Suicide

Kamalraj N

2023-02-19

```
#importing library
library(ggplot2)
library(lattice)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(rmarkdown)
#importing data
df<-read.csv("suicide.csv")
#Basic insights and Preprocessing
head(df)

##   Indicator  Geography Year Strata Strata.Name Numerator Denominator
Rate
## 1   Suicide CALIFORNIA 2000    Sex      Female       682    17080000
3.993
## 2   Suicide CALIFORNIA 2000    Sex        Male      2431    16920000
14.370
## 3   Suicide CALIFORNIA 2000    Sex       Total      3113    34000000
9.156
## 4   Suicide CALIFORNIA 2001    Sex      Female       685    17340000
3.950
## 5   Suicide CALIFORNIA 2001    Sex        Male      2571    17170000
14.970
## 6   Suicide CALIFORNIA 2001    Sex       Total      3256    34510000
9.434
##   Age.Adjusted.Rate
## 1             4.067
## 2            15.890
## 3             9.564
## 4             4.005
## 5            16.410
## 6             9.813
```

```
tail(df)
```

```
##      Indicator  Geography      Year Strata Strata.Name Numerator
Denominator
## 1399   Suicide   Mariposa 2015-2017    Sex       Total        18
54200
## 1400   Suicide      Butte 2015-2017    Sex       Total       128
676000
## 1401   Suicide Santa Cruz 2015-2017    Sex       Total       136
829300
## 1402   Suicide     Sonoma 2015-2017    Sex       Total       205
1512000
## 1403   Suicide    Trinity 2015-2017    Sex       Total        16
40490
## 1404   Suicide     Sutter 2015-2017    Sex       Total        42
294800
##       Rate Age.Adjusted.Rate
## 1399 33.21              33.12
## 1400 18.94              18.31
## 1401 16.40              15.74
## 1402 13.55              12.19
## 1403 39.51              43.87
## 1404 14.25              13.87
```

```
#Renaming a column
df <- rename(df, location = Geography)
df <- rename(df, Strata_Name = Strata.Name)
df <- rename(df, Age_Adjusted_Rate = Age.Adjusted.Rate)
head(df)
```

```
##   Indicator   location Year Strata Strata_Name Numerator Denominator
Rate
## 1   Suicide CALIFORNIA 2000    Sex      Female       682    17080000
3.993
## 2   Suicide CALIFORNIA 2000    Sex        Male      2431    16920000
14.370
## 3   Suicide CALIFORNIA 2000    Sex       Total      3113    34000000
9.156
## 4   Suicide CALIFORNIA 2001    Sex      Female       685    17340000
3.950
## 5   Suicide CALIFORNIA 2001    Sex        Male      2571    17170000
14.970
## 6   Suicide CALIFORNIA 2001    Sex       Total      3256    34510000
9.434
##   Age_Adjusted_Rate
## 1             4.067
## 2            15.890
## 3             9.564
## 4             4.005
```

```
## 5            16.410
## 6             9.813

#checking the dimensation
dim(df)

## [1] 1404    9

#To summarize the dataset in statistical measures
summary(df)

##   Indicator           location             Year              Strata
##  Length:1404        Length:1404        Length:1404        Length:1404
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Strata_Name          Numerator        Denominator             Rate
##  Length:1404        Min.   :   11.0   Min.   :    26970   Min.   : 2.088
##  Class :character   1st Qu.:   31.0   1st Qu.:   299825   1st Qu.: 6.022
##  Mode  :character   Median :   75.5   Median :   666100   Median :11.040
##                     Mean   :  369.4   Mean   :  3740042   Mean   :12.700
##                     3rd Qu.:  181.8   3rd Qu.:  2083750   3rd Qu.:17.023
##                     Max.   :12560.0   Max.   :118100000   Max.   :48.890
##  Age_Adjusted_Rate
##  Min.   : 2.184
##  1st Qu.: 6.042
##  Median :10.910
##  Mean   :12.246
##  3rd Qu.:16.295
##  Max.   :52.250

#randomly choosing index
sample_index=sample(1:nrow(df),20,replace = FALSE)
sample_index

##  [1]  907  140 1140  657 1255  922  684 1018  766 1125   79  931  229  412
205
## [16]  761 1179 1100  241  271

#To display the structure
str(df)

## 'data.frame':    1404 obs. of  9 variables:
##  $ Indicator        : chr  "Suicide" "Suicide" "Suicide" "Suicide" ...
##  $ location         : chr  "CALIFORNIA" "CALIFORNIA" "CALIFORNIA"
"CALIFORNIA" ...
##  $ Year             : chr  "2000" "2000" "2000" "2001" ...
##  $ Strata           : chr  "Sex" "Sex" "Sex" "Sex" ...
##  $ Strata_Name      : chr  "Female" "Male" "Total" "Female" ...
##  $ Numerator        : int  682 2431 3113 685 2571 3256 710 2500 3210 733
```

```
...
##  $ Denominator      : int  17080000 16920000 34000000 17340000 17170000
34510000 17550000 17380000 34940000 17780000 ...
##  $ Rate             : num  3.99 14.37 9.16 3.95 14.97 ...
##  $ Age_Adjusted_Rate: num  4.07 15.89 9.56 4 16.41 ...

#Filter(used to subset a data frame)
f<-df%>%
filter(Year<=2005)
head(f)

##   Indicator   location Year Strata Strata_Name Numerator Denominator
Rate
## 1   Suicide CALIFORNIA 2000    Sex      Female       682    17080000
3.993
## 2   Suicide CALIFORNIA 2000    Sex        Male      2431    16920000
14.370
## 3   Suicide CALIFORNIA 2000    Sex       Total      3113    34000000
9.156
## 4   Suicide CALIFORNIA 2001    Sex      Female       685    17340000
3.950
## 5   Suicide CALIFORNIA 2001    Sex        Male      2571    17170000
14.970
## 6   Suicide CALIFORNIA 2001    Sex       Total      3256    34510000
9.434
##   Age_Adjusted_Rate
## 1             4.067
## 2            15.890
## 3             9.564
## 4             4.005
## 5            16.410
## 6             9.813

# Subset the data to include only the columns of interest
df_subset <- df %>%
select(location,Year,Strata,Strata_Name,Numerator,Denominator,Rate,Age_Adjust
ed_Rate)
head(df_subset)

##     location Year Strata Strata_Name Numerator Denominator   Rate
## 1 CALIFORNIA 2000    Sex      Female       682    17080000  3.993
## 2 CALIFORNIA 2000    Sex        Male      2431    16920000 14.370
## 3 CALIFORNIA 2000    Sex       Total      3113    34000000  9.156
## 4 CALIFORNIA 2001    Sex      Female       685    17340000  3.950
## 5 CALIFORNIA 2001    Sex        Male      2571    17170000 14.970
## 6 CALIFORNIA 2001    Sex       Total      3256    34510000  9.434
##   Age_Adjusted_Rate
## 1             4.067
## 2            15.890
## 3             9.564
## 4             4.005
```

```
## 5            16.410
## 6             9.813
```

```
# Subset the data to include only the Suicide Rate column
df_rate <- df %>%
select(Rate)
head(df_rate)
```

```
##     Rate
## 1  3.993
## 2 14.370
## 3  9.156
## 4  3.950
## 5 14.970
## 6  9.434
```

```
#summarize the sucide rate
summary(df_rate)
```

```
##       Rate
##  Min.   : 2.088
##  1st Qu.: 6.022
##  Median :11.040
##  Mean   :12.700
##  3rd Qu.:17.023
##  Max.   :48.890
```

