USA Leading Causes of Death

2023-03-11

Description Of Project:

The leading sources of death in the USA by racial and gender since 2007. The cause of death is established using the USA death certificate, which is produced for each death that occurs in USA.

The most recent instance was on 9/24/2019. The Bureau of Vital Statistics and the USA Department of Health and Mental Hygiene have repressed rates based on tiny numbers (RSE $>$ 30) and aggregate counts under 5.

Data Source Link

https://catalog.data.gov/dataset/new-york-city-leading-causes-of-death/resource/845a3736-6ce4-46da-a82d-8a5f9add81f1

Project Purpose

By examining the aforementioned information to determine the trend of Race, Ethnicity, Death Rate, and Gender of each year, we will be able to extract the useful information from the dataset.

Project Results

1) To calculate the input statistics.
2) Determination of linear regression for Year, Race, Nationality, Death Rate, and Gender.
3) The development of linear regression for Year, Race, Ethnicity, Mortality, and Gender.
4) Making a data visualization for the variables Year, Race, Ethnicity, Death Rate, and Gender.
5) Examining the year, race, ethnicity, death rate, and gender data groups.

Packages And Libraries

```
install.packages("tidyr", repos = "http://cran.us.r-project.org")
install.packages("factoextra", repos="http://cran.us.r-project.org")
install.packages("dplyr", repos = "http://cran.us.r-project.org")
install.packages("knitr", repos = "http://cran.us.r-project.org")
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
library(factoextra)
library(tidyr)
library(dplyr)
library(magrittr)
library(lubridate)
library(stringr)
library(knitr)
library(ggplot2)
```

Code for USA Leading Causes of Death

```
# Reading the data
main_df =  read.csv("C:/Users/Public/USA_Leading_Causes_of_Death.csv")

# Making all headings lowercase and Renaming the columns spaces with "_".

main_df = main_df %>%select_all(~gsub("\\s+|\\.", "_", .)) %>% select_all(tolower)

# Structure before transformation
str(main_df)
```

```
## 'data.frame':    1272 obs. of  7 variables:
##  $ year                   : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ leading_cause          : chr  "Diseases of Heart (I00-I09, I11, I13, I20-I51)" "Malignant Neoplası
##  $ sex                    : chr  "Male" "Male" "Male" "Male" ...
##  $ race_ethnicity         : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
##  $ deaths                 : chr  "1603" "1164" "423" "245" ...
##  $ death_rate             : chr  "136.8029917" "99.3379179" "36.09960418" "20.9087542" ...
##  $ age_adjusted_death_rate: chr  "176.783287" "121.5817693" "35.70789583" "25.40934387" ...
```

```
# Transforming the column values

main_df$sex[main_df$sex == "M"] = "Male"
main_df$sex[main_df$sex == "F"] = "Female"

# Data Casting the values with numeric data type
main_df$deaths = as.numeric(main_df$deaths)
```

```
## Warning: NAs introduced by coercion
```

```
main_df$death_rate = as.numeric(main_df$death_rate)
```

```
## Warning: NAs introduced by coercion
```

```r
main_df$age_adjusted_death_rate = as.numeric(main_df$age_adjusted_death_rate)
```

```
## Warning: NAs introduced by coercion
```

```r
# Structure after transformation
str(main_df)
```

```
## 'data.frame':    1272 obs. of  7 variables:
##  $ year                   : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ leading_cause          : chr  "Diseases of Heart (I00-I09, I11, I13, I20-I51)" "Malignant Neoplasr
##  $ sex                    : chr  "Male" "Male" "Male" "Male" ...
##  $ race_ethnicity         : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
##  $ deaths                 : num  1603 1164 423 245 182 ...
##  $ death_rate             : num  136.8 99.3 36.1 20.9 15.5 ...
##  $ age_adjusted_death_rate: num  176.8 121.6 35.7 25.4 19.9 ...
```

```r
# Removing the invalid records from the data frame
main_df_na = na.omit(main_df)

# Statistics of the core data frame
kable(summary(select(main_df_na,year,deaths,death_rate,age_adjusted_death_rate)),row.names = FALSE,capt
```

