

Stat 405 Mortality

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

December 5th, 2022

Introduction

Masaki Kobayashi's *Harakiri* (1962) is a Japanese film about a samurai who asks to commit ritual suicide at a lord's palace. Throughout the film, the audience learns the story of what brings the samurai to make the request. At the palace, the samurai argues with and disrespects the lord's samurais, in revenge for past wrongs. These layers of disrespect lead to conflict and the main samurai kills many of the lord's in combat. The film ends with the lord's history where the events of the film are manipulated and recorded incorrectly to preserve honor. Contrastingly, when a person dies in the modern United States, their causes of death and characteristics are meticulously recorded with substantial effort put in to accuracy — not for honor — but for statistics. We are here to do those statistics.

Resources: [Website](#) [PDF](#) [Slides](#) [Github](#)

Primary Dataset

The National Bureau of Economic Research creates and distributes a dataset of US mortality for every year since 1959. This dataset is unique for both its breadth and depth. Each row in the dataset represents a single death, and each column represents a different demographic characteristic of the deceased. The information is derived from death certificates filed by medical professionals in the 50 states plus Washington DC. We made the decision to use the 2019 edition of the dataset since we did not want to focus on COVID-19. 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. In 2019, there were 2,861,523 deaths total. The following are the 10 most common groups split up by race, age, education, and sex:

count	race	age	education	sex
91357	White	85-89 years	High school graduate or GED	F
88154	White	90-94 years	High school graduate or GED	F
76361	White	80-84 years	High school graduate or GED	F
61108	White	75-79 years	High school graduate or GED	F
61038	White	80-84 years	High school graduate or GED	M
60104	White	75-79 years	High school graduate or GED	M
57254	White	85-89 years	High school graduate or GED	M
53786	White	70-74 years	High school graduate or GED	M
50247	White	65-69 years	High school graduate or GED	M
48484	White	70-74 years	High school graduate or GED	F

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 combine the secondary dataset by matching up subgroups between the death dataset and the risk factor dataset and using aggregate statistics and free text to analyze how risk factors can be matched with causes of death.

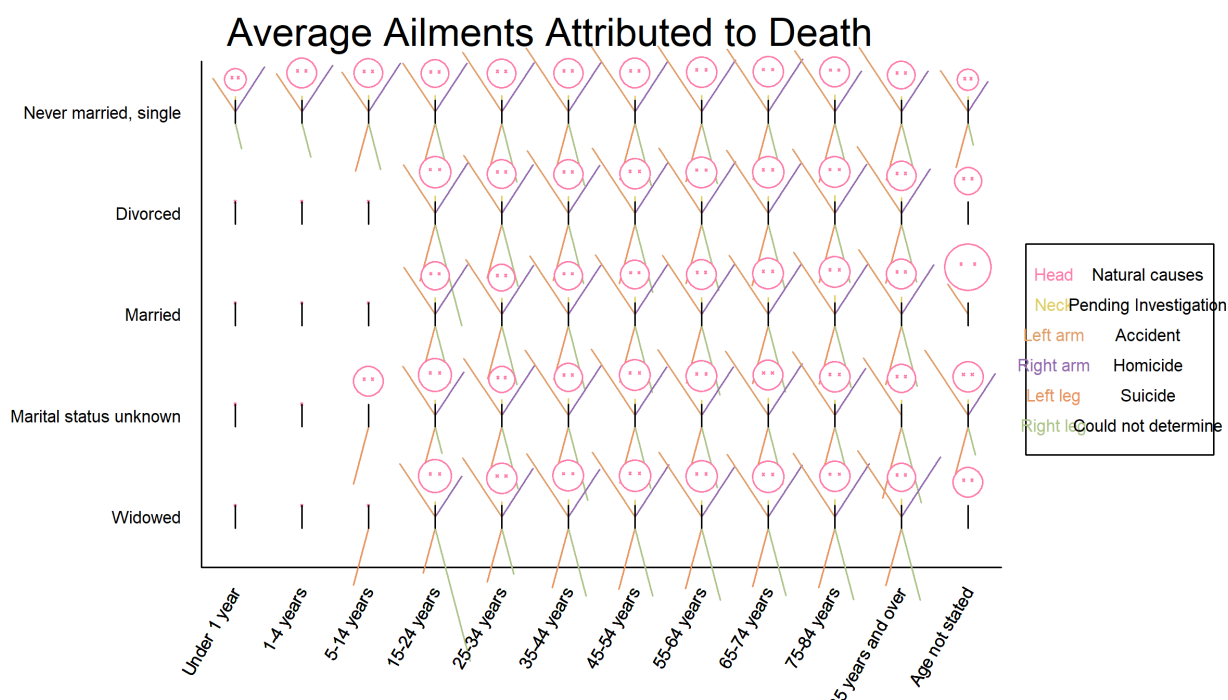
Questions

The main question we are asking is: given someone is dead, can we predict how they died? We will approach this by looking at factors such as age, gender, place of death, educational, health conditions, and race, and build models for manner of death.

Additionally, we observe if there are any irregularities or trends in mortality when looking at specific factors such as the relationship between manner of death across months, deaths by age, deaths by gender and manner of death by education level. This will both inform us on trends in mortality among different demographic factors while also enhancing our predictions about how someone died based on the factors they identify with.

Killer Plot

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. Rather than determining size by relative count, size is determined by record count. A record is an ailment that is recorded on a death certificate. These follow the ICD-10 standard. Some examples of records are C159: Malignant neoplasm of esophagus and K766: Portal hypertension.



The following table displays the average number of records for deaths in the given age ranges. Most age groups die with about 3 records. The average gets lower for younger age groups and older age groups, peaking in the middle age ranges. The age range with the highest record average is 25-34 years with an average record count of 3.304.