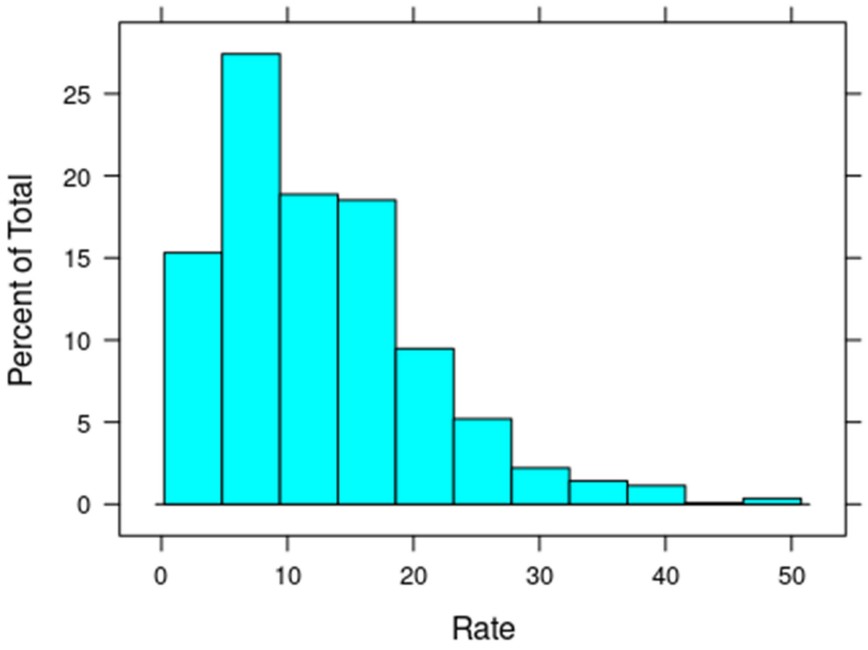
```
#2
# Subset the data to include only the Suicide Rate column
df_rate <- df %>%
select(Rate)
head(df_rate)
```

```
##     Rate
## 1  3.993
## 2 14.370
## 3  9.156
## 4  3.950
## 5 14.970
## 6  9.434
```