Table 1: USA Cases of Death Statistics

year	deaths	death_rate	age_adjusted_death_rate
Min. :2007	Min. : 5.0	Min. : 2.40	Min. : 2.50
1st Qu.:2009	1st Qu.: 102.0	1st Qu.: 11.95	1st Qu.: 12.00
Median :2011	Median : 207.0	Median : 18.50	Median : 20.00
Mean :2012	Mean : 577.3	Mean : 53.52	Mean : 53.21
3rd Qu.:2013	3rd Qu.: 472.5	3rd Qu.: 66.07	3rd Qu.: 77.90
Max. :2019	Max. :7050.0	Max. :491.40	Max. :414.59

```r
# Grouping the required columns
df_grpby = group_by(main_df_na,leading_cause, year, sex)

det_tot = main_df_na %>% group_by(year) %>%
  summarise(Totadeaths=sum(deaths),.groups = 'drop') %>% as.data.frame()

# Linear regression of deaths VS years
summary(lm(deaths~year, data = main_df_na))
```

```
##
## Call:
## lm(formula = deaths ~ year, data = main_df_na)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -602.6 -469.3 -370.1  -98.8 6434.4
##
## Coefficients:
```

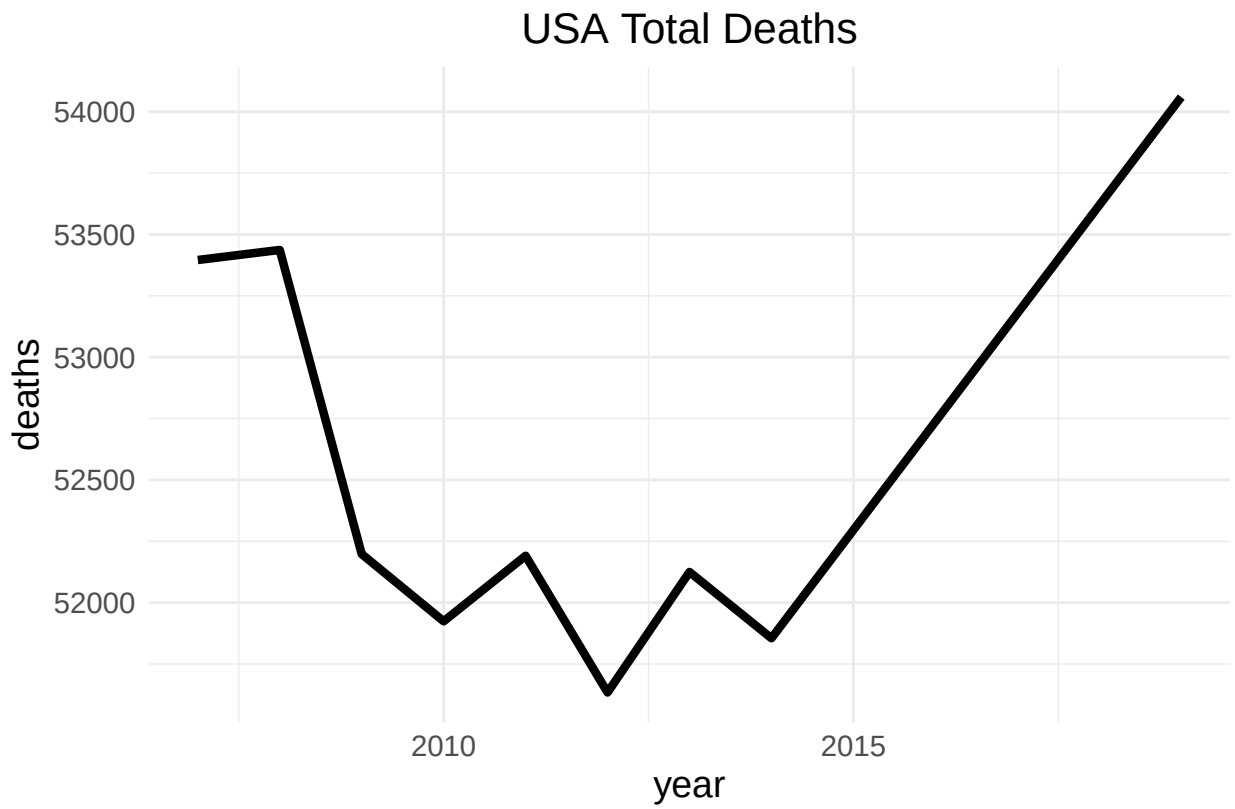
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17183.345  18651.411   0.921    0.357
## year           -8.255      9.272  -0.890    0.374
##
## Residual standard error: 958.5 on 817 degrees of freedom
## Multiple R-squared:  0.0009693,  Adjusted R-squared:  -0.0002535
## F-statistic: 0.7927 on 1 and 817 DF,  p-value: 0.3735
```