Age	N	Average Record Count
Under 1 year	21012	2.181
1-4 years	3701	2.781
5-14 years	5541	2.826

Age	N	Average Record Count
15-24 years	29979	3.010
25-34 years	59543	3.304
35-44 years	83472	3.295
45-54 years	161212	3.219
55-64 years	376411	3.195
65-74 years	557075	3.200
75-84 years	689088	3.149
85 years and over	874198	2.982
Age not stated	291	2.577

The following table displays the average number of records for deaths in the given marital statuses. The average record count for all groups is around 3.1. There is not much variation by group, and the marital status with the highest record count is “Marital status unknown” with an average record count of 3.257. The range of these averages is 0.226, so there is very little variation by marital status.

Marital Status	N	Average Record Count
Divorced	478548	3.214
Married	1038238	3.134
Never married, single	403235	3.137
Marital status unknown	23155	3.257
Widowed	918347	3.031

The following table displays the average number of records for deaths by manner of death. Most manner of death groups die with about 3 records. The average is much lower for “Pending Investigation”, with an average of 1.311. The highest average record is for the manner of death “Accident” with an average of 4.005. Every other manner of death yields an average record count close to 3.

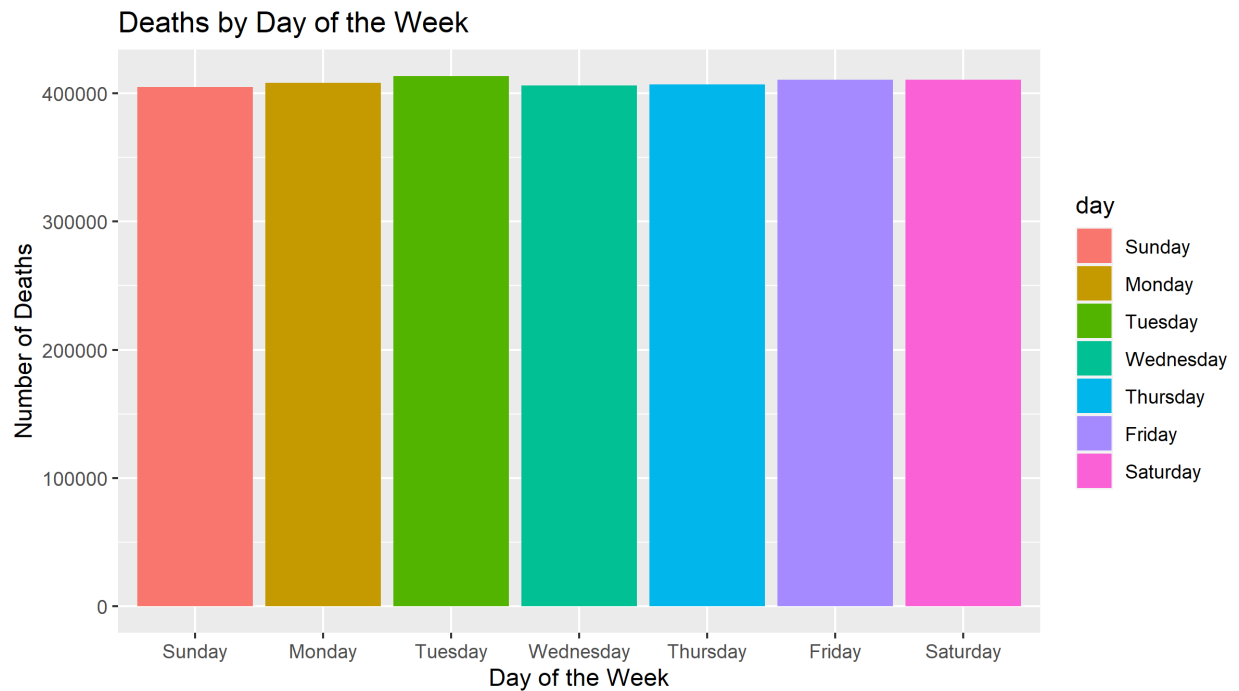
Manner	N	Average Record Count
Accident	173608	4.005
Suicide	47764	2.930
Homicide	20310	3.236

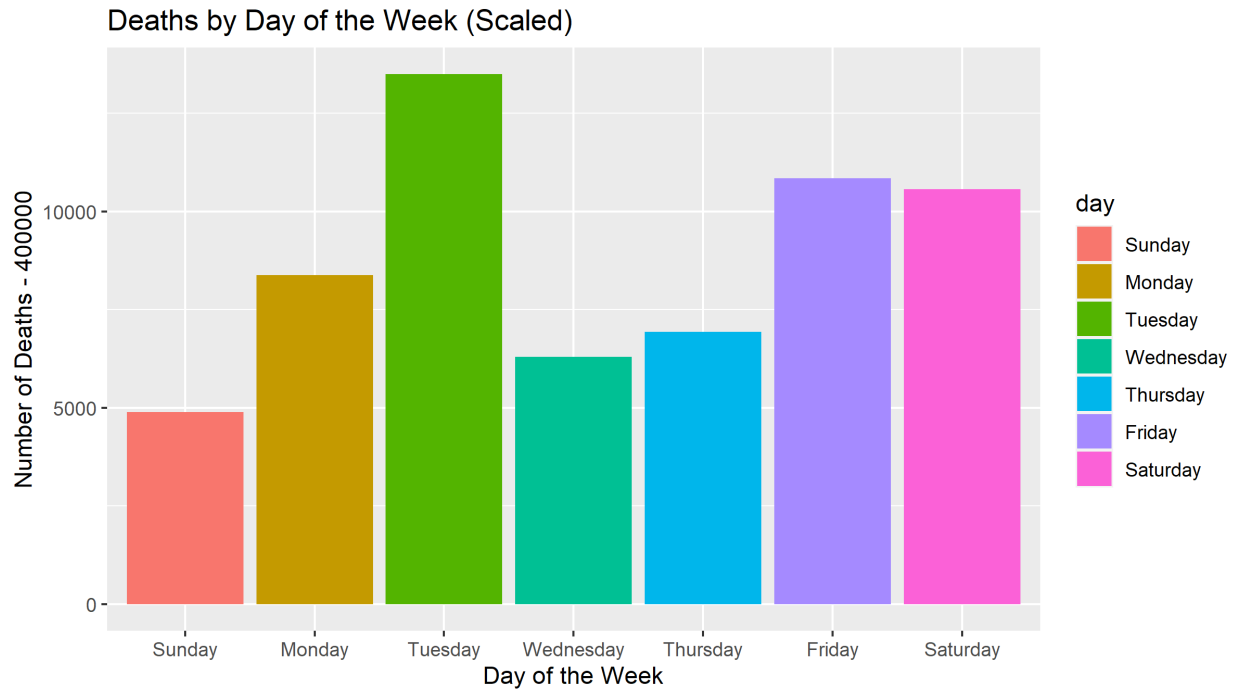
Manner	N	Average Record Count
Pending Investigation	4484	1.311
Could Not Determine	11800	2.884
Natural	2327811	3.028
Not Specified	275746	3.358