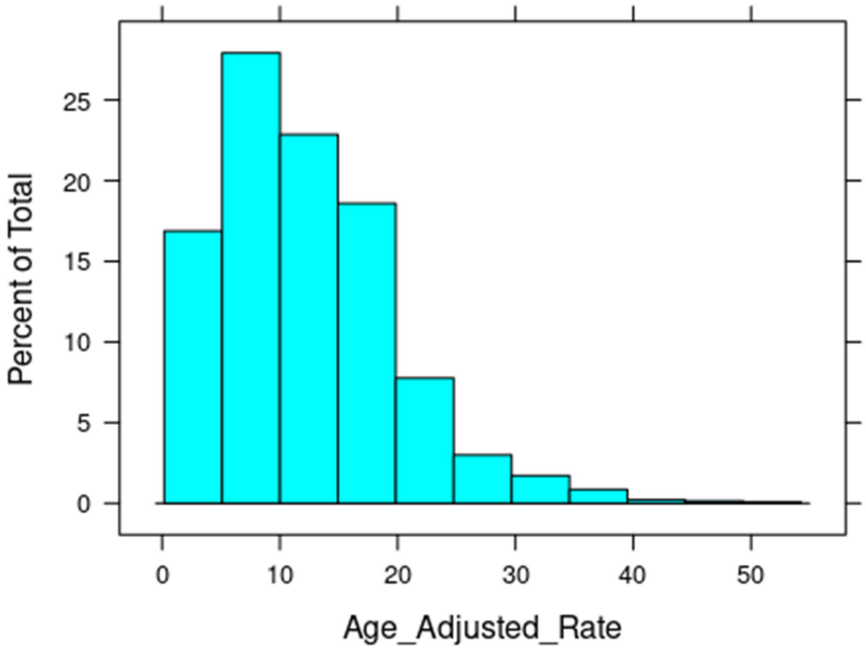
```
#summarize the sucide rate
summary(df_rate)
```

```
##       Rate
##  Min.   : 2.088
##  1st Qu.: 6.022
##  Median :11.040
##  Mean   :12.700
##  3rd Qu.:17.023
##  Max.   :48.890
```