```
# Polynomial regression of deaths VS years
summary(lm(deaths ~ poly(year, 2, raw = TRUE),data = main_df_na))
```

```
##
## Call:
## lm(formula = deaths ~ poly(year, 2, raw = TRUE), data = main_df_na)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -583.3 -470.4 -371.6  -94.3 6462.7
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -5.286e+06  9.549e+06  -0.554    0.580
## poly(year, 2, raw = TRUE)1  5.261e+03  9.487e+03   0.555    0.579
## poly(year, 2, raw = TRUE)2 -1.309e+00  2.356e+00  -0.555    0.579
##
## Residual standard error: 958.9 on 816 degrees of freedom
## Multiple R-squared:  0.001347,   Adjusted R-squared:  -0.001101
## F-statistic: 0.5502 on 2 and 816 DF,  p-value: 0.577
```

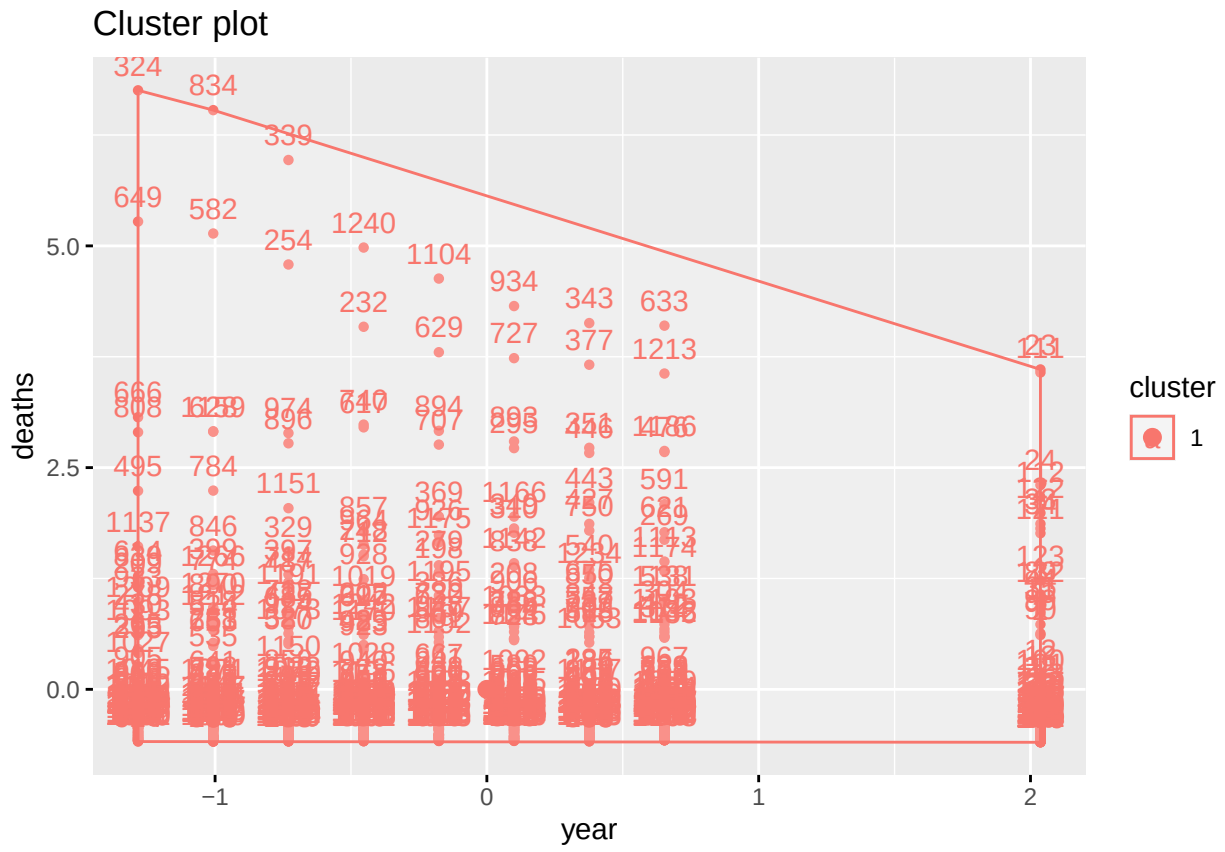
```
# Plotting the graph for deaths VS years
ggplot(det_tot,aes(x=year,y=Totadeaths)) +geom_line(size=1.5) +
    labs(x = "year", y = "deaths",caption="USA Death Cases From 2007 until 2019") +
    ggtitle(paste0("USA Total Deaths"))+
    theme_minimal()+
    theme(legend.position = "right",
          plot.caption = element_text(hjust = 0.5),
          plot.title = element_text(hjust = 0.5,size=16),
          text = element_text(size=14))+
    scale_fill_brewer(palette="Set3")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```



