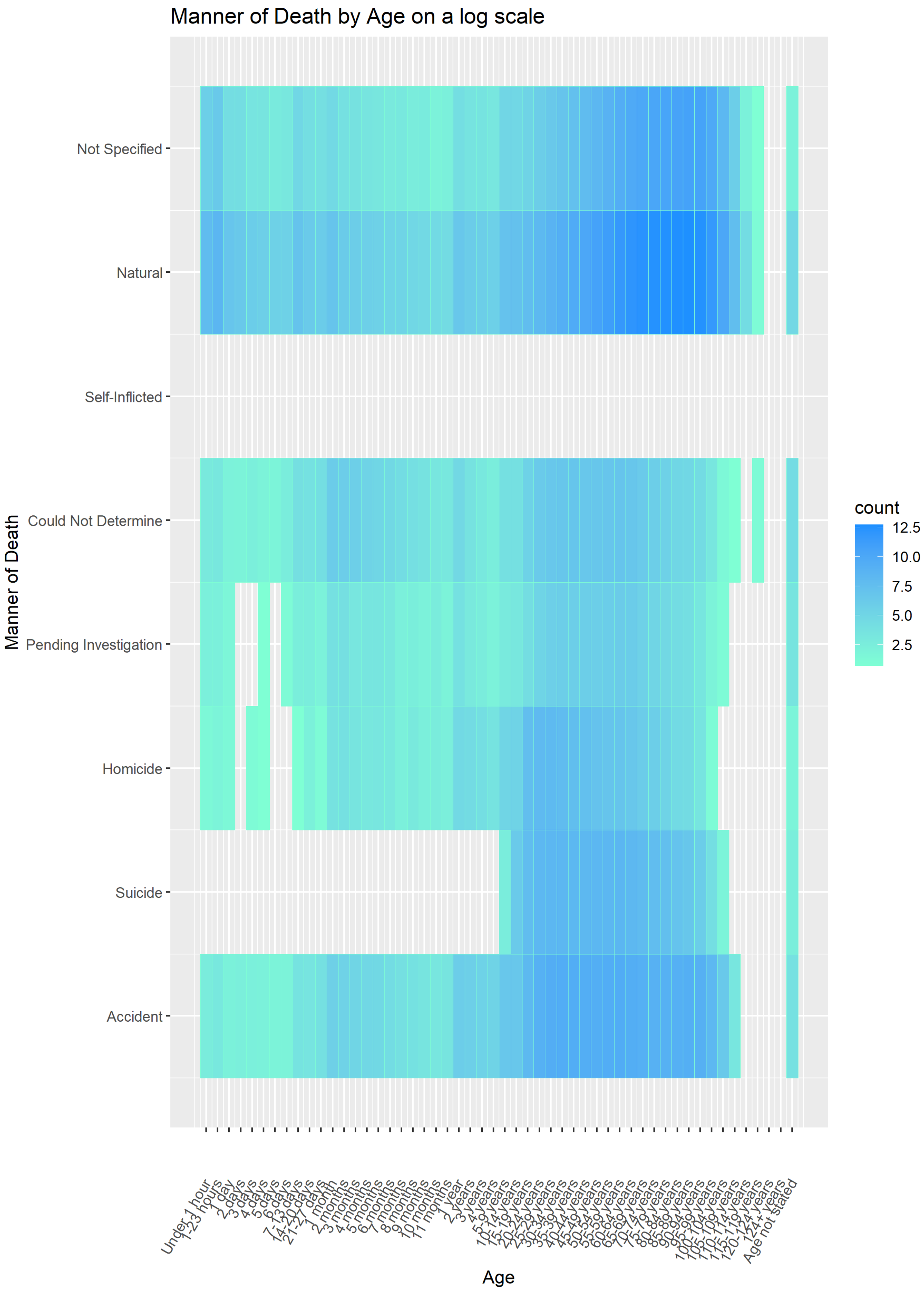
Deaths by Age

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
Number of Deaths
300000
200000
100000
0
Age Range
Under 1 hour
1-23 hours
1 day
2 days
3 days
4 days
5 days
6 days
7-13 days
14-20 days
21-27 days
1 month
2 months
3 months
4 months
5 months
6 months
7 months
8 months
9 months
10 months
11 months
1 year
2 years
3 years
4 years
5-9 years
10-14 years
15-19 years
20-24 years
25-29 years
30-34 years
35-39 years
40-44 years
45-49 years
50-54 years
55-59 years
60-64 years
65-69 years
70-74 years
75-79 years
80-84 years
85-89 years
90-94 years
95-99 years
100-104 years
105-109 years
110-114 years
115-119 years
Age not stated
age