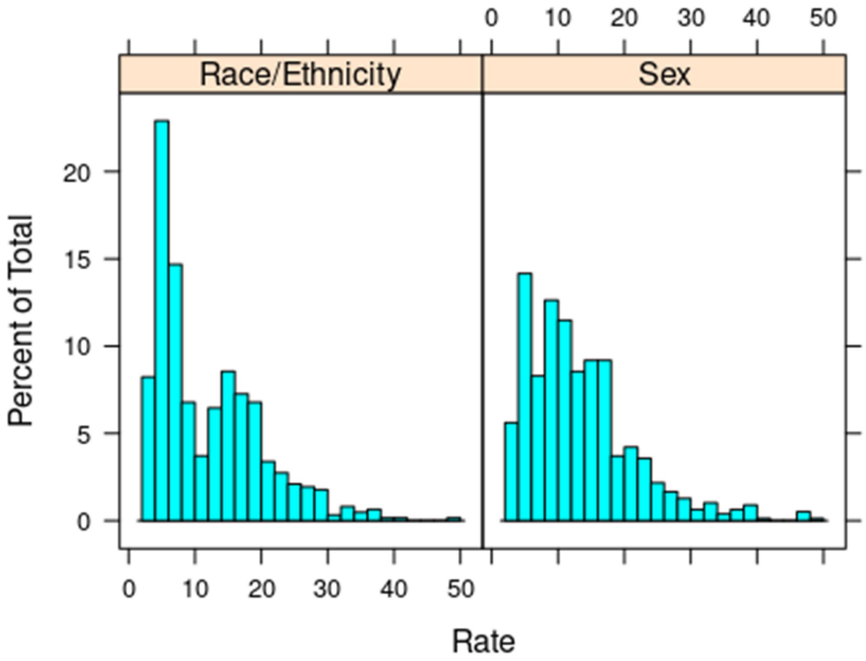
```
#histogram ##########
histogram(~Rate,data=df_rate)
```



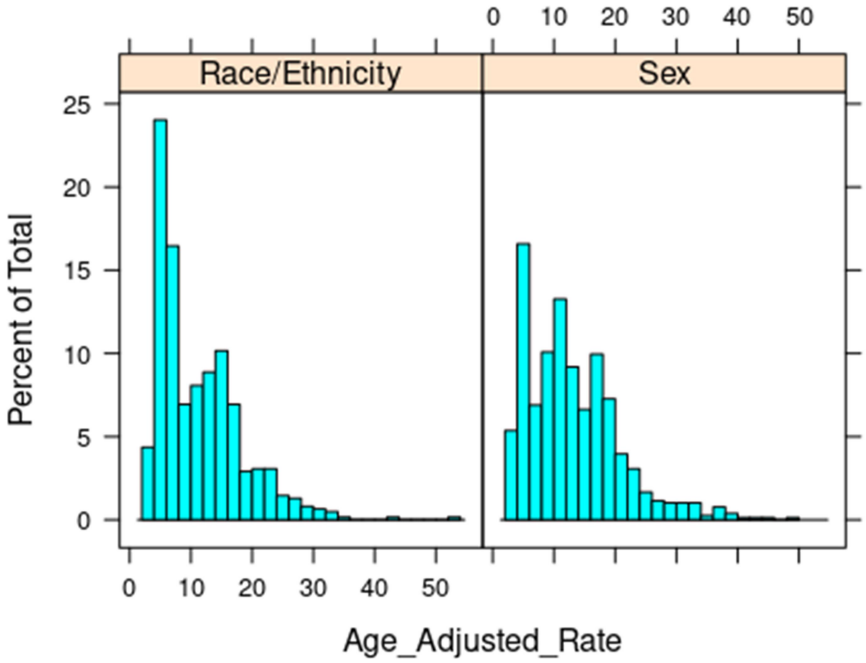

```
histogram(~Age_Adjusted_Rate,data=df)
```



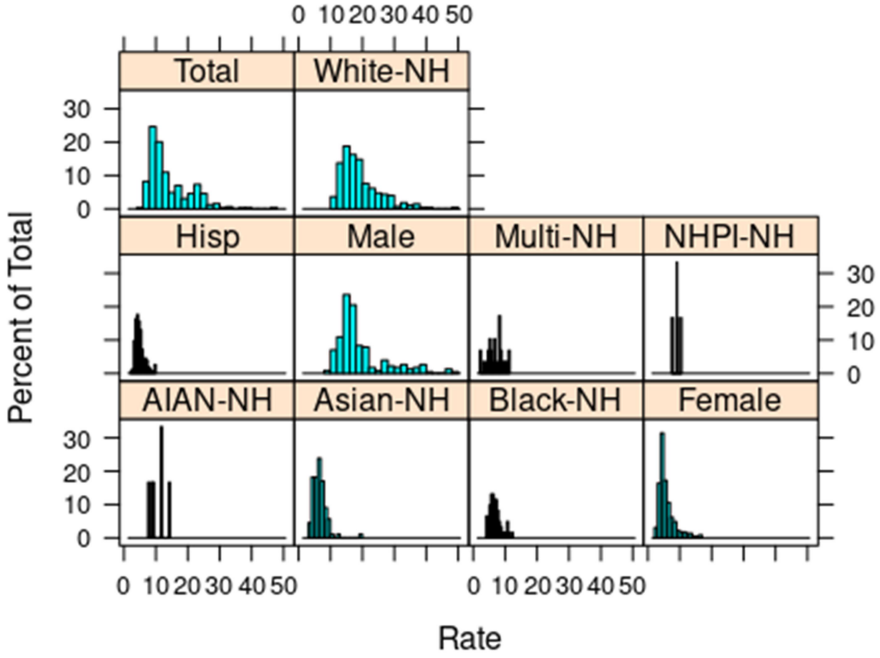

```
histogram(~Rate|Strata,data=df,breaks=20)
```



