Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Group Lab 3

U.S. traffic fatalities: 1980-2004

In this lab, you are asked to answer the question **"Do changes in traffic laws affect traffic fatalities?"** To do so, you will conduct the tasks specified below using the data set *driving.Rdata*, which includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws.

Specifically, this data set contains data for the 48 continental U.S. states from 1980 through 2004. Various driving laws are indicated in the data set, such as the alcohol level at which drivers are considered legally intoxicated. There are also indicators for "per se" laws—where licenses can be revoked without a trial—and seat belt laws. A few economics and demographic variables are also included. The description of the each of the variables in the dataset is come with the dataste.

```
library(foreign)
library(gplots)
library(ggplot2)
library(stats)
library(Hmisc)
library(car)
library(dplyr)
library(corrplot)
#library(corrgram)
library(lattice)
library(plm)
library(viridis)
library(tsibble)
library(forecast)

#for US State Map
library(tidyverse)
library(maps)
library(mapproj)
```

Exercises:

1. (30%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

```r
#driving <- miceadds::load.Rdata2( filename="driving.Rdata")

#load data
load('driving.RData')
driving.df <- data

#
dim(driving.df)
```

```
## [1] 1200   56
```

```r
#describe the variables
desc
```

```
##        variable                                          label
## 1          year                                1980 through 2004
## 2         state           48 continental states, alphabetical
## 3          sl55                                speed limit == 55
## 4          sl65                                speed limit == 65
## 5          sl70                                speed limit == 70
## 6          sl75                                speed limit == 75
## 7        slnone                                   no speed limit
## 8      seatbelt    =0 if none, =1 if primary, =2 if secondary
## 9        minage                            minimum drinking age
## 10      zerotol                              zero tolerance law
## 11          gdl                  graduated drivers license law
## 12        bac10                        blood alcohol limit .10
## 13        bac08                        blood alcohol limit .08
## 14        perse administrative license revocation (per se law)
## 15       totfat                       total traffic fatalities
## 16      nghtfat                     total nighttime fatalities
## 17      wkndfat                       total weekend fatalities
## 18    totfatpvm         total fatalities per 100 million miles
## 19   nghtfatpvm     nighttime fatalities per 100 million miles
## 20   wkndfatpvm       weekend fatalities per 100 million miles
## 21     statepop                                state population
## 22    totfatrte        total fatalities per 100,000 population
## 23   nghtfatrte    nighttime fatalities per 100,000 population
## 24   wkndfatrte       weekend accidents per 100,000 population
## 25   vehicmiles               vehicle miles traveled, billions
## 26         unem                     unemployment rate, percent
## 27    perc14_24          percent population aged 14 through 24
## 28     sl70plus                           sl70 + sl75 + slnone
## 29       sbprim                     =1 if primary seatbelt law
## 30      sbsecon                   =1 if secondary seatbelt law
## 31          d80                              =1 if year == 1980
## 32          d81
## 33          d82
```

```
## 34          d83
## 35          d84
## 36          d85
## 37          d86
## 38          d87
## 39          d88
## 40          d89
## 41          d90
## 42          d91
## 43          d92
## 44          d93
## 45          d94
## 46          d95
## 47          d96
## 48          d97
## 49          d98
## 50          d99
## 51          d00
## 52          d01
## 53          d02
## 54          d03
## 55          d04                              =1 if year == 2004
## 56 vehicmilespc
```

