# Death in the US

**Group 7**

Mounika Bandam

Ruhui Luan

Yue Jiang

Wejdan Althobaiti

## Abstract

Death is unpredictable. Thousands of people die every year and leave their families and friends heart broken. Every year in the United States, the Centers of Disease Control and Prevention (CDC) releases the country's most detailed report on death. Using this report to analyze death in USA is our goal since understanding any pattern that might be helpful to increase life expectations. After data analysis, we came up with interesting results. The most common manners of death were Natural, not specified, accident, suicide, and homicide respectively. The most percentage of death was in high school graduate who mostly died because of accidents. After analyzing Homicide data, we found that men are more than women, and singles are more than married although that singles represent only 12% in the all data. In addition, white people who represents only 11% in the data, represents 47% of people who died by Homicide. Also, the word cloud illustrates the most common causes of death in our data such as Malignant diseases, Atherosclerosis, and road accidents. After constructing three networks, we found that causes of death and manner of death is too similar in all states since it is one country, so we suggest building a network between different countries which will probably illustrates how countries vary. The race network

shows strong similarity in the manner of death among the earlier settler in the US. It also depicts weaker connection between the latter immigrants or the people who live away from the mainland US and earlier settlers.

# 1. Introduction

Death has never been an easy incident since losing our beloved ones is painful, inevitable, and unpredictable. Everyone dies in the long run. However, there are a lot of possibilities to postpone death**.** Annually in the United States, the Centers of Disease Control and Prevention (CDC) releases the country’s most detailed report on death under the National Vital Statistics Systems. This death dataset includes every person who died in the country for a particular year. In addition, it includes detailed information about each death incident such as causes of death and the demographic background of the deceased. To understand the columns of death dataset, see Appendix 1. It is essential to put the sensitive nature of the topic aside and analyze death dataset since that would help to understand the complex circumstances of death across the United States. Moreover, analyzing this data would define life expectancy for individuals and compare death in the USA to the rest of the world.

# 2. Data

## 2.1 Data Collecting

Finding the dataset was not an easy task. Several datasets have been reviewed by the group until an interesting topic have been found. Our group got the dataset from Kaggle. The death dataset is a collection of spreadsheets in a csv format. Twenty-four csv files are included. Death Records, for example is the main file in which contains an individual death record in each row. Each death record has a one-to-many relationship with the Entity Axis Conditions and Record Axis Conditions tables via a Death Record Id key. Both of these conditions tables contain ICD-10 codes that indicate cause of death for each person. The other csv files are explaining the ambiguous columns in the main csv file. For example, number one in the Manner of Death column means accident.

### 2.2 Data Cleaning

Having that large number of csv files is hard to analyze. Consequently, cleaning the data was a tough process that took us a long time. Following are, the steps which have been done to clean the data:

- Going through twenty-four csv files to extract useful information.
- Going through thirty-eight columns in the main csv file to choose interesting variables.
- Creating a one csv file with fifteen columns to work on. To understand the columns of death dataset, see Appendix 1.
- Converting a variable from one type to another
- Omitting missing values (NAs)
- Detecting outliers (Tukey's box and-whisker method)

The dataset is not network in nature; therefore, we will find the connection among the data.

Following are attributes that will be considered in the project:

- Resident Status
- Education 2003 Revision
- Month of Death
- Sex
- Age
- Marital Status
- Manner of Death
- Place of Injury
- Cause
- Race
- State
- Population

## 3. Data Visualizations

### 3.1 General Data Analysis

- **Manner of Death**

In the data death has 8 manners,

1. Accident
2. Suicide
3. Homicide
4. Pending investigation

5. Could Not determine
6. Self-inflicted
7. Natural
8. Unspecified

The "Natural" death was the most common death manner among all other manners. See the following **figure 3.1.1**. The accident comes following, and there were 120,804 records in the data. The third rank of manner is Suicide, which has 39,023 records and Homicide was 15,356.