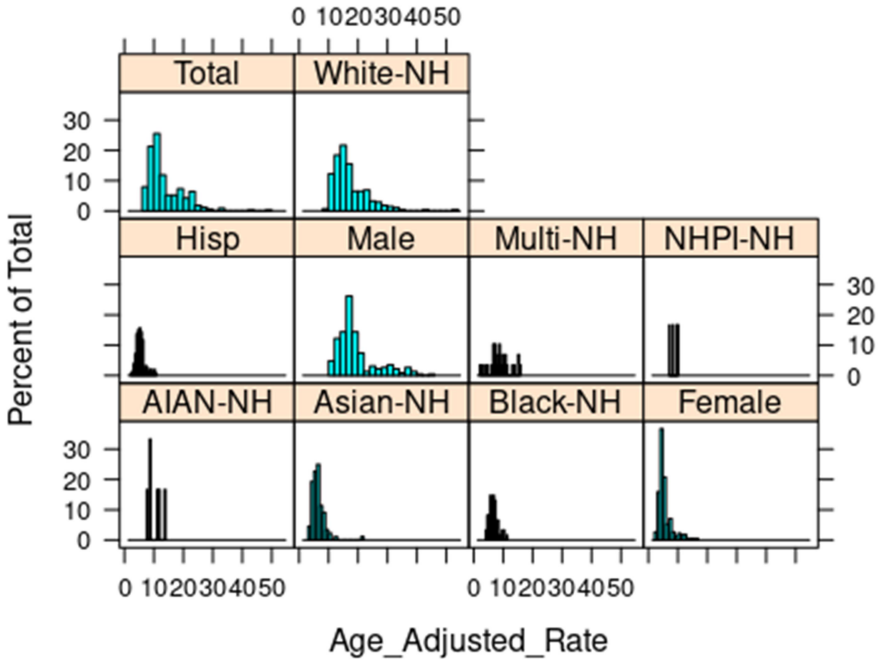

```
histogram(~Age_Adjusted_Rate|Strata,data=df,breaks=20)
```



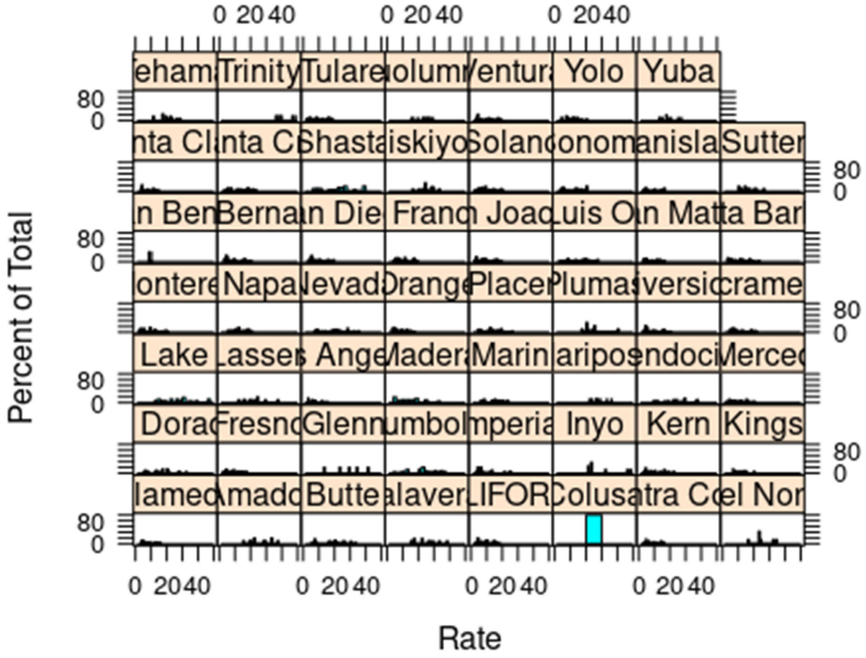

```
histogram(~Rate|Strata_Name,data=df,breaks=20)
```