USA Death Cases From 2007 until 2019

```
# Clustering for deaths VS years
df_USAdet = select(main_df_na,year,deaths)
USAdetcls = kmeans(df_USAdet, centers = 1)
fviz_cluster(USAdetcls, data = df_USAdet)
```




```
# Linear regression of death rate VS years
summary(lm(death_rate~year, data = main_df_na))
```

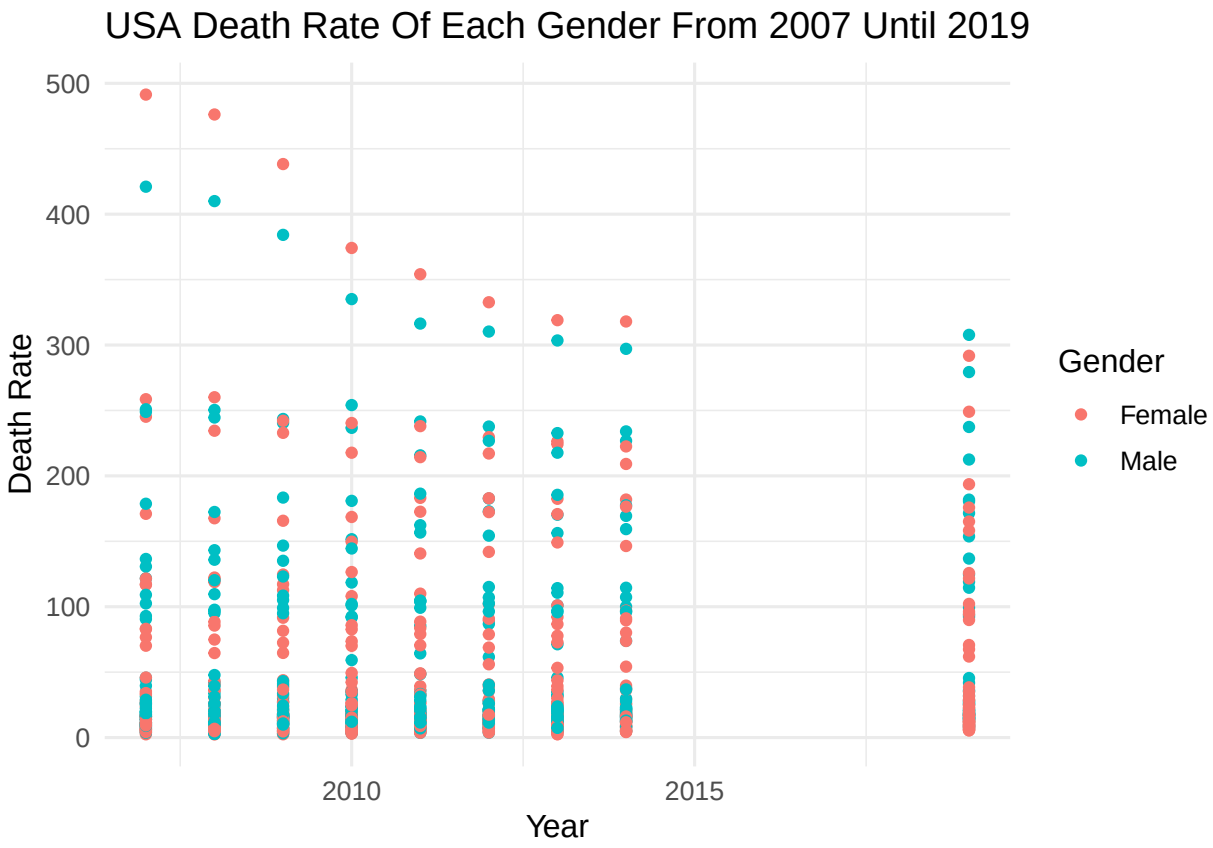
```
##
## Call:
## lm(formula = death_rate ~ year, data = main_df_na)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51.10 -41.57 -35.03  12.53 437.91
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.776e+01  1.472e+03   0.026    0.980
## year        7.837e-03  7.319e-01   0.011    0.991
##
## Residual standard error: 75.67 on 817 degrees of freedom
## Multiple R-squared:  1.403e-07,  Adjusted R-squared:  -0.001224
## F-statistic: 0.0001146 on 1 and 817 DF,  p-value: 0.9915
```

```
# Polynomial regression of death rate VS years
summary(lm(death_rate ~ poly(year, 2, raw = TRUE),data = main_df_na))
```