Exploration

Deaths by Weekday

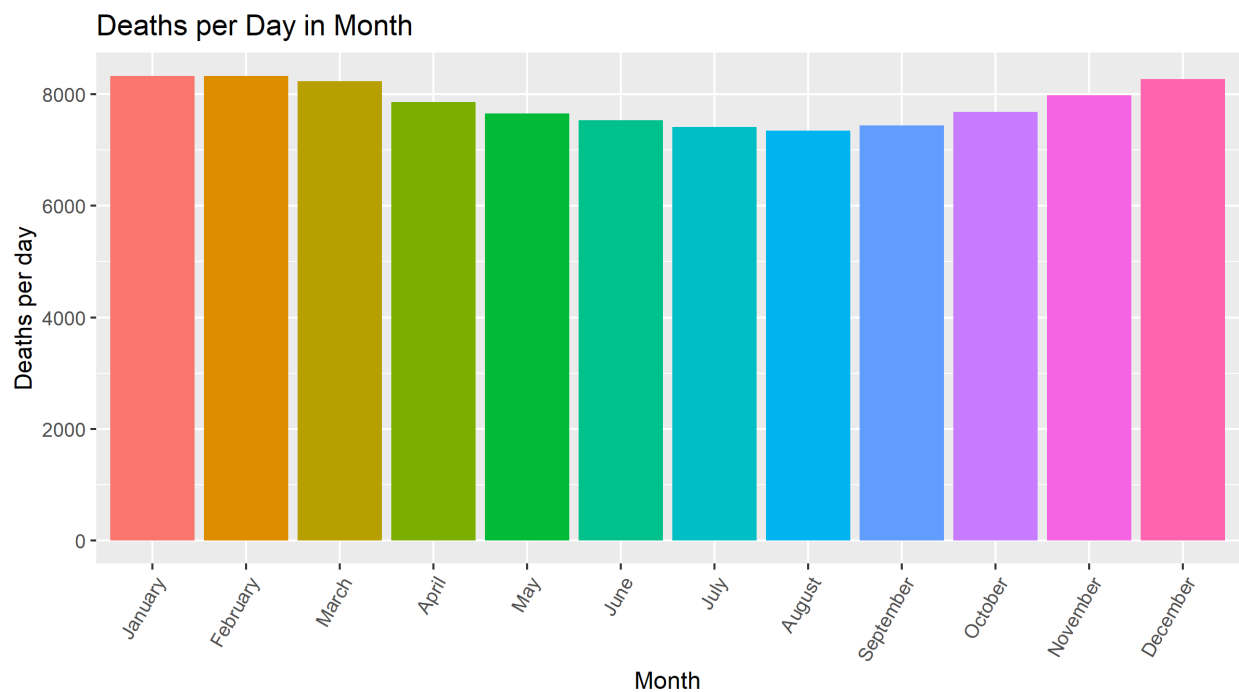
First, we plotted weekday of death versus death counts. There were the most deaths on Tuesday. However, days have an average of 7839.789 deaths and 2019 had an extra Tuesday so after adjusting for that, the most deaths were on Fridays.