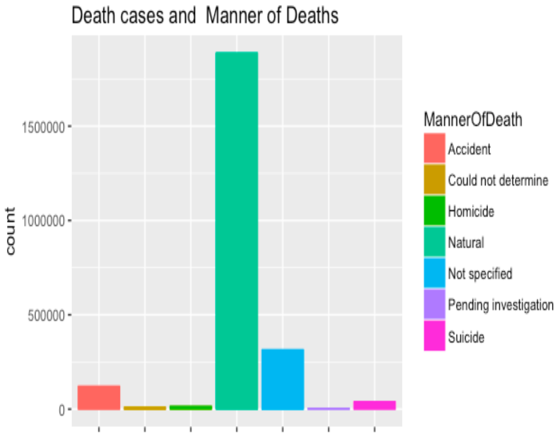


**Figure 3.1.1**

- **Education level**

Knowing the education level of people who died is also important since we can connect that to manners of death and draw useful conclusions. In the data, the most common education level is "high school graduate or GED completed". The second highest education is "some college credit" and "8th grade or less". "Master's degree" is one of the three lowest percentage in the data. See **Figure 3.1.2** the distributions of education level across the manner of death.

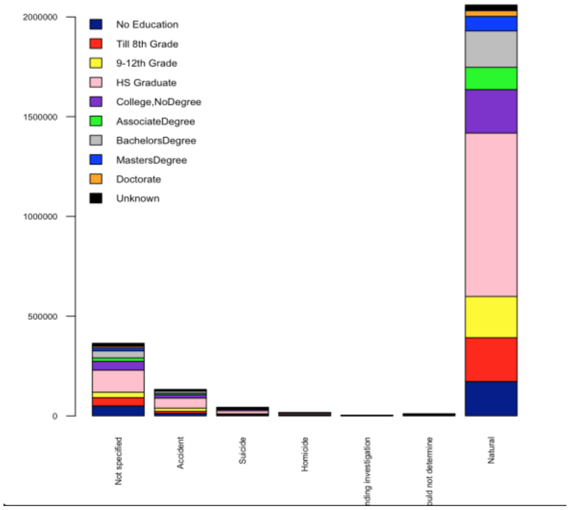
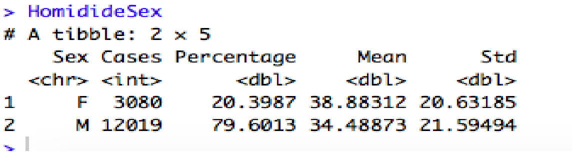


**Figure 3.1.2**

## 3.2 Homicide Vs Race, Sex, and Education level

- **Homicide cases per Sex**

There are 79% of the Homicide cases were men while women represented only 20%. The mean of ages in men is 34 while in women is 38 years old.



```
> HomidideSex
# A tibble: 2 x 5
    Sex Cases Percentage     Mean      Std
  <chr> <int>      <dbl>    <dbl>    <dbl>
1     F  3080    20.3987 38.88312 20.63185
2     M 12019    79.6013 34.48873 21.59494
> 
```

**Figure 3.2.1**

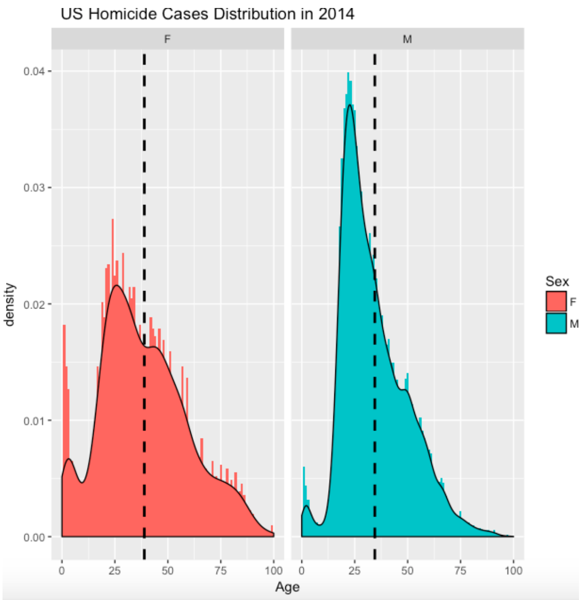
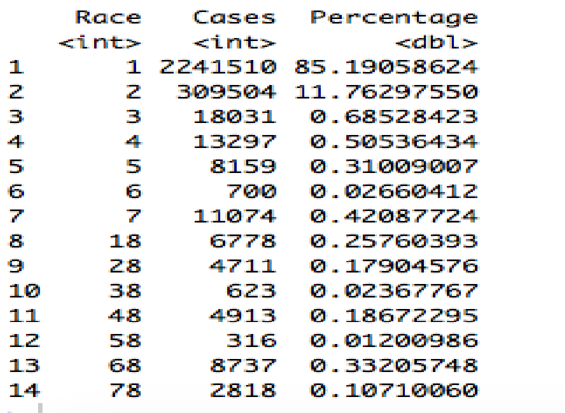


