Midterm - Leading Causes of Death: United States

Jiayi Nie

10/19/2021

## Intro

The dataset used for this project presents the age-adjusted death rates for the 10 leading causes of death in the United States beginning in 1999 till 2017. Causes of death classified by the International Classification of Diseases. Cause of death statistics are based on the underlying cause of death. Access from CDC (<https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu>)

Since it’s hard to demonstrate all 51 states clearly, introduce United States regions dataset which contains all 51 states and groups them into 5 regions according to their geographic position on the continent: the Northeast, Southwest, West, Southeast, and Midwest. (ref: <https://www.nationalgeographic.org/maps/united-states-regions/>) (data source:<https://www.kaggle.com/omer2040/usa-states-to-region>)

Also use US state location dataset in this project. (data resource: <https://www.kaggle.com/washimahmed/usa-latlong-for-state-abbreviations>)

In this project we will discuss and explore the leading death cause in United State and how doese it change as year past. The purpose of this study is try to figure out the death rate change pattern for the whole country and assess what cause of death should we pay more attention to. In other word, how did the leading cause of death nowadays (in 2017) different from 18 years ago (in 1999) in each region. The research questions and hypothesis for this project to examine and analyze are:

* Examine whether the death rate of all-cause death rate follow decrease trend in the whole United State and each state. If there is exception, state them.
* Illustrate the leading cause of death with highest death number in 2017 and 1999 respectively, state how the leading death cause changes in following aspects:
  + What’s the most common death cause in 1999 among whole US, how about 2017?
  + What’s the region with highest death number in 2017?

Then from time, region(state) and cause of death three aspect draw our conclusion.

## Method

The dataset resource are included in the intro part. In this part we will use EDA checklist to clean and wrangle our data. Primarily used ggplot and leaflet do the data visualization to explore data.

### Data loading

if (!file.exists("bi63-dtpu.csv"))  
 download.file(  
 url = "https://data.cdc.gov/api/views/bi63-dtpu/rows.csv",  
 destfile = "bi63-dtpu.csv",  
 method = "libcurl",  
 timeout = 60  
 )  
data <- data.table::fread("bi63-dtpu.csv")  
  
region <- data.table::fread("USAregion2.csv")  
  
location <- data.table::fread("USAlocation.csv")

### Data wrangling and cleaning

Checking the dimension, headers and footers of the data

knitr::kable(dim(data))

|  |
| --- |
| x |
| 10868 |
| 6 |

knitr::kable(head(data))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | 113 Cause Name | Cause Name | State | Deaths | Age-adjusted Death Rate |
| 2017 | Accidents (unintentional injuries) (V01-X59,Y85-Y86) | Unintentional injuries | United States | 169936 | 49.4 |
| 2017 | Accidents (unintentional injuries) (V01-X59,Y85-Y86) | Unintentional injuries | Alabama | 2703 | 53.8 |
| 2017 | Accidents (unintentional injuries) (V01-X59,Y85-Y86) | Unintentional injuries | Alaska | 436 | 63.7 |
| 2017 | Accidents (unintentional injuries) (V01-X59,Y85-Y86) | Unintentional injuries | Arizona | 4184 | 56.2 |
| 2017 | Accidents (unintentional injuries) (V01-X59,Y85-Y86) | Unintentional injuries | Arkansas | 1625 | 51.8 |
| 2017 | Accidents (unintentional injuries) (V01-X59,Y85-Y86) | Unintentional injuries | California | 13840 | 33.2 |

knitr::kable(tail(data))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | 113 Cause Name | Cause Name | State | Deaths | Age-adjusted Death Rate |
| 1999 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney disease | Vermont | 56 | 9.2 |
| 1999 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney disease | Virginia | 1035 | 16.9 |
| 1999 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney disease | Washington | 278 | 5.2 |
| 1999 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney disease | West Virginia | 345 | 16.4 |
| 1999 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney disease | Wisconsin | 677 | 11.9 |
| 1999 | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Kidney disease | Wyoming | 30 | 6.8 |

It includes 10.9K rows and 6 column:column information:Year, 113 cause name, Cause name, State, Deaths, Age-adjusted death rate

Check the variable type of the data

knitr::kable((str(data)))