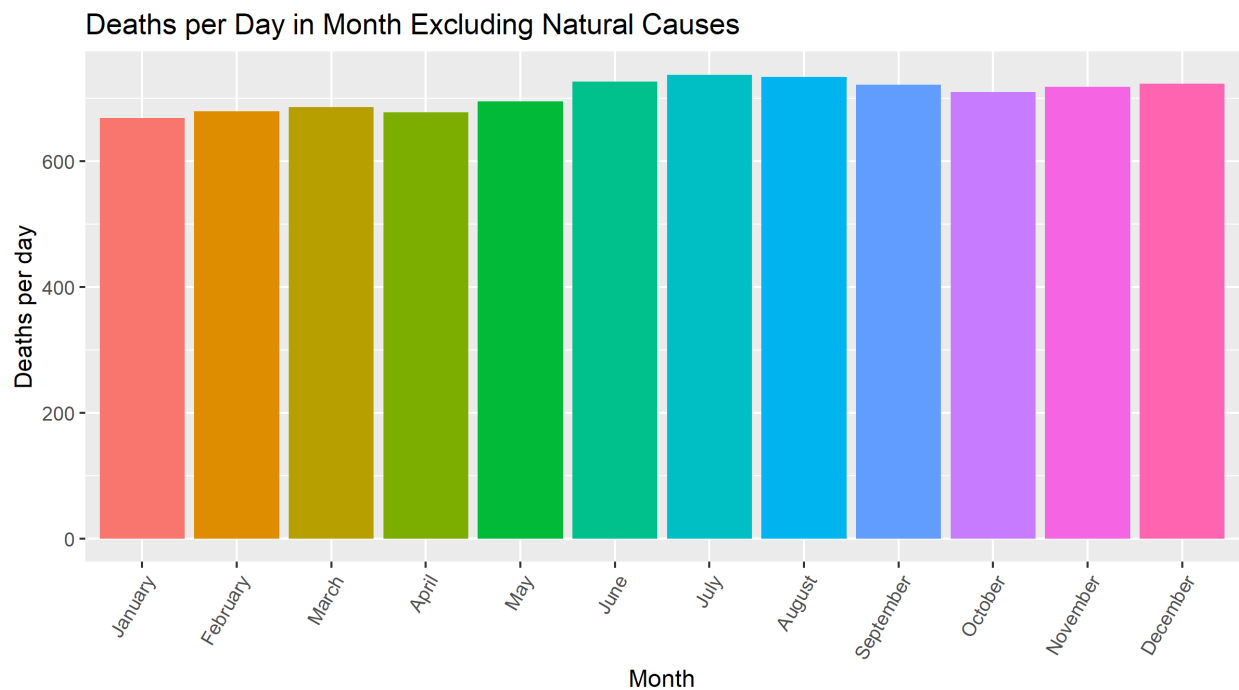
Deaths by Month

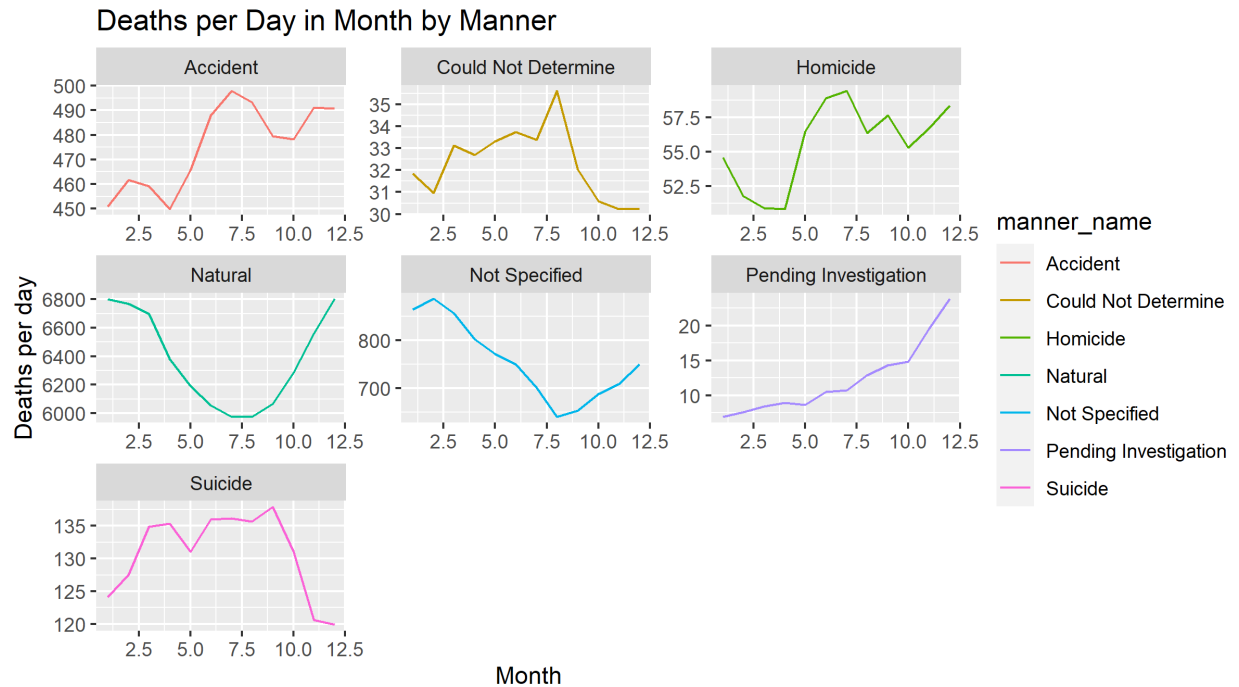
The most deaths occur in the coldest and darkest months of the year which are February, January, December, and March. Summer months have lower deaths by around 10-11%.



Interestingly, the large number of deaths in the winter months and lower numbers in summer can be explained by natural deaths. Since most deaths are due to natural causes, even a small increase in deaths due to natural causes can have a large impact on the total number of deaths. We speculate that the reasoning for increased deaths due to natural causes in winter months is because people spend more time inside with cold weather which leads to increased disease transmission.