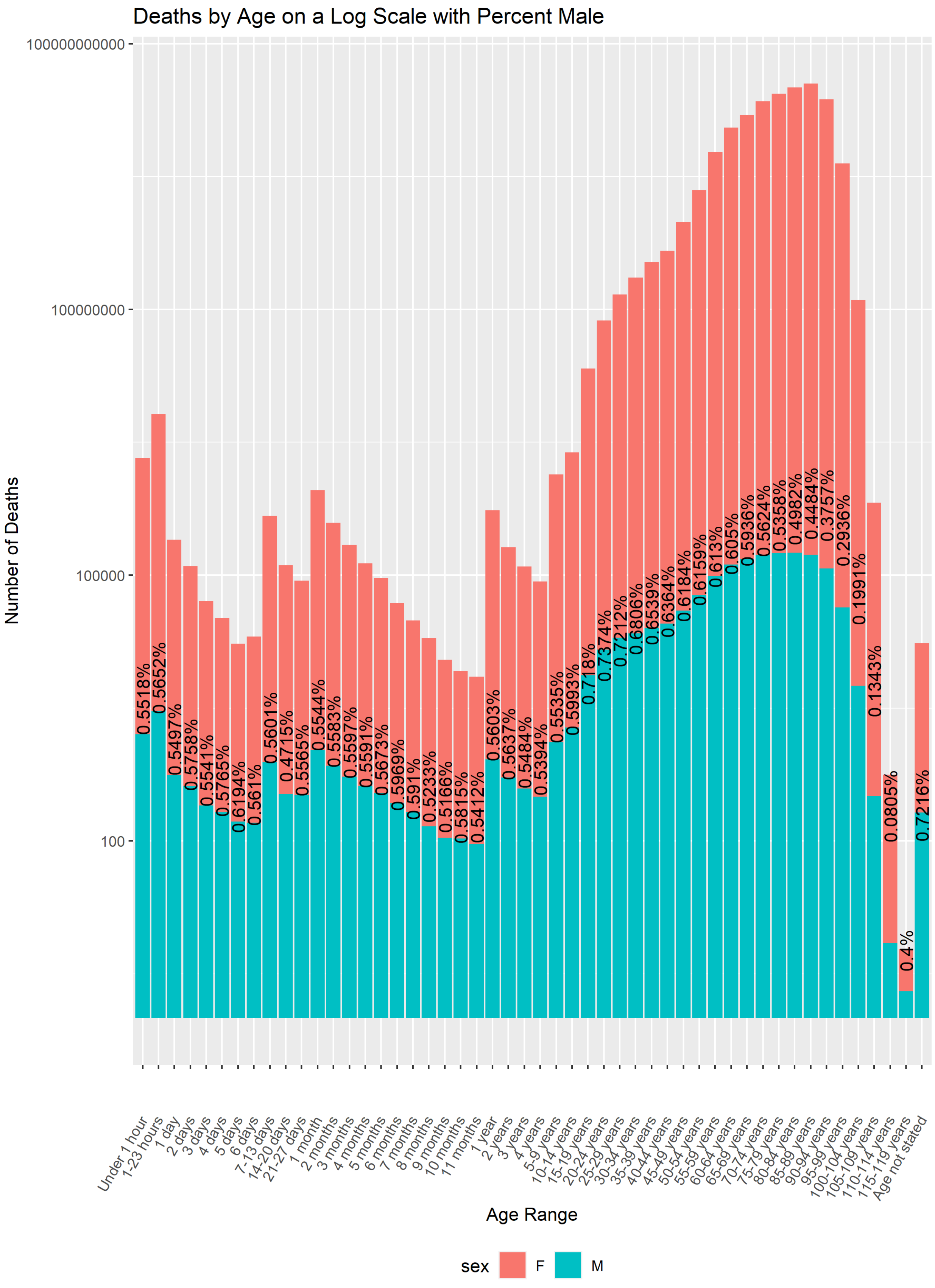
Deaths by Manner

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.


Manner of Death by Age on a log scale
Manner of Death
Not Specified
Natural
Self-Inflicted
Could Not Determine
Pending Investigation
Homicide
Suicide
Accident
count
12.5
10.0
7.5
5.0
2.5
Under 1 hour
1-23 hours
1 day
2 days
3 days
4 days
5 days
6 days
7-13 days
14-20 days
21-27 days
1 month
2 months
3 months
4 months
5 months
6 months
7 months
8 months
9 months
10 months
11 months
1 year
2 years
3 years
4 years
5-9 years
10-14 years
15-19 years
20-24 years
25-29 years
30-34 years
35-39 years
40-44 years
45-49 years
50-54 years
55-59 years
60-64 years
65-69 years
70-74 years
75-79 years
80-84 years
85-89 years
90-94 years
95-99 years
100-104 years
105-109 years
110-114 years
115-119 years
120-124 years
124+ years
Age not stated
Age