Figure 3.2-2

- **Homicide cases per Race**

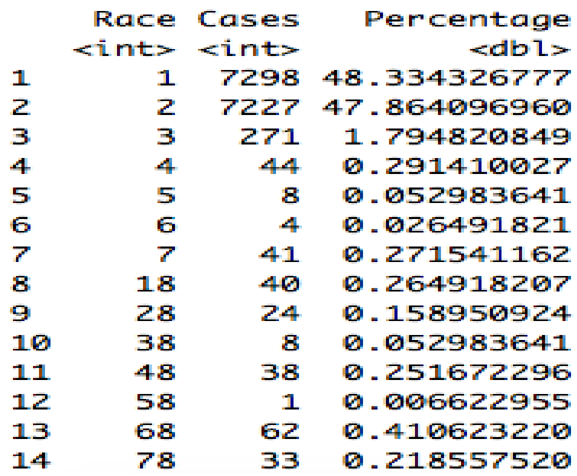
A significant percentage of Homicide cases were black (code 2) (around 47%), although in all the death cases black represents around 11%. White represents 85% of death in all data and only 48% in Homicide manner of death, See the following **tables and Figure 3.2.3**

**Race in All the data**



```
    Race   Cases   Percentage
   <int>   <int>        <dbl>
1      1 2241510  85.19058624
2      2  309504  11.76297550
3      3   18031   0.68528423
4      4   13297   0.50536434
5      5    8159   0.31009007
6      6     700   0.02660412
7      7   11074   0.42087724
8     18    6778   0.25760393
9     28    4711   0.17904576
10    38     623   0.02367767
11    48    4913   0.18672295
12    58     316   0.01200986
13    68    8737   0.33205748
14    78    2818   0.10710060
```

**Race in Homicide Cases only**



```
    Race Cases    Percentage
   <int> <int>         <dbl>
1      1  7298  48.334326777
2      2  7227  47.864096960
3      3   271   1.794820849
4      4    44   0.291410027
5      5     8   0.052983641
6      6     4   0.026491821
7      7    41   0.271541162
8     18    40   0.264918207
9     28    24   0.158950924
10    38     8   0.052983641
11    48    38   0.251672296
12    58     1   0.006622955
13    68    62   0.410623220
14    78    33   0.218557520
```

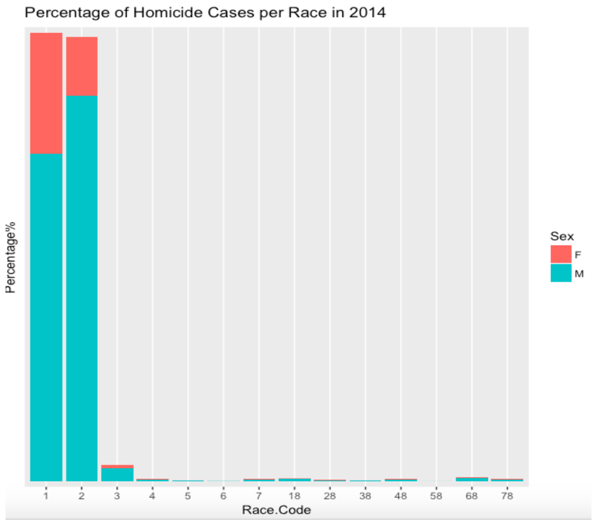


Figure 3.2.3

- **Homicide cases per Race and Education**

Since the Race percent of people, who died because of Homicide, shows that the most common death race groups by Homicide were White, Black, American Indian, and Asian respectively, we filtered the data accordingly to investigate in these races. See **Figure 3.2.4** which illustrates Homicide cases per Race and Education where the education level represented by colored circles. Each Education level has its own color. It can be clearly seen in the graph the blue circles are prevalent which stand for high school graduate. No master or professional degree can be noticed from the graph.