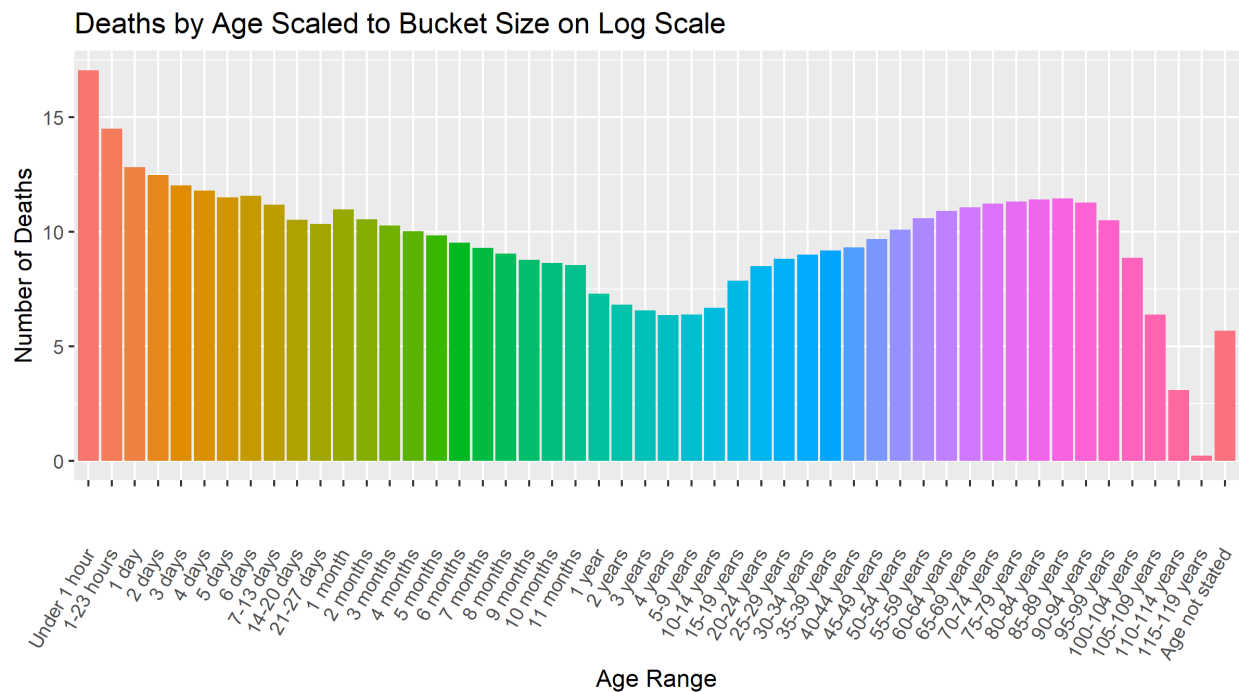
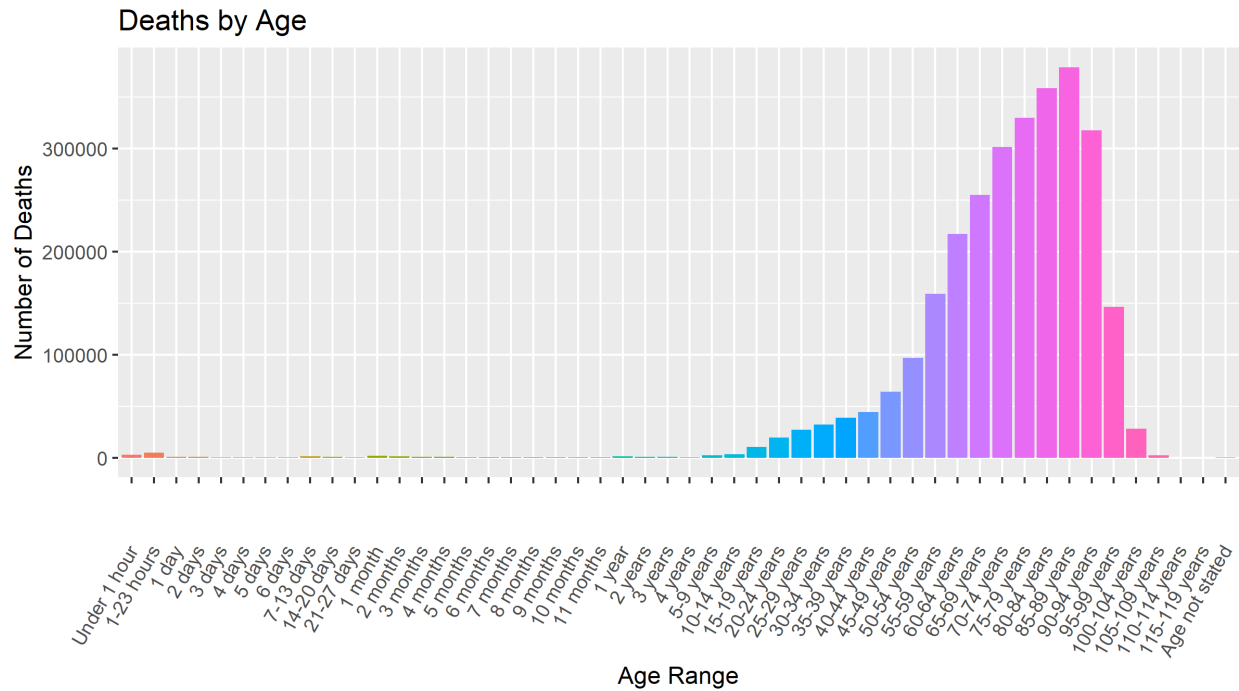
Month	Death Count	Death Count Excluding Natural
April	7860	677.7
August	7349	733.8
December	8277	723.1
February	8333	679.6
January	8331	668.5
July	7415	737.7
June	7534	727.1
March	8239	686.3
May	7659	695.4
November	7986	718.3
October	7683	710.1
September	7442	721.5