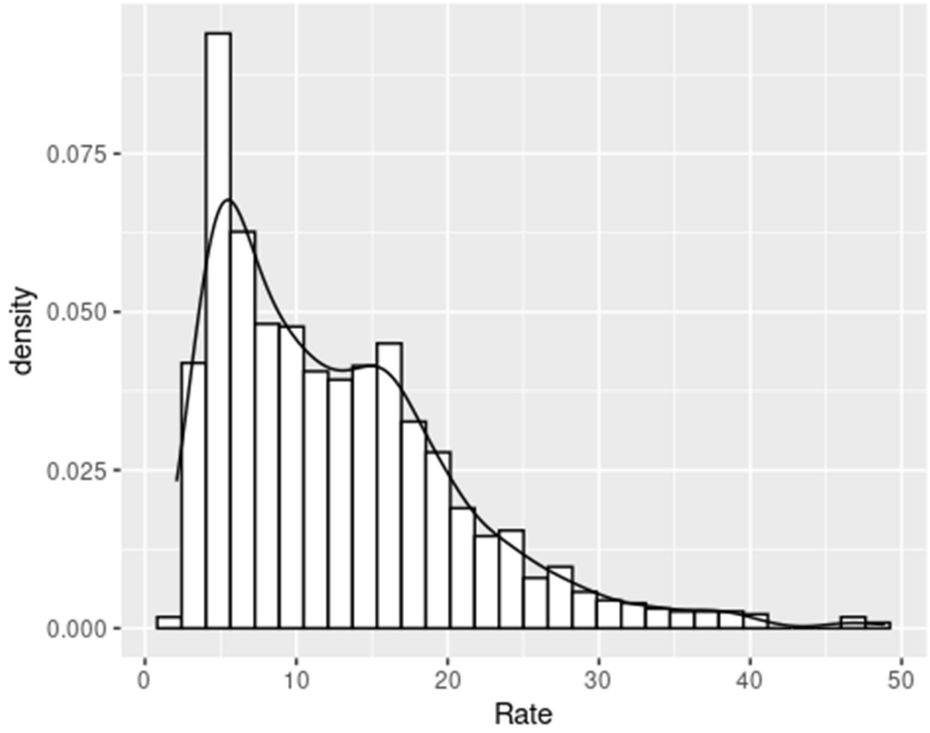

```
histogram(~Age_Adjusted_Rate|Strata_Name,data=df,breaks=20)
```




```
histogram(~Rate|location,data=df,breaks=20)
```