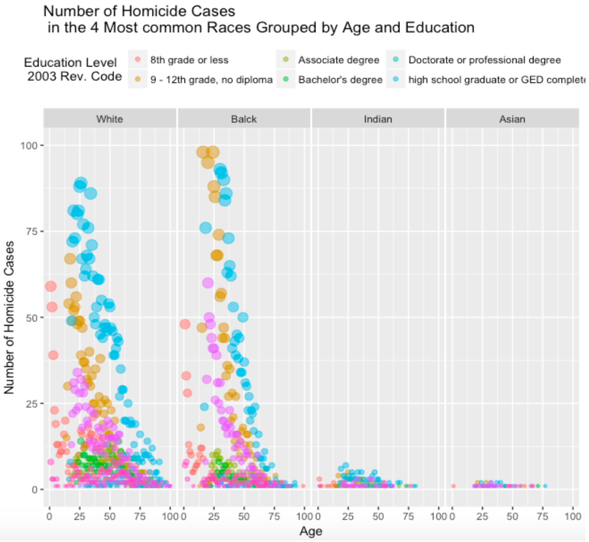
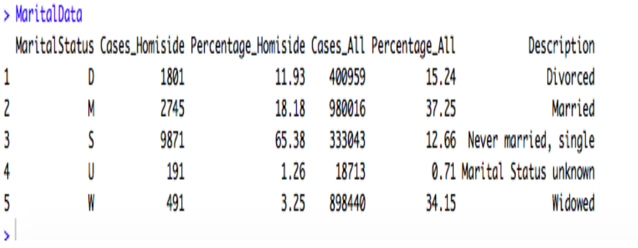


**Figure 3.2.4**

- **Homicide cases per Marital Status and Sex**

It is remarkable that the percentage of single in all data is only 12.66%, it is 65% when it comes to Homicide. On the other side, married people are the larger group in the all data, which compose of 37%, but they represent only 18% when it comes to Homicide. See **Figure 3.2.5**



```
> MaritalData
  MaritalStatus Cases_Homiside Percentage_Homiside Cases_All Percentage_All           Description
1             D           1801               11.93    400959          15.24              Divorced
2             M           2745               18.18    980016          37.25               Married
3             S           9871               65.38    333043          12.66  Never married, single
4             U            191                1.26     18713           0.71 Marital Status unknown
5             W            491                3.25    898440          34.15               Widowed
> 
```

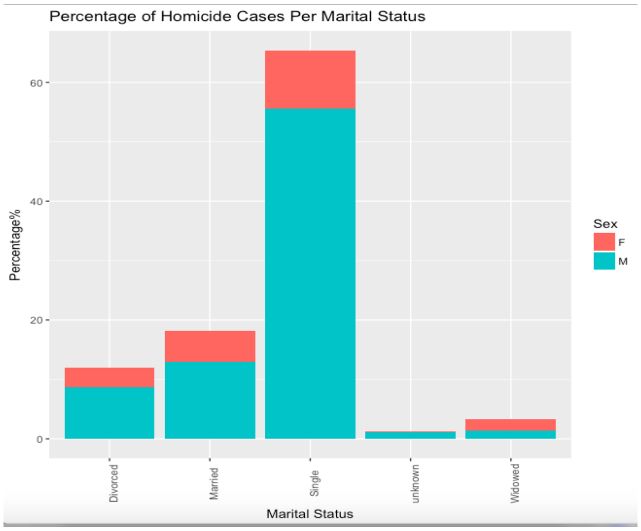


**Figure 3.2.5**

- **Homicide cases per Marital Status and Race**

It can be clearly seen that although the Black are the highest percentage in Homicide, White women who died by Homicide are more than black women. It is also interesting that in all the four common races, the highest percentage in Marital Status are single, married, divorced respectively.