Deaths by Age

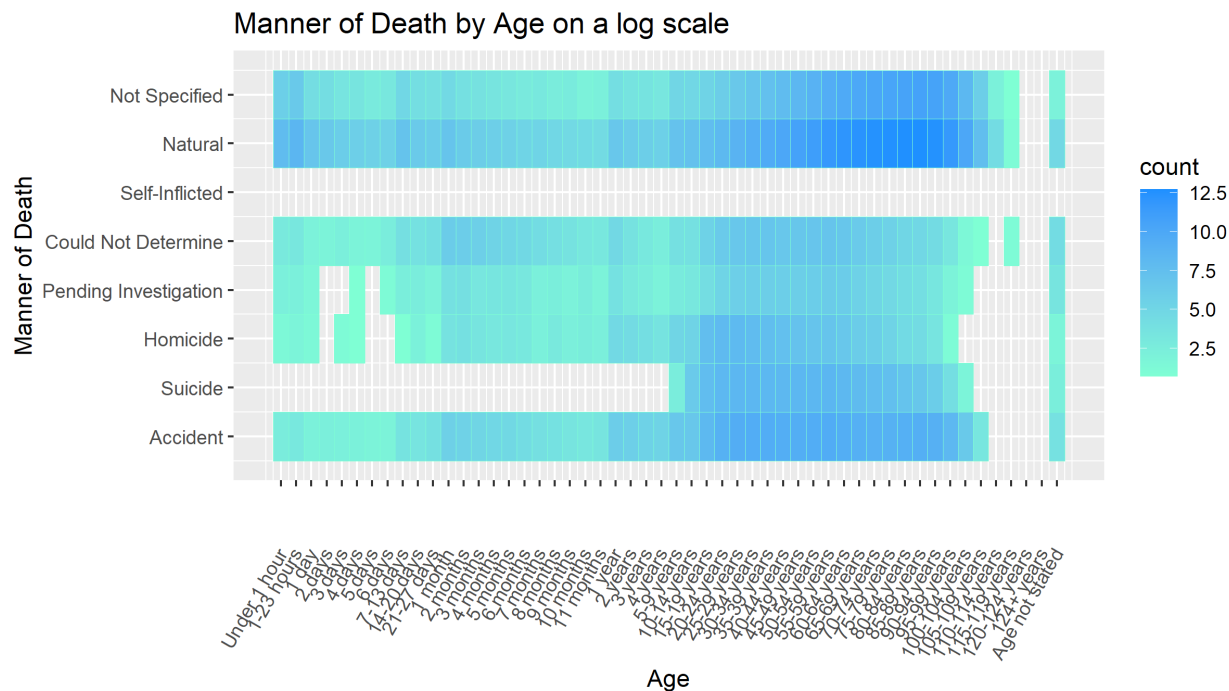
Next, we plotted age versus death counts. Deaths were most prevalent among older age groups such as those between 70 and 84, although deaths start increasing more quickly at age 60. There is also a spike in those less than 1 day old. However, those greater than 1 day old do not frequently die. We also created a version of the plot scaled to bucket size. For privacy reasons, the NBER does not release ages of deaths but rather different buckets that the ages fall into. These bucket are of different lengths of time so we created a rescaled version on a log scale to better showcase the data.



Deaths by Manner

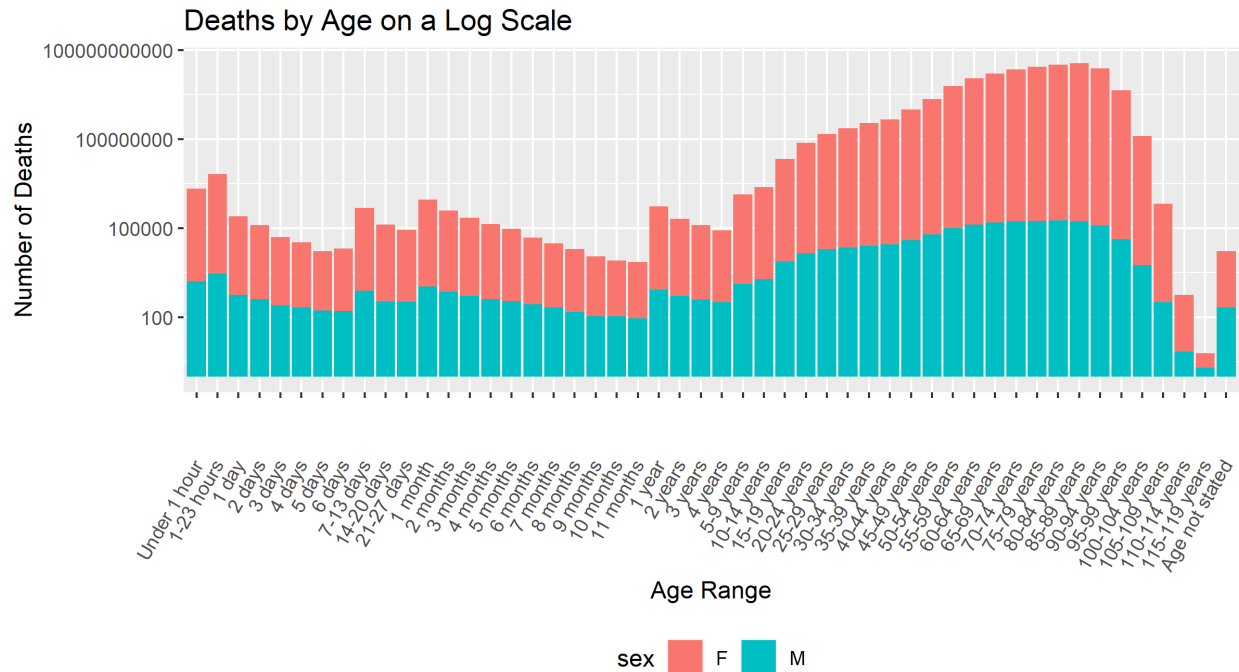
Here, we plotted the manner of death versus age and counted how many people of an aged died by manner of death. The majority of people die from natural causes, especially those aged 60+ and less than 1 day old.

This plot informs us about 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 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.