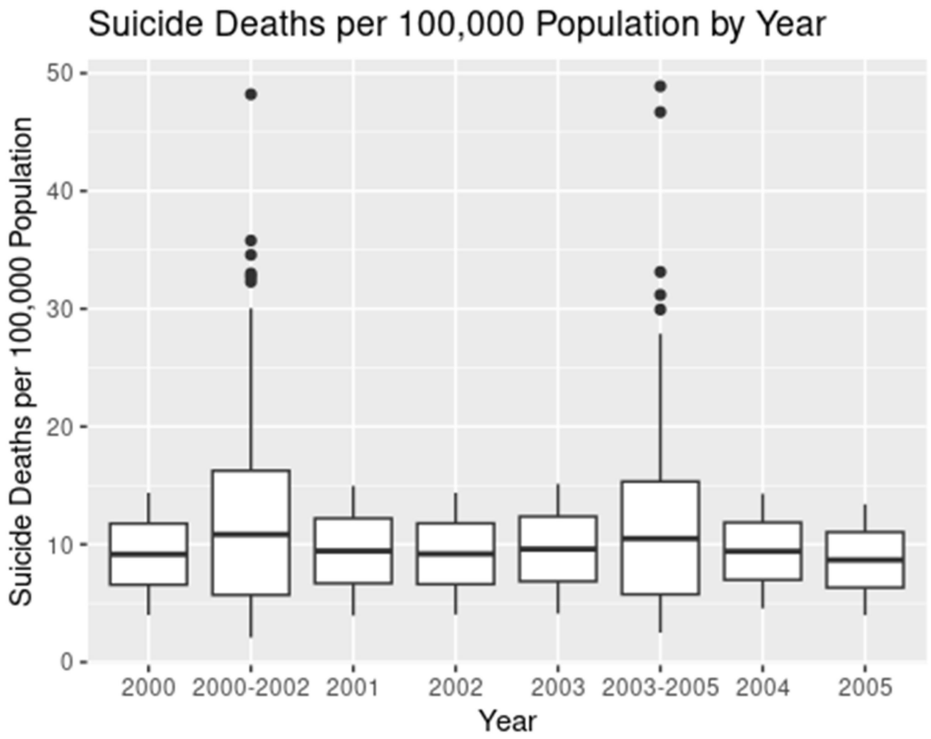


```
# Create a histogram with a density curve overlaid
ggplot(df_rate, aes(x = Rate)) +
geom_histogram(aes(y = ..density..), color = "black", fill = "white") +
geom_density(alpha = .2)

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2
3.4.0.
## ℹ Please use `after_stat(density)` instead.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

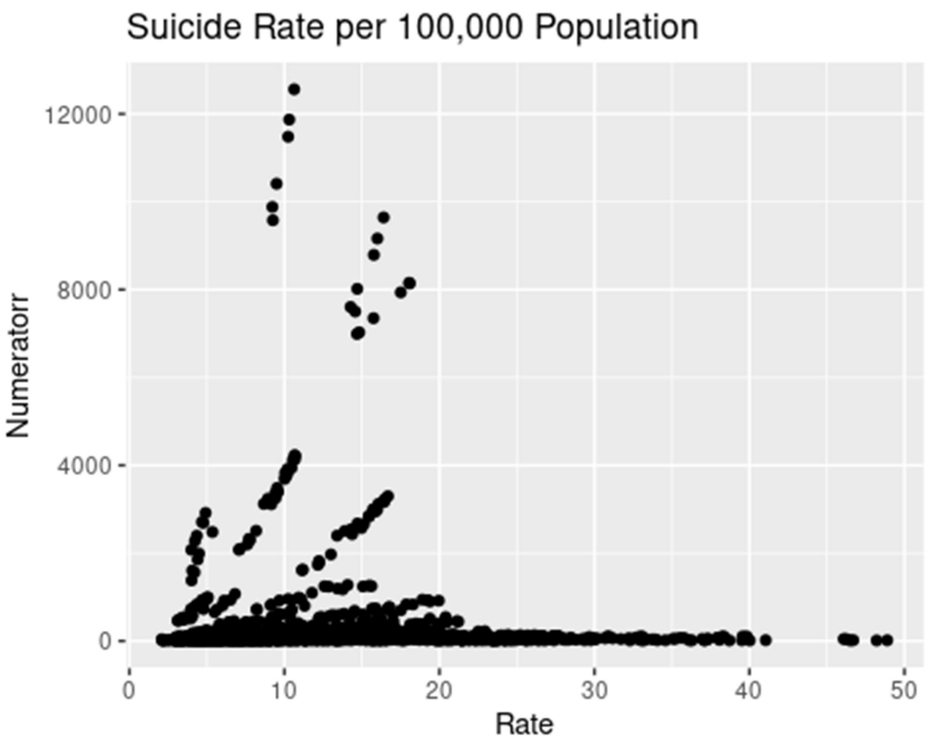



```
#Boxplot ####
f1<-df%>%
filter(Year<=2005)
head(f1)