Classes ‘data.table’ and ‘data.frame’: 10868 obs. of 6 variables: $ Year : int 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 … $ 113 Cause Name : chr “Accidents (unintentional injuries) (V01-X59,Y85-Y86)” “Accidents (unintentional injuries) (V01-X59,Y85-Y86)” “Accidents (unintentional injuries) (V01-X59,Y85-Y86)” “Accidents (unintentional injuries) (V01-X59,Y85-Y86)” … $ Cause Name : chr “Unintentional injuries” “Unintentional injuries” “Unintentional injuries” “Unintentional injuries” … $ State : chr “United States” “Alabama” “Alaska” “Arizona” … $ Deaths : int 169936 2703 436 4184 1625 13840 3037 2078 608 427 … $ Age-adjusted Death Rate: num 49.4 53.8 63.7 56.2 51.8 33.2 53.6 53.2 61.9 61 … - attr(\*, “.internal.selfref”)=

|| || || ||

Check how many category in key colunmns (Cause name and State)

cate1 <- unique(data$`Cause Name`)  
knitr::kables(list(as.array(cate1),length(cate1)))

Unintentional injuries All causes Alzheimer’s disease Stroke CLRD Diabetes Heart disease Influenza and pneumonia Suicide Cancer Kidney disease

11

cate2<- unique(data$State)  
knitr::kables(list(as.data.frame(cate2),length(cate2)))

c(“United States”, “Alabama”, “Alaska”, “Arizona”, “Arkansas”, “California”, “Colorado”, “Connecticut”, “Delaware”, “District of Columbia”, “Florida”, “Georgia”, “Hawaii”, “Idaho”, “Illinois”, “Indiana”, “Iowa”, “Kansas”, “Kentucky”, “Louisiana”, “Maine”, “Maryland”, “Massachusetts”, “Michigan”, “Minnesota”, “Mississippi”, “Missouri”, “Montana”, “Nebraska”, “Nevada”, “New Hampshire”, “New Jersey”, “New Mexico”, “New York”, “North Carolina”, “North Dakota”, “Ohio”, “Oklahoma”, “Oregon”, “Pennsylvania”, “Rhode Island”, “South Carolina”, “South Dakota”, “Tennessee”, “Texas”, “Utah”, “Vermont”, “Virginia”, “Washington”, “West Virginia”, “Wisconsin”, “Wyoming”)

52 There are 10 causes of death (1 extra is “All-cause”) and 51 states (including District of Columbia) and the whole US included in this dataset.

Check NA

d1 = summary(is.na(data))  
d2 = summary(data)  
knitr::kable(rbind(d1,d2))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Year | 113 Cause Name | Cause Name | State | Deaths | Age-adjusted Death Rate |
|  | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical |
|  | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10868 |
|  | Min. :1999 | Length:10868 | Length:10868 | Length:10868 | Min. : 21 | Min. : 2.6 |
|  | 1st Qu.:2003 | Class :character | Class :character | Class :character | 1st Qu.: 612 | 1st Qu.: 19.2 |
|  | Median :2008 | Mode :character | Mode :character | Mode :character | Median : 1718 | Median : 35.9 |
|  | Mean :2008 | NA | NA | NA | Mean : 15460 | Mean : 127.6 |
|  | 3rd Qu.:2013 | NA | NA | NA | 3rd Qu.: 5756 | 3rd Qu.: 151.7 |
|  | Max. :2017 | NA | NA | NA | Max. :2813503 | Max. :1087.3 |

There is no NA. Next create a new variable with state region category and location information

d\_region<-merge(  
 x=data,  
 y=region,  
 by.x= "State",  
 by.y = "State",  
 all.x = TRUE,  
 all.y = FALSE  
)  
  
alldata <- merge(  
 x=d\_region,  
 y=location,  
 by.x = "State",  
 by.y = "City",  
 all.x=TRUE,  
 all.y=FALSE  
)

We already know that there is a “United States” category in “State” in our original data, check if there is 209 NA in our new colunm to test the merge result.

knitr::kable(summary(is.na(alldata)))