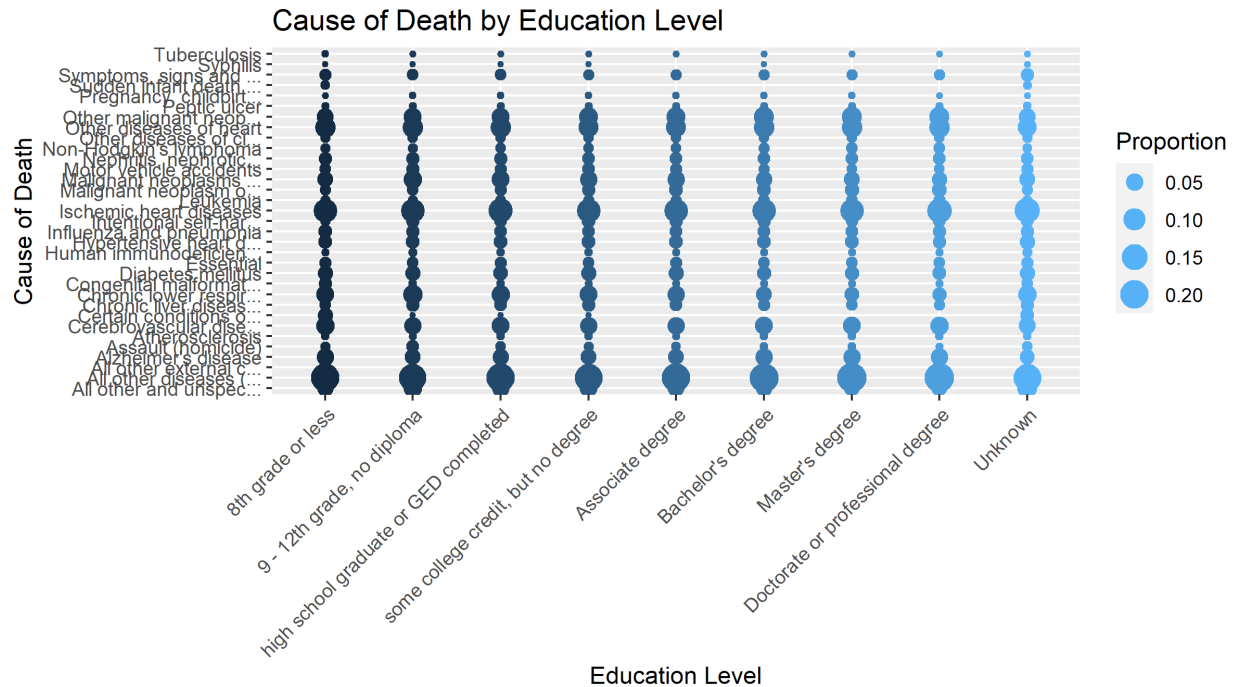
Deaths by Age and Gender

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. This demonstrates 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 has the proportion of male deaths to be above 50%. The proportion of male deaths goes down after 80 years of age, and is because women tend to live longer. What this plot may help to inform us about is the differences in male and females lives and life expectancies. 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

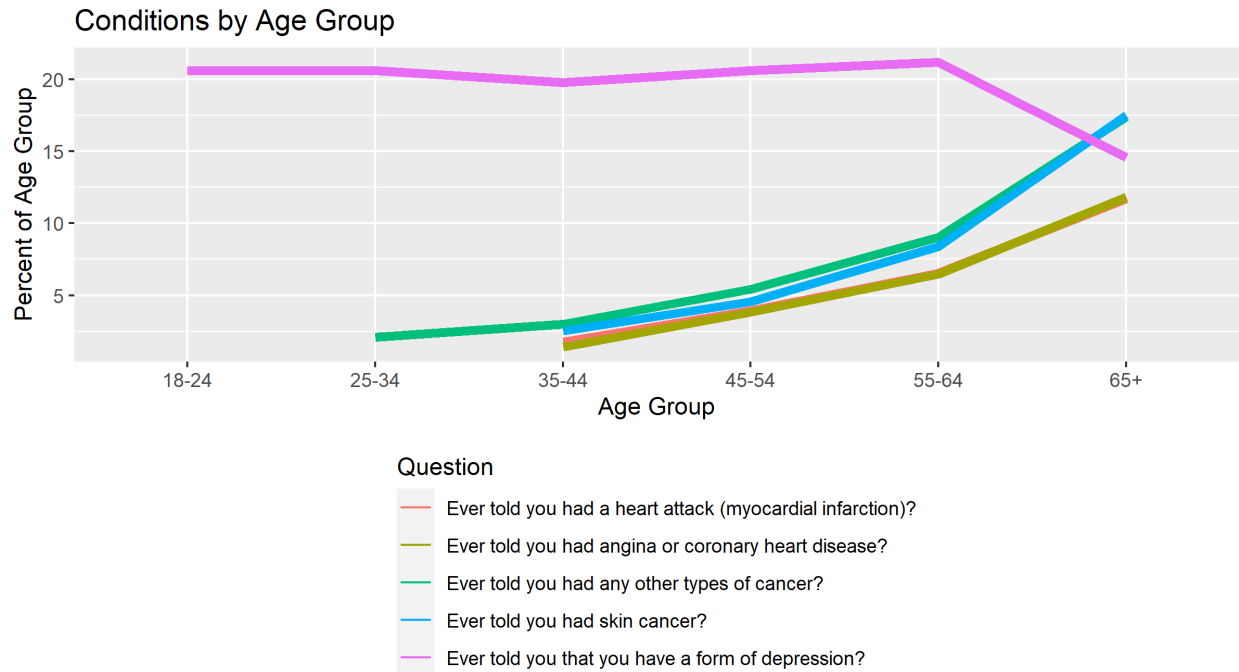
For most causes of death, level of education does not have an impact on what proportion of people have that cause of death. The largest difference belongs to “Certain conditions originating in the perinatal period” with high occurrences in those with 8th grade or less education and those with unknown 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.