See **Figure 3.2.6**

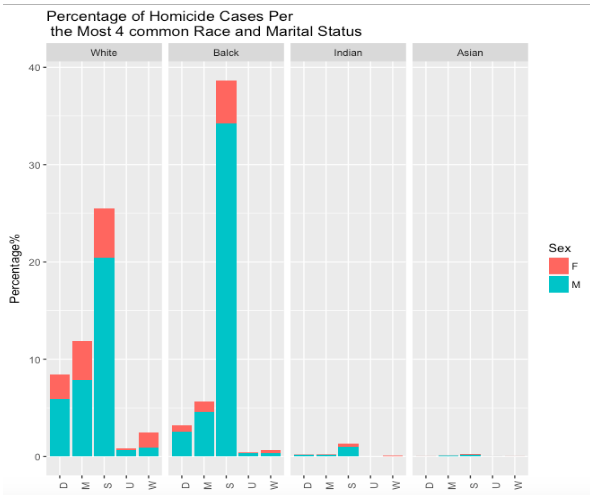
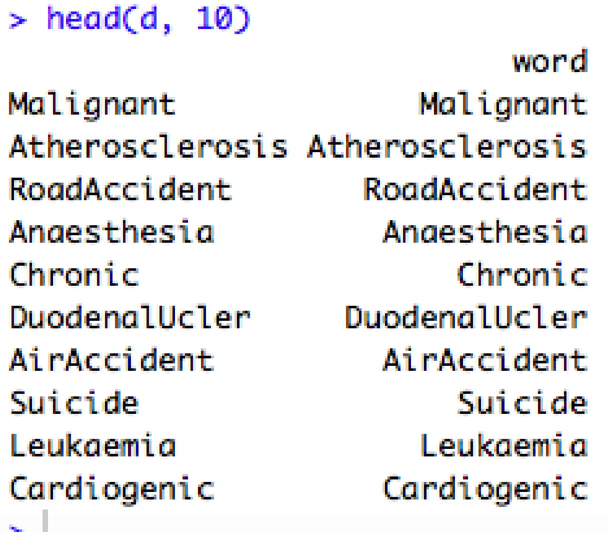


**Figure 3.2.6**

## 3.3 Causes of Death Word Cloud

Text mining for the most common causes of death is one of the essential parts in our project. It is important to understand the most frequent causes for precautions and different science purposes. We used word cloud and tm packages in **Word Cloud on Causes of Death.R** file (for mining) to display which causes occurred frequently and led to death. Using word cloud, helped us to know the causes of death clearly since the colors and sizes of words with the same frequency are identical. The cause “Malignant” diseases is the most frequent word in the death dataset then “Atherosclerosis” which leads to heart attacks, and “Road Accident”. See **Figure 3.3.1** to see the most ten frequent causes of death**,** and a Word Cloud of all causes.



```
> head(d, 10)
                                        word
Malignant                          Malignant
Atherosclerosis              Atherosclerosis
RoadAccident                    RoadAccident
Anaesthesia                      Anaesthesia
Chronic                              Chronic
DuodenalUcler                  DuodenalUcler
AirAccident                      AirAccident
Suicide                              Suicide
Leukaemia                          Leukaemia
Cardiogenic                      Cardiogenic
>
```