```
#examine the dat
rbind(head(data,2),tail(data,2))
```

```
##      year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl
## 1    1980     1    1    0    0    0      0        0     18       0   0
## 2    1981     1    1    0    0    0      0        0     18       0   0
## 1199 2003    51    0    0    0    1      0        2     21       1   0
## 1200 2004    51    0    0    0    1      0        2     21       1   0
##      bac10 bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm
## 1        1     0     0    940     422     236      3.20      1.437
## 2        1     0     0    933     434     248      3.35      1.558
## 1199     0     1     1    165      62      32      1.79      0.673
## 1200     0     1     1    164      67      31      1.77      0.723
##      wkndfatpvm statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem
## 1         0.803  3893888     24.14      10.84       6.06   29.37500  8.8
## 2         0.890  3918520     24.07      11.08       6.33   27.85200 10.7
## 1199      0.347   501242     32.92      12.37       6.38    9.21788  4.4
## 1200      0.335   507000     32.35      13.21       6.11    9.26600  3.7
##      perc14_24 sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88
## 1         18.9        0      0       0   1   0   0   0   0   0   0   0   0
## 2         18.7        0      0       0   0   1   0   0   0   0   0   0   0
## 1199      15.1        1      0       1   0   0   0   0   0   0   0   0   0
## 1200      14.9        1      0       1   0   0   0   0   0   0   0   0   0
##      d89 d90 d91 d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04
```

```
## 1      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 2      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 1199   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
## 1200   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
##      vehicmilespc
## 1        7543.874
## 2        7107.785
## 1199    18390.080
## 1200    18276.135
```

```
#check for nulls
apply(data, 2, function(x) any(is.na(x)))
```

```
##         year        state         sl55         sl65         sl70
##        FALSE        FALSE        FALSE        FALSE        FALSE
##         sl75       slnone     seatbelt       minage      zerotol
##        FALSE        FALSE        FALSE        FALSE        FALSE
##          gdl        bac10        bac08        perse       totfat
##        FALSE        FALSE        FALSE        FALSE        FALSE
##      nghtfat      wkndfat    totfatpvm   nghtfatpvm   wkndfatpvm
##        FALSE        FALSE        FALSE        FALSE        FALSE
##     statepop    totfatrte   nghtfatrte   wkndfatrte   vehicmiles
##        FALSE        FALSE        FALSE        FALSE        FALSE
##         unem    perc14_24     sl70plus       sbprim      sbsecon
##        FALSE        FALSE        FALSE        FALSE        FALSE
##          d80          d81          d82          d83          d84
##        FALSE        FALSE        FALSE        FALSE        FALSE
##          d85          d86          d87          d88          d89
##        FALSE        FALSE        FALSE        FALSE        FALSE
##          d90          d91          d92          d93          d94
##        FALSE        FALSE        FALSE        FALSE        FALSE
##          d95          d96          d97          d98          d99
##        FALSE        FALSE        FALSE        FALSE        FALSE
##          d00          d01          d02          d03          d04
##        FALSE        FALSE        FALSE        FALSE        FALSE
## vehicmilespc
##        FALSE
```

```
#check for gaps in panel
table(data$state)
```

```
##
##  1  3  4  5  6  7  8 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
## 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