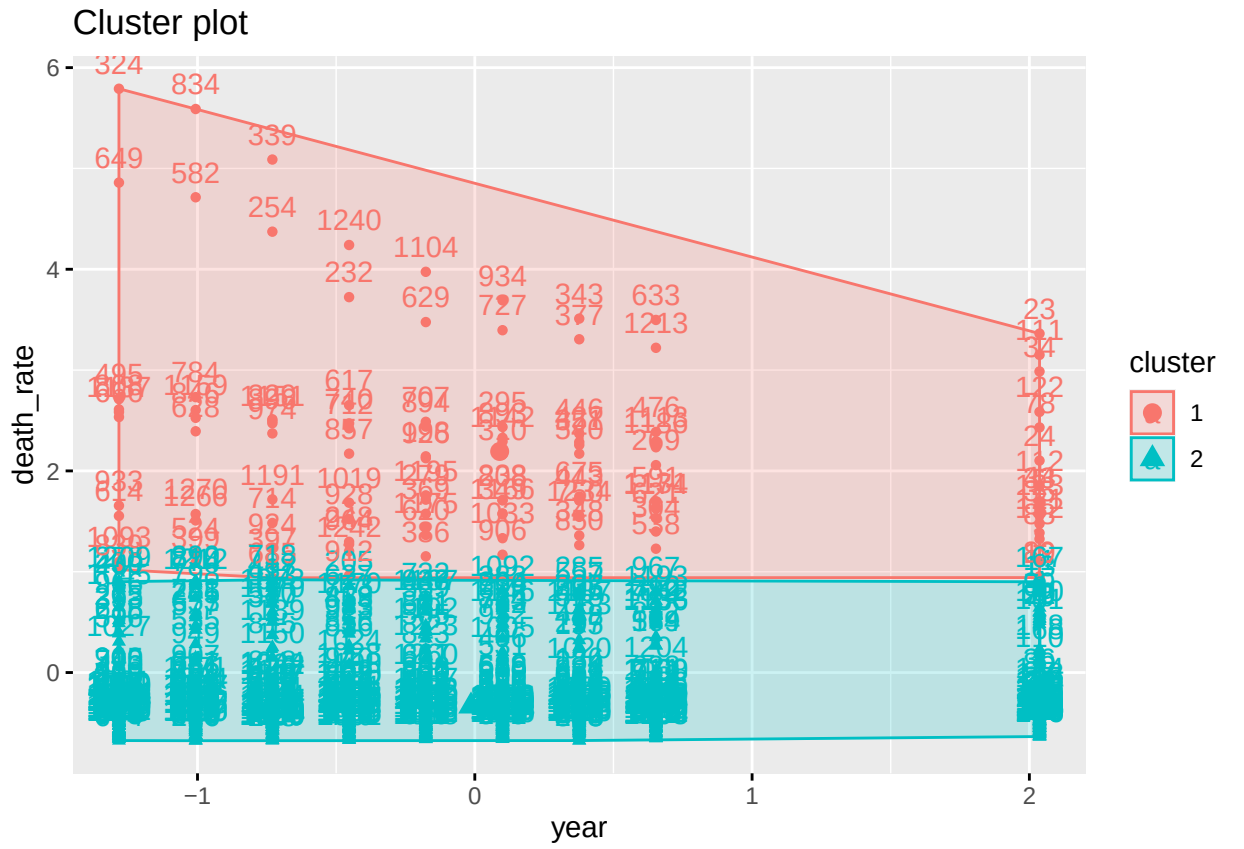
```
##
## Call:
```

```
## lm(formula = death_rate ~ poly(year, 2, raw = TRUE), data = main_df_na)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51.49 -41.63 -34.92  12.28 437.41
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.338e+04  7.540e+05   0.124    0.901
## poly(year, 2, raw = TRUE)1 -9.273e+01  7.491e+02  -0.124    0.902
## poly(year, 2, raw = TRUE)2  2.303e-02  1.861e-01   0.124    0.902
##
## Residual standard error: 75.71 on 816 degrees of freedom
## Multiple R-squared:  1.892e-05,  Adjusted R-squared:  -0.002432
## F-statistic: 0.007721 on 2 and 816 DF,  p-value: 0.9923
```

```r
# Plotting the graph for death rate VS years
ggplot(main_df_na,aes(x=year, y=death_rate, color = sex)) +
  ggtitle("USA Death Rate Of Each Gender From 2007 Until 2019") +
  xlab("Year") +
  ylab("Death Rate") +
  theme_minimal(base_size = 12) +
  geom_point() +
  scale_color_discrete(name = "Gender")
```



```
# Clustering for death rate VS years
df_USAdetrate = select(main_df_na,year,death_rate)
USAdetratecls = kmeans(df_USAdetrate, centers = 2)
fviz_cluster(USAdetratecls, data = df_USAdetrate)
```