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | State | Year | 113 Cause Name | Cause Name | Deaths | Age-adjusted Death Rate | State Code | Region | Division | State.y | Latitude | Longitude |
|  | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical |
|  | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10868 | FALSE:10659 | FALSE:10659 | FALSE:10659 | FALSE:10659 | FALSE:10659 | FALSE:10659 |
|  | NA | NA | NA | NA | NA | NA | TRUE :209 | TRUE :209 | TRUE :209 | TRUE :209 | TRUE :209 | TRUE :209 |

## Data exploration and Preliminary Results

### Figure out “what’s the death rate change pattern in the whole US” from time duration, region and causes three aspects

### Let’s go time duration first, illustrate the all-causes death rate change pattern in the US.

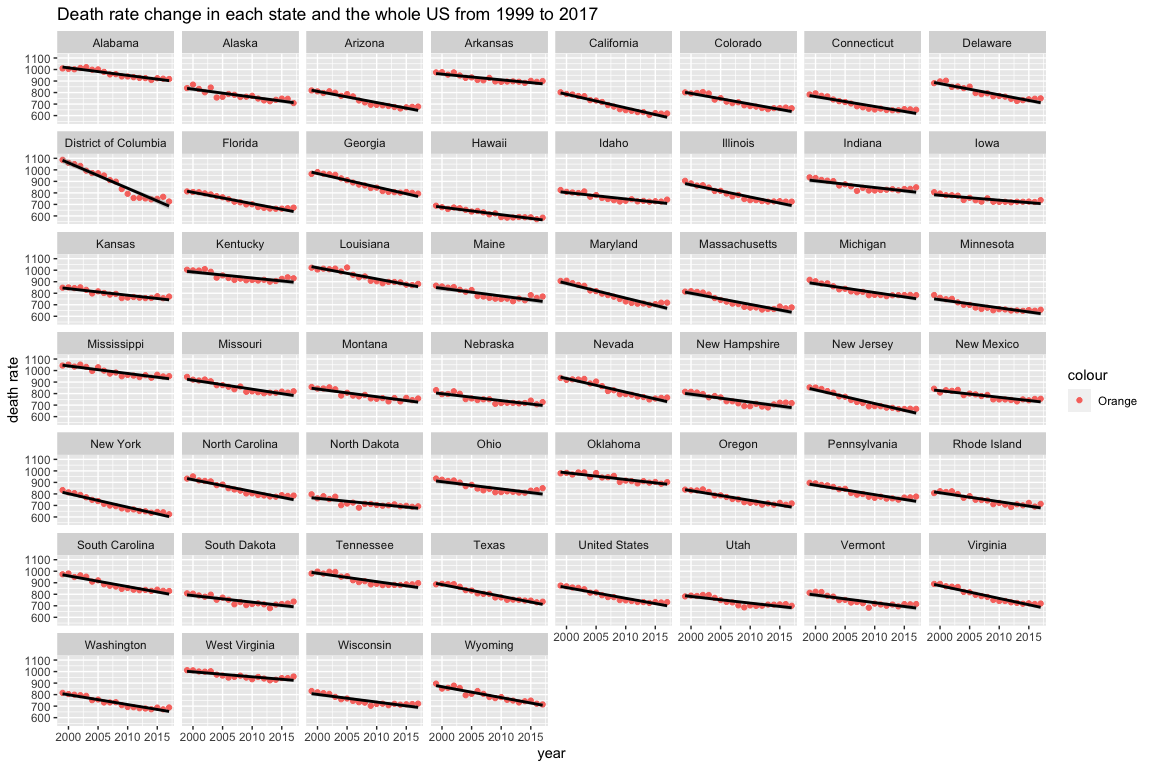
First extra all causes data and create a new dataset

allcause<- alldata %>%  
 filter(alldata$`Cause Name` == "All causes")

Then use ggplot to draw line chart (year vs death rate) and grouped by states, should have 52 lines in total

ggplot(allcause,mapping = aes(x=allcause$`Year`, y = allcause$`Age-adjusted Death`,color = "Orange")) +  
 geom\_point()+  
 geom\_smooth(method="lm", color = "Black")+  
 facet\_wrap(~State)+  
 labs(title = "Death rate change in each state and the whole US from 1999 to 2017",x="year", y ="death rate", size = 8)

## `geom\_smooth()` using formula 'y ~ x'



This graph indicates all 51 states and the whole United states death rate change during 18 years. According to this graph, all 51 states follow a downward trend which meet our hypothesis.