##   Indicator   location Year Strata Strata_Name Numerator Denominator
Rate
## 1   Suicide CALIFORNIA 2000    Sex      Female       682    17080000
3.993
## 2   Suicide CALIFORNIA 2000    Sex        Male      2431    16920000
14.370
## 3   Suicide CALIFORNIA 2000    Sex       Total      3113    34000000
9.156
## 4   Suicide CALIFORNIA 2001    Sex      Female       685    17340000
3.950
## 5   Suicide CALIFORNIA 2001    Sex        Male      2571    17170000
```

```
14.970
## 6   Suicide CALIFORNIA 2001    Sex       Total      3256    34510000
9.434
##   Age_Adjusted_Rate
## 1             4.067
## 2            15.890
## 3             9.564
## 4             4.005
## 5            16.410
## 6             9.813

summary(f1)

##   Indicator           location             Year              Strata
##  Length:442         Length:442         Length:442         Length:442
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Strata_Name          Numerator        Denominator             Rate
##  Length:442         Min.   :  11.00   Min.   :    26970   Min.   : 2.088
##  Class :character   1st Qu.:  28.25   1st Qu.:   314700   1st Qu.: 5.667
##  Mode  :character   Median :  71.50   Median :   688150   Median :10.470
##                     Mean   : 346.50   Mean   :  3745997   Mean   :11.758
##                     3rd Qu.: 177.75   3rd Qu.:  2146750   3rd Qu.:15.685
##                     Max.   :9881.00   Max.   :107100000   Max.   :48.890
##  Age_Adjusted_Rate
##  Min.   : 2.184
##  1st Qu.: 5.858
##  Median :10.635
##  Mean   :11.656
##  3rd Qu.:15.547
##  Max.   :52.250

#Boxplot ####
library(tidyr)
# Create box plot by year(>2005)
suicide_data_long <- pivot_longer(f1, cols = c("Rate"), names_to = "Measure",
values_to = "Value")
ggplot(suicide_data_long, aes(x = Year, y = Value, group = Year)) +
geom_boxplot() +
labs(title = "Suicide Deaths per 100,000 Population by Year",
x = "Year",
y = "Suicide Deaths per 100,000 Population")
```




```
#Scatterplot

ggplot(df, aes(x = Rate, y =Numerator)) +
  geom_point() +
  labs(title = "Suicide Rate per 100,000 Population",
       x = "Rate", y = "Numeratorr")
```


Suicide Rate per 100,000 Population
Numeratorr
12000
8000
4000
0
0
10
20
30
40
50
Rate

Description:

- Indicator - LGHC Indicator
- Geography - Data is either state- or county-level. Location is based on residence, not occurrence of event.
- Year - Year in which incident occurred. Due to statistical stability and small cell size concerns, county-level data is in 3-year aggregates.
- Strata - General demographic category under which responses have been stratified.
- Strata Name - Individual strata within a demographic category.
- Numerator - Number of incidents. For 3-year aggregates, number is the sum total of all 3 years (not an average).
- Denominator - Population at risk, taken from California Department of Finance projections. For 3-year aggregates, number is the sum total of all 3 years (not an average).
- Rate - Number of incidents per 100,000 populations at risk.
- Age Adjusted Rate - Age adjustment, also called age standardization, is a technique used to allow statistical populations to be compared when the age profiles of the populations are quite different.

Assumption:

- The data is related to suicide deaths per 100,000 populations.
- The data is reported for a certain time period and location(s).
- The data is collected from reliable sources such as official government records.
- The data is complete and accurate to the best of the reporting authority's knowledge.
- The data includes all reported cases of suicide deaths in the given population and location(s).
- The data may be subject to variations due to differences in data collection methods and reporting practices across different locations.
- The dataset may be used to study the trends and patterns of suicide deaths in the given population and location(s) over time.
- The dataset may be useful for policymakers, researchers, and mental health professionals in designing and implementing suicide prevention strategies.

Inferences:

- The fluctuating trend in suicide death rates may reflect changes in societal, economic, and cultural factors that influence suicide risk, such as access to mental health care, substance abuse, and social isolation.
- The differences in suicide death rates by demographic groups may indicate the need for targeted prevention and intervention efforts to address the underlying risk factors and promote mental health and wellbeing.
- The dataset highlights the importance of reliable and timely surveillance
- Data on suicide deaths to inform public health policy and practice and reduce the burden of suicide in the population.