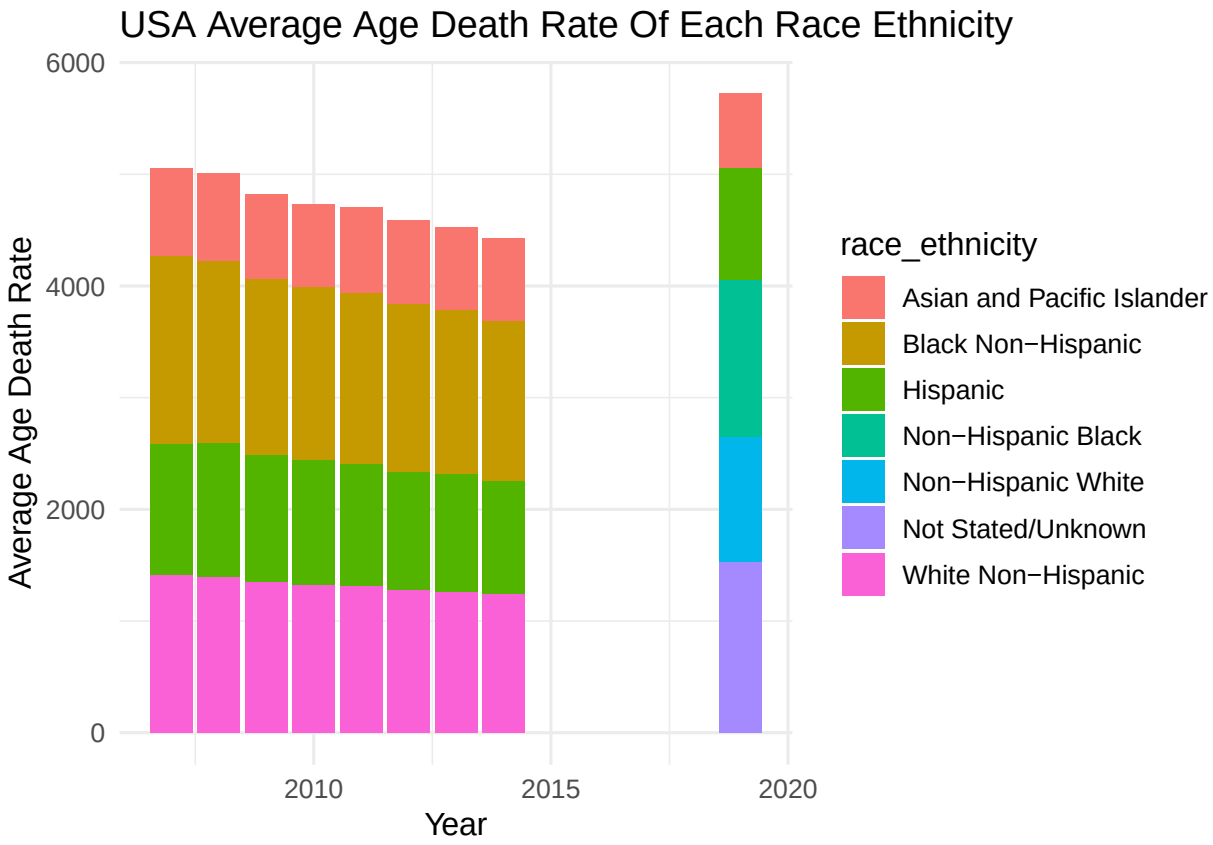

```
# Linear regression of age adjusted death rate VS years
summary(lm(age_adjusted_death_rate~year, data = main_df_na))
```

```
##
## Call:
## lm(formula = age_adjusted_death_rate ~ year, data = main_df_na)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.04 -41.27 -33.01  24.80 364.61
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  934.9747  1343.8957   0.696    0.487
## year          -0.4383     0.6681  -0.656    0.512
##
## Residual standard error: 69.06 on 817 degrees of freedom
## Multiple R-squared:  0.0005267,  Adjusted R-squared:  -0.0006967
## F-statistic: 0.4305 on 1 and 817 DF,  p-value: 0.5119
```