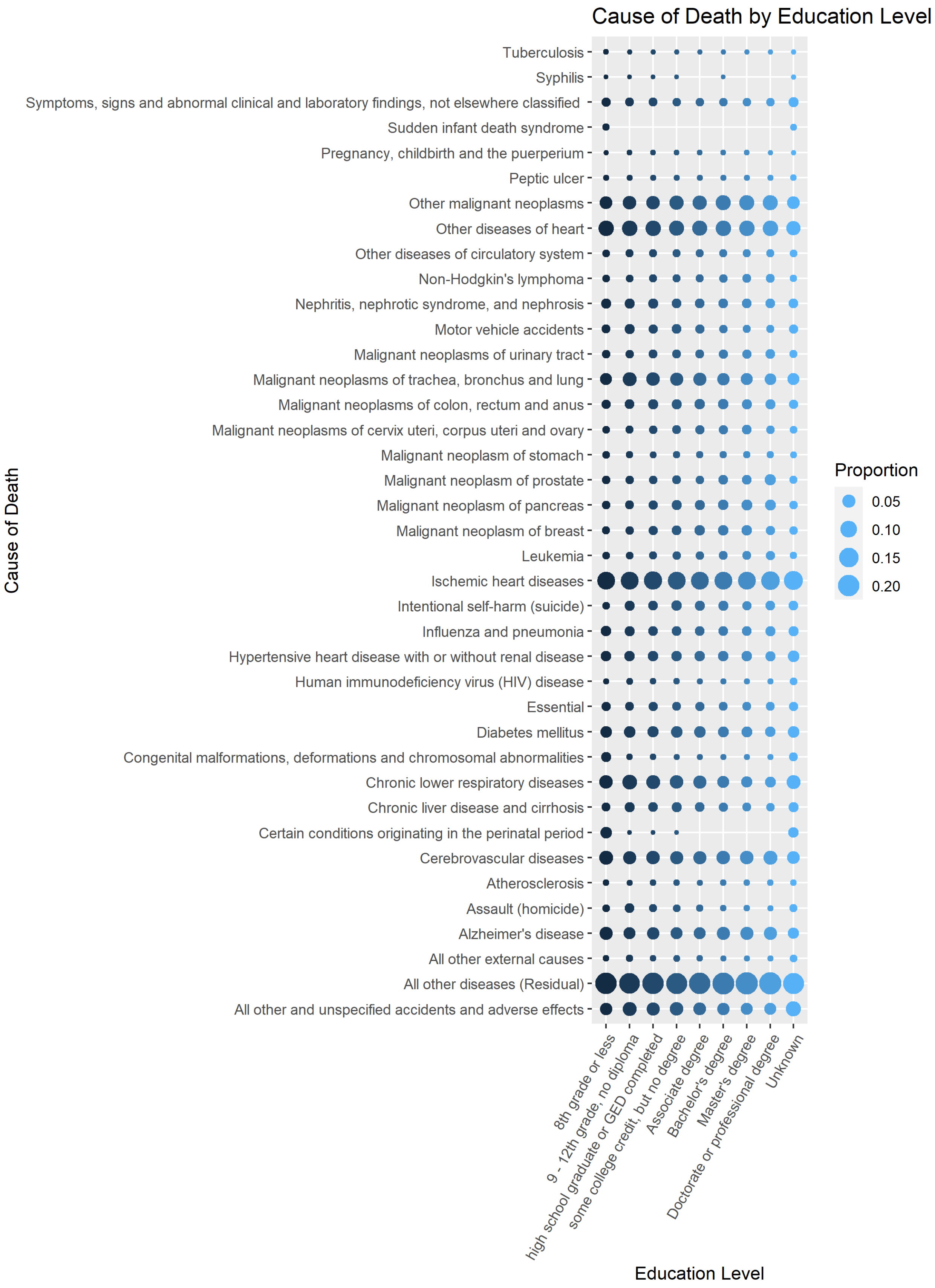
Deaths by Age and Gender

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.