### Next from the region prospective, use the location information to make a leaflet that indicates the leading cause of death with highest death number in 2017 and 1999 respectively

#### First look back to 1999

data1999<- alldata %>%  
 group\_by(State) %>%  
 filter(Year ==1999, `Cause Name` != "All causes") %>%  
 filter(Deaths == max(Deaths))

pal <- colorFactor(palette = "RdYlBu",domain=data1999$Deaths)  
  
leaflet(data1999) %>%   
 addProviderTiles('CartoDB.Positron') %>%  
 addCircles(  
 lat = ~Latitude, lng=~Longitude,  
 color = ~ pal(data1999$Deaths),  
 label = ~paste0(Deaths, ' Leading cause: ', `Cause Name` ),  
 opacity =1, fillOpacity=1, radius=600  
 )

Take more closer look at 1999 year data

knitr::kable(table(data1999$`Cause Name`))

|  |  |
| --- | --- |
| Var1 | Freq |
| Cancer | 1 |
| Heart disease | 51 |

region1999<- data1999 %>%  
 group\_by(Region) %>%  
 summarise(sum(Deaths))%>%  
 replace\_na(list(Region = "US total"))  
knitr::kable(region1999)

|  |  |
| --- | --- |
| Region | sum(Deaths) |
| Midwest | 176295 |
| Northeast | 171784 |
| Southeast | 195472 |
| Southwest | 68933 |
| West | 112778 |
| US total | 725192 |

Now we know that in 1999, the most risk death cause is **heart disease** with 50 states had this leading cause. And the region with highest death number is **Southeast**, which include Alabama, Arkansas, District of Columbia, Delaware, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia West Virginia

#### Now we pay more attention on most recent record in 2017.

data2017<- alldata %>%  
 group\_by(State) %>%  
 filter(Year ==2017, `Cause Name` != "All causes") %>%  
 filter(Deaths == max(Deaths))

pal2 <- colorFactor(c('cyan3','goldenrod2'),domain=data2017$Deaths)  
  
leaflet(data2017) %>%   
 addProviderTiles('CartoDB.Positron') %>%  
 addCircles(  
 lat = ~Latitude, lng=~Longitude,  
 color = ~ pal2(data2017$Deaths),  
 label = ~paste0(Deaths, ' Leading cause: ', `Cause Name` ),  
 opacity =1, fillOpacity=1, radius=600  
 )

Take closer look at 2017 data

knitr::kable(table(data2017$`Cause Name`))

|  |  |
| --- | --- |
| Var1 | Freq |
| Cancer | 13 |
| Heart disease | 39 |

region2017<- data2017 %>%  
 group\_by(Region) %>%  
 summarise(sum(Deaths))%>%  
 replace\_na(list(Region = "US total"))  
knitr::kable(region2017)

|  |  |
| --- | --- |
| Region | sum(Deaths) |
| Midwest | 149135 |
| Northeast | 136893 |
| Southeast | 184949 |
| Southwest | 72412 |
| West | 111289 |
| US total | 647457 |

Now we know that in 2017, the most risk death cause is still **heart disease**. And the region with highest death number is still **Southeast**, which include Alabama,

### Brief Conclusion

Now we could draw a preliminary conclusion to our question:

* It’s apparently that all death rate went down in the past 18 years in all 51 states
* Compared to 1999, the death number decrease in 2017. However, there is an exception that the death number in region **southwest** increase from 68933 to 72412. Possible reason might be the increase population in this region. Could take closer look combined with some external information next step.
* From the geographic aspect, the pattern did not change. The order of death number in these five region is still **Southeast > Midwest > Northwest > West > Southwest** It’s worth noting that, southwest has significant low number compared to other 4. Could combined with population data and state death rate data to check whether the low number comes from low population or low disease incidence. (we don’t know the region population, so could not sum all death rate up for now)
* From the disease cause aspect, in 1999, the most fatal death cause is heart disease which becoming leading cause in 50 states. In 2017, Cancer catchs up a little. This might indicate that nowadays cancer should pay more attention to.