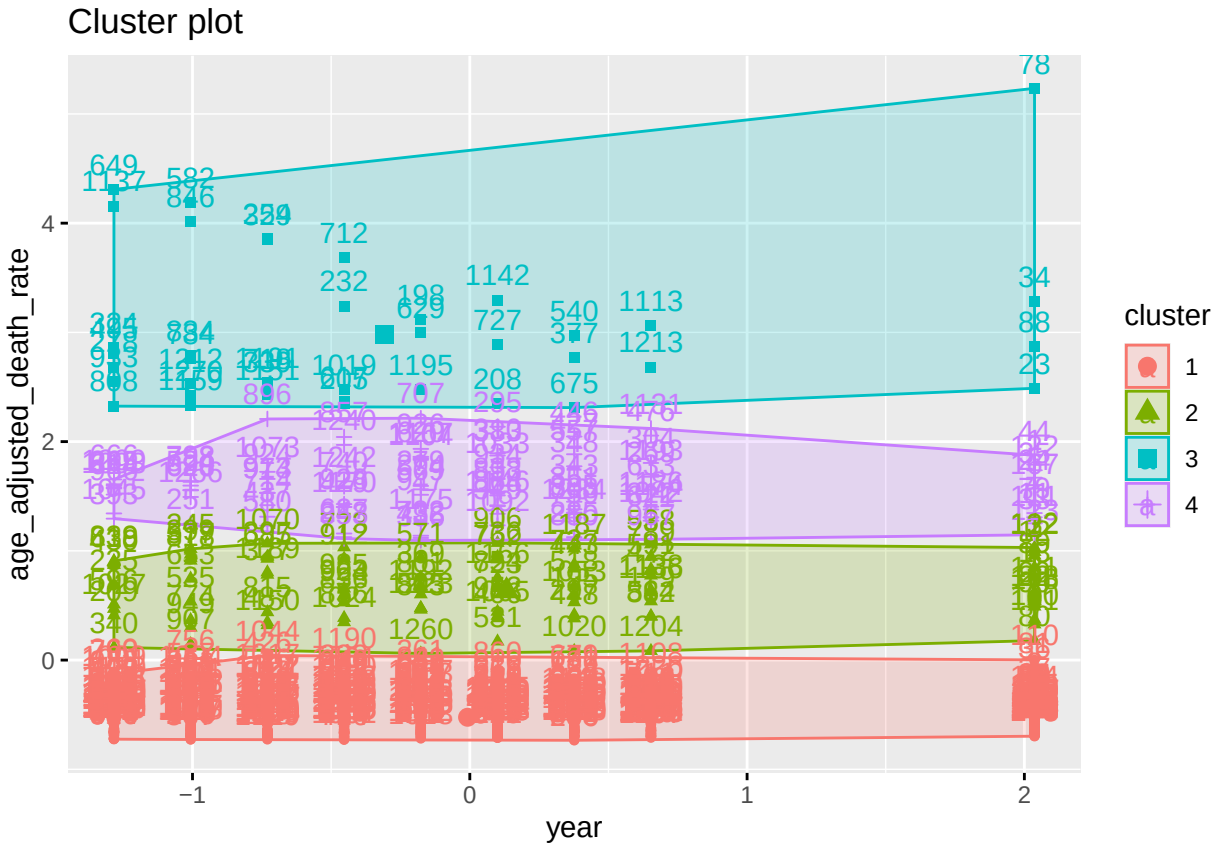
```r
# Polynomial regression of age adjusted death rate VS years
summary(lm(age_adjusted_death_rate ~ poly(year, 2, raw = TRUE),data = main_df_na))
```

```
##
## Call:
## lm(formula = age_adjusted_death_rate ~ poly(year, 2, raw = TRUE),
##     data = main_df_na)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.60 -41.16 -32.82  24.02 363.16
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.917e+05  6.881e+05   0.424    0.672
## poly(year, 2, raw = TRUE)1 -2.893e+02  6.837e+02  -0.423    0.672
## poly(year, 2, raw = TRUE)2  7.175e-02  1.698e-01   0.423    0.673
##
## Residual standard error: 69.1 on 816 degrees of freedom
## Multiple R-squared:  0.0007453,  Adjusted R-squared:  -0.001704
## F-statistic: 0.3043 on 2 and 816 DF,  p-value: 0.7377
```

```r
# Plotting the graph for age adjusted death rate VS years
ggplot(df_grpby,aes(x=year, y=age_adjusted_death_rate,fill=race_ethnicity)) +
  ggtitle("USA Average Age Death Rate Of Each Race Ethnicity") +
  xlab("Year") +
  ylab("Average Age Death Rate") +
  theme_minimal(base_size = 12) +
  geom_bar(stat="identity") +
  scale_color_discrete(name = "Race Ethnicity")
```




```
# clustering for age adjusted death rate VS years
df_USAraceeth = select(main_df_na,year,age_adjusted_death_rate)
USAraceeth = kmeans(df_USAraceeth, centers = 4)
fviz_cluster(USAraceeth, data = df_USAraceeth)
```


Cluster plot
age_adjusted_death_rate
year
cluster
1
2
3
4