Analysis

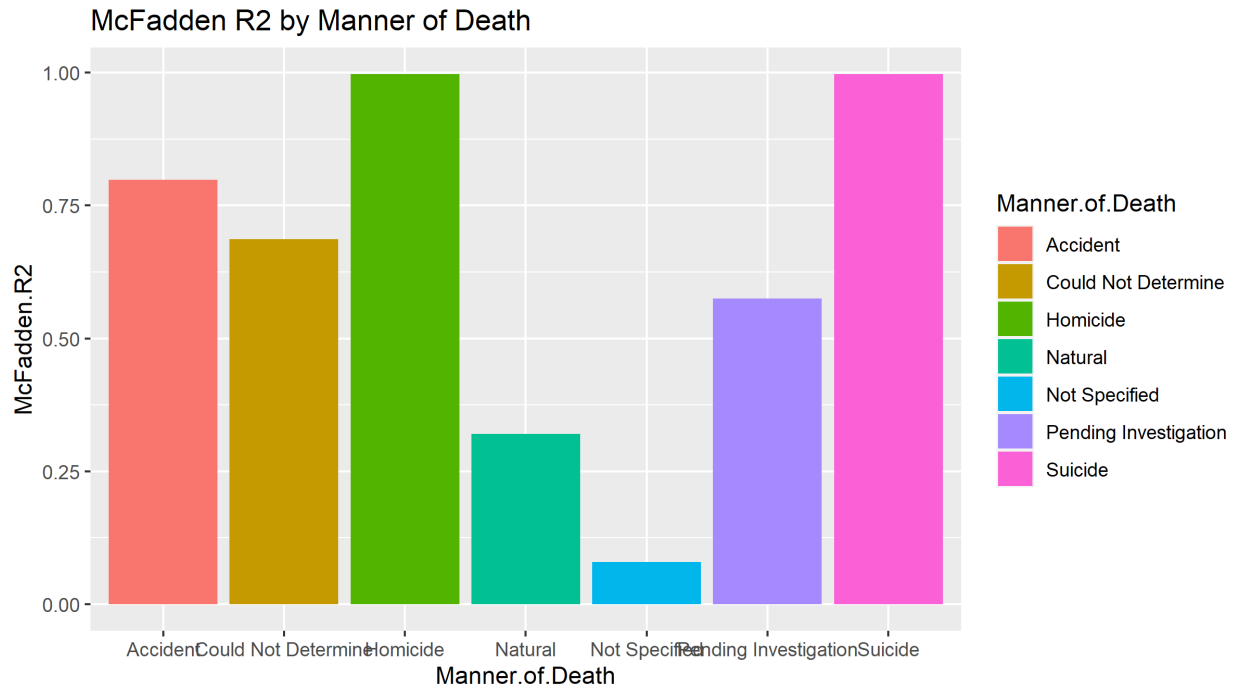
Free Text Analysis: Selected Health Conditions by Age Group

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 are more likely to have been diagnosed with cancer.



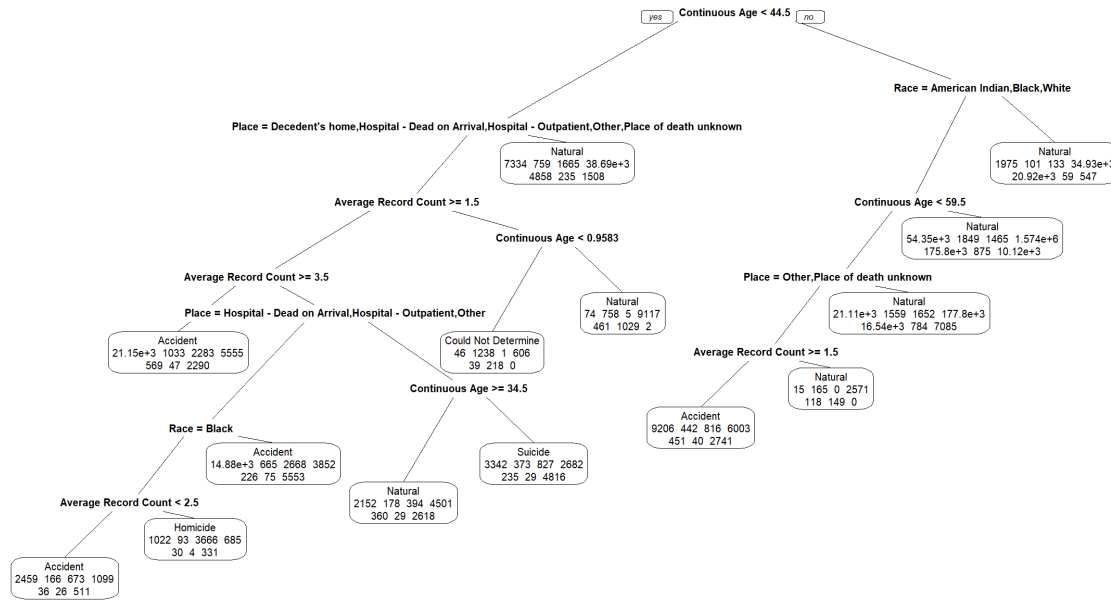
Logistic Classification

The first model we tried to answer the question of “given someone is dead, how did they die?” was logistic classification. We created indicator variables for each of the manners of death before fitting models containing age, record count, sex, education, place of death, cause of death, and race to each of the indicator variables. Below is a plot of the McFadden R^2 values which demonstrates that the logistic classification regression models were much more effective for manners of death like suicide and homicide than for others like not specified and natural.



Decision Tree & Random Forest

We then moved on to a decision tree and random forest classification model. The decision tree predicts with 82.11% accuracy and the random forest predicts with 88.89% accuracy. Each node contains a yes or no classification for either a factor or a logical comparator for a continuous variable. Each leaf contains the classified manner of death and 7 numbers. Each of the 7 numbers corresponds to the count of each manner that was classified to that. The first number is accident, second is suicide, third is homicide, fourth is pending investigation, fifth is could not determine, sixth is natural, and seventh is not specified.



Ultimately, we built a model that can effectively classify a manner of death given someone is dead. Additionally, we derived interesting insights from the data and how deaths are associated with weekdays, marital status, time of year, record counts and more.