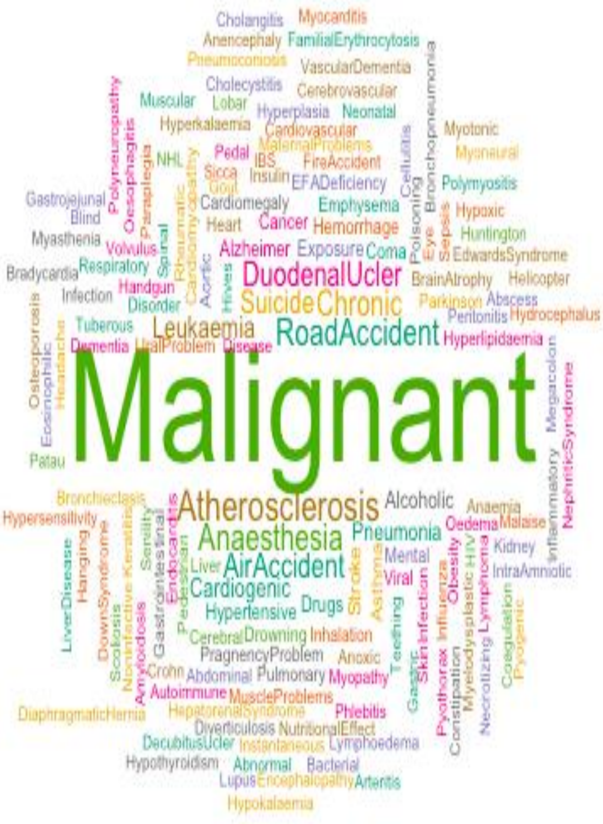


**Figure 3.3.1**

## 3.4 States Networks by Manner of Death & Causes

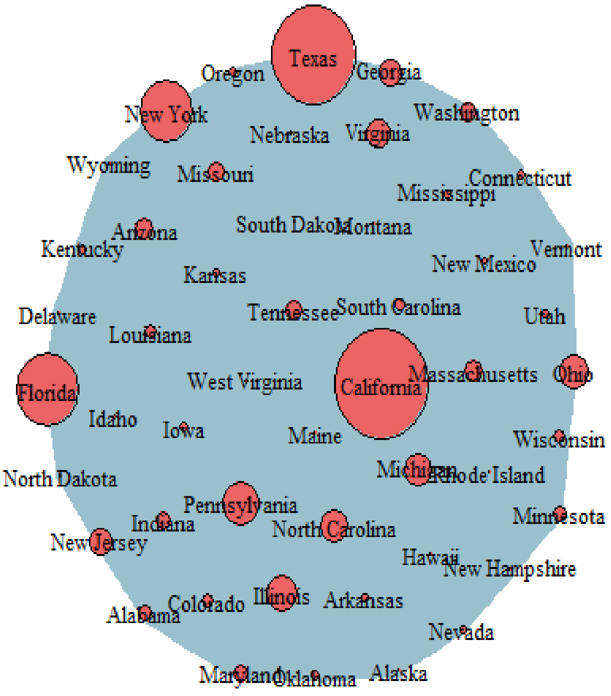


**Figure 3.4.1**

In **Figure 3.4.1** you can see **States Network by Manner of Death**. It can be clearly seen that edges are overlapped since they are weighted. If a Manner of Death has been found in two states, so they should be connected, and if more than one have been found the edge between two states will be thicker. The nodes sizes represent the number of population divided by number of death in each state.

In **Figure 3.4.2** you can see **States Network by Causes**. It is the same case with Manner of death Network. Since almost all states have a lot of common causes between each other, edges cannot be identified. The nodes sizes represent the number of population divided by number of death in each state.

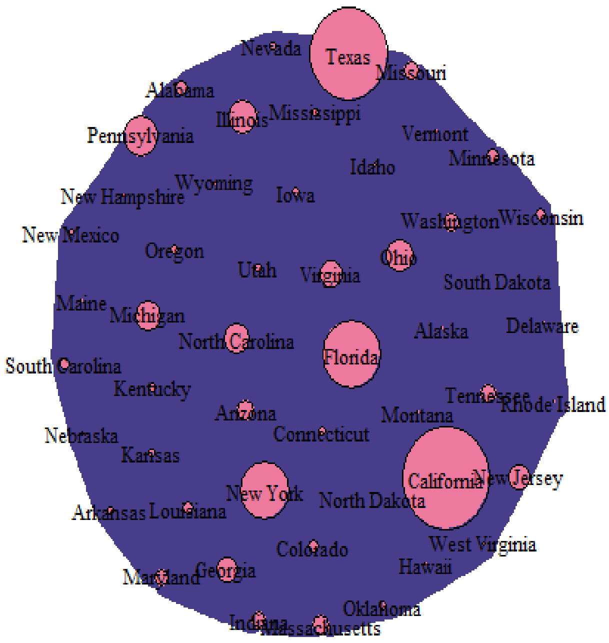


**Figure 3.4.2**

## 3.5 Race Network by Manner of Death

The following network illustrates the Race in our data. Each race is a node that will be connected to another race if the record indicates that a manner of death such as “accident” has been occurred in both races. The nodes sizes reflect the number of dead people in a race. The thickness of edges depicts the number of manner of death that were common between two races. The more common manner they have, the thicker the edges get. See **Figure 3.5.1**