Deaths by Age on a Log Scale with Percent Male

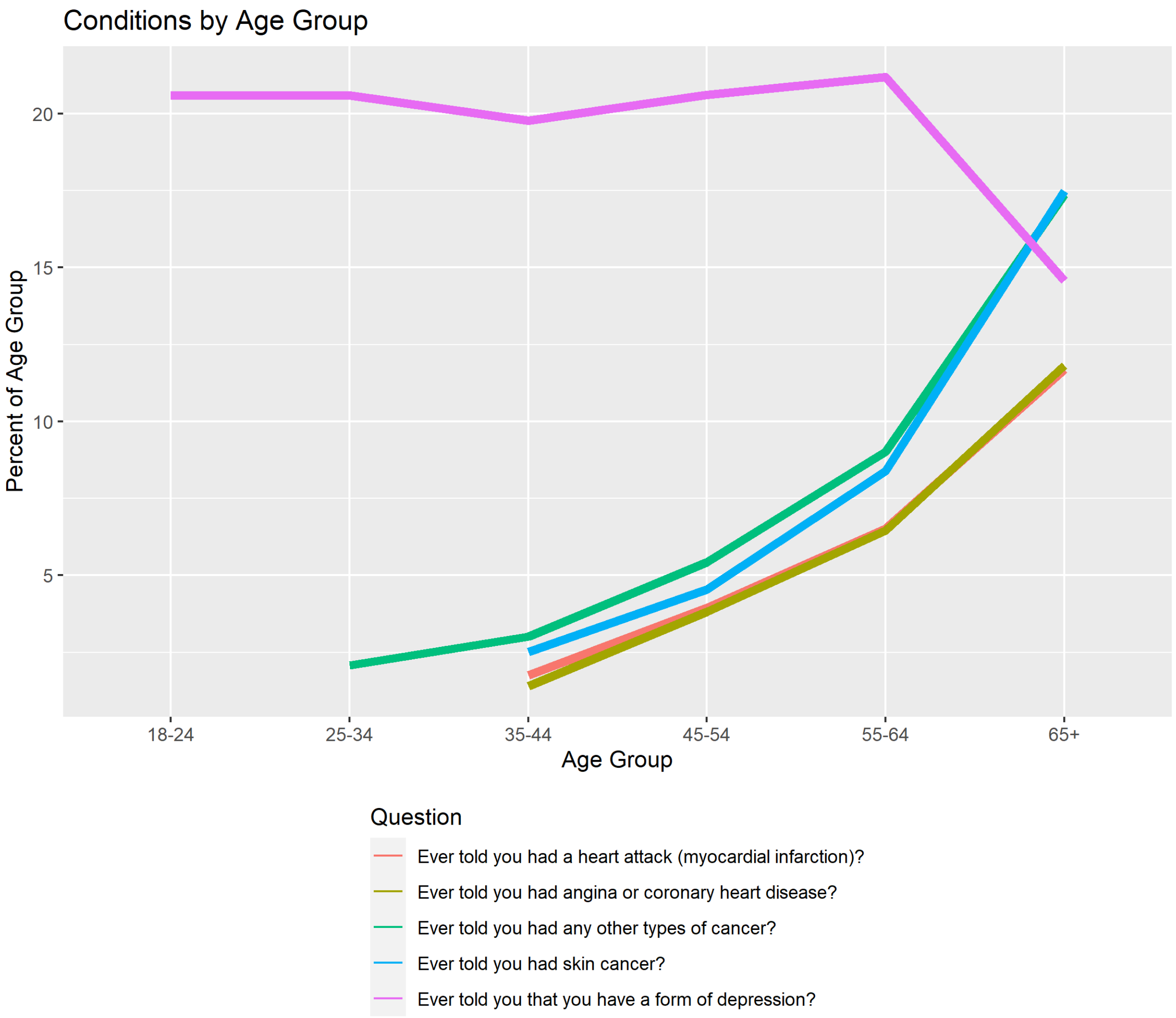
Cause of Death by Education

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.


Cause of Death by Education Level
Cause of Death
Tuberculosis
Syphilis
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
Sudden infant death syndrome
Pregnancy, childbirth and the puerperium
Peptic ulcer
Other malignant neoplasms
Other diseases of heart
Other diseases of circulatory system
Non-Hodgkin's lymphoma
Nephritis, nephrotic syndrome, and nephrosis
Motor vehicle accidents
Malignant neoplasms of urinary tract
Malignant neoplasms of trachea, bronchus and lung
Malignant neoplasms of colon, rectum and anus
Malignant neoplasms of cervix uteri, corpus uteri and ovary
Malignant neoplasm of stomach
Malignant neoplasm of prostate
Malignant neoplasm of pancreas
Malignant neoplasm of breast
Leukemia
Ischemic heart diseases
Intentional self-harm (suicide)
Influenza and pneumonia
Hypertensive heart disease with or without renal disease
Human immunodeficiency virus (HIV) disease
Essential
Diabetes mellitus
Congenital malformations, deformations and chromosomal abnormalities
Chronic lower respiratory diseases
Chronic liver disease and cirrhosis
Certain conditions originating in the perinatal period
Cerebrovascular diseases
Atherosclerosis
Assault (homicide)
Alzheimer's disease
All other external causes
All other diseases (Residual)
All other and unspecified accidents and adverse effects
8th grade or less
9 - 12th grade, no diploma
high school graduate or GED completed
some college credit, but no degree
Associate degree
Bachelor's degree
Master's degree
Doctorate or professional degree
Unknown
Education Level
Proportion
0.05
0.10
0.15
0.20

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 then 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.


Conditions by Age Group
Percent of Age Group
20
15
10
5
18-24
25-34
35-44
45-54
55-64
65+
Age Group
Question
Ever told you had a heart attack (myocardial infarction)?
Ever told you had angina or coronary heart disease?
Ever told you had any other types of cancer?
Ever told you had skin cancer?
Ever told you that you have a form of depression?