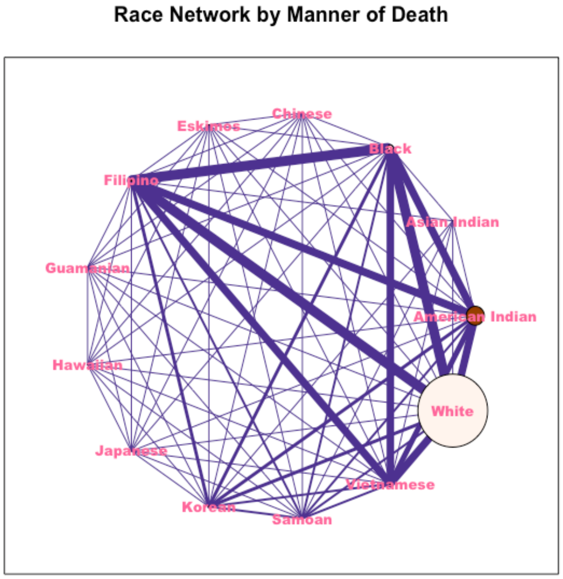


**Figure 3.5.1**

## 4. Conclusion

Based on the general analysis and the social network analysis method we have executed on the data. We have got the following conclusions:

(1) High school graduate was the highest education level in the death records regardless the gender and race. The proportion of the master or professional degrees are rather low. However, whether we can take these result as the evidence that the higher the education level one can get, the lower the rate of death is still need more investigation and data to support.

(2) White represents 85% while black only takes 11% in the all death records. However, white was 48% and black 47% in Homicide.

(3) Although singles represent 12% in all death records, they represent 65% in homicide only.

(4) The most common causes of death in US 2014 were: malignant diseases, heart problems, and road accidents.

(5) Strong similarity of the cause of death and the manner of death in all the 50 states.

(6) Strong similarity of the manner of death among the earlier settler in the United States. Weaker connection between the latter immigrants or the people live away from the mainland US and earlier settlers. Moreover, the races that have closed-culture show

weak linkages with other races.

## 5. Future Developments

First, using our analysis we can spread the awareness across the US to be more careful of the most common manner of deaths such as accident, suicide and homicide. Second, parenthetical companies can use our analysis to look for ideas on the new medicines for the highest ranked diseases that led to death. And further, we can utilize our analysis into the research of the death trends of the whole world, which will help the demographic scientists to have a more through view on the feature of death.

## References

- Kaggle Website - https://www.kaggle.com/datasets
- Analyzing the Social Web, *Golbeck J*, first edition, 2013
- Network visualization with R, *Ognyanova K*
- List of U.S. states by population in 2014 - https://simple.wikipedia.org/wiki/List_of_U.S._states_by_population

## Appendix 1

### Columns of Death Records Dataset

Primary table containing a single row per death record with the below columns,

- Id (*integer primary key*) - Main identifier, used for joining with Death Record Id in Entity Axis Conditions and Record Axis Conditions tables.
- ResidentStatus (*integer*) - (e.g. 1 = Residents, 2 = Intrastate resident and so on)
- Education2003Revision (*integer*) - Years of education using the 2003 revision code (e.g. 8 = Doctorate or professional degree)
- MonthOfDeath (*integer*) - Month of death (e.g. 1 = January, 12 = December)
- Sex (*text*) - (M = Male, F = Female)

- MaritalStatus (*text*) - (e.g. M = married, D = divorced, W = widowed)
- MannerOfDeath (*integer*) - (e.g. 1 = Accident, 2 = Suicides)
- ActivityCode (*integer*) - (e.g. 0 = While engaged in sports activity, 1 = While engaged in leisure activity)
- PlaceOfInjury (*integer*) - (e.g. 0 = Home, 1 = Residential institution)
- Causes (*text*) - ICD-10 code for the underlying cause of death (e.g. I251 = Atherosclerotic heart disease)
- Race (*integer*) - Reported race (e.g. 1 = White, 2 = Black)
- State(*text)* – All the fifty USA states.
- Population(*integer)-* Reported population for each state.

## Appendix 2

**Race Code**

**Code – Description**

0 - Other races

1 - White

2 - Black

3 - American Indian (includes Aleuts and Eskimos)

4 - Chinese

5 - Japanese

6 - Hawaiian (includes Part-Hawaiian)

7 - Filipino

8 - Other Asian or Pacific Islander

18 - Asian Indian

28 - Korean

38 - Samoan

48 - Vietnamese

58 - Guamanian

68 - Other Asian or Pacific Islander in areas reporting codes 18-58

78 – Combined other Asian or Pacific Islander, includes codes 18-68 for areas that do not report them separately