```
table(data$year)
```

```
##
```

```
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
##   48   48   48   48   48   48   48   48   48   48   48   48   48   48   48
## 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
##   48   48   48   48   48   48   48   48   48   48
```

The Dataset is panel data, that contains observations about different US states from year 1980 to 2004. There are 1200 observations in total, with 56 columns . The data has 25 observations each, one per year, from 48 continental states except state ids 2,9 and 12 (which we will later identify as Alaska, District Of Columbia and Hawaii that are not part of continental US States). All variables are observed for all states and over all time periods, hence the panel is balanced.

The fields in the data can be divided into *dimensions*, *attributes* and *measures*. Dimensions are the unique combination that identifies the data record (aka the index of the panel data), which are `Year` and `State`. Attributes are the features of the dimensions, which for us is the traffic laws and demographic variables. Both types - dimensions and attributes - are explanatory variables. Measures are the quantitative variables of interest that includes fatality numbers. For us, the measure is *totfatrte*.

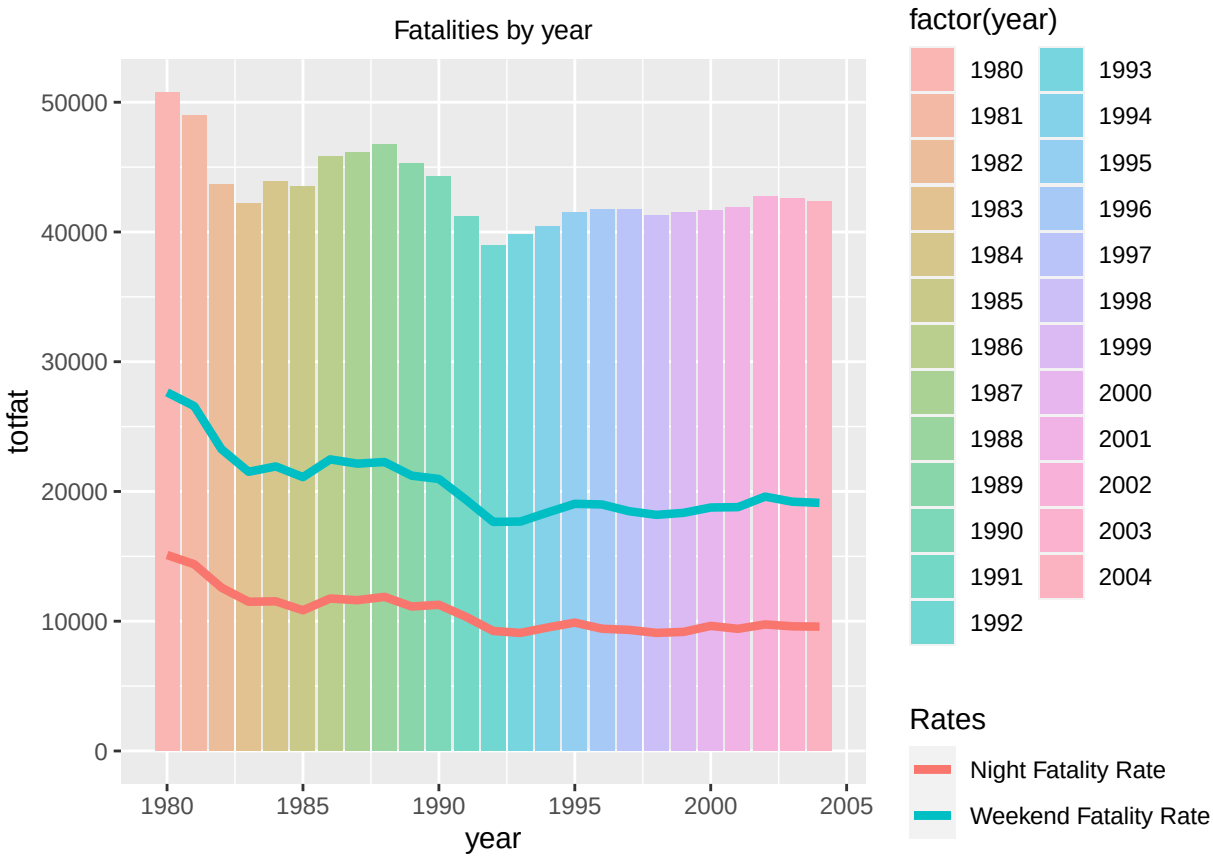
It will be useful to add more context around the state information, in addition to the state id. Since we know the id is alphabetical, we get the aphabetical list of US states with two letter abbreviated code, and match with the state variable in fatailty data.

```
#get state name
us.states = read.csv("usstates.csv", header = TRUE, sep = ",", dec = ".")

data.with.name <- merge(data, us.states, by=c("state","state"))
```

For EDA, to determine both the common and individual driving behaviors of US States across time, we'll analyze the aggregate of traffic laws in US across time. Then we'll focus on the fatality progression of top and bottom ranked US states across years. We'll also evaluate how the fatality pattern is different between years 1980 and 2004.

Below we analyze the fatality rate change by year.

```
#fatality change by year
traffic.yearly.aggr <- data %>%
  group_by(year) %>%
  summarise_at(vars(totfat, nghtfat,wkndfat), funs(sum))

ggplot(traffic.yearly.aggr, aes(year, totfat)) +
  geom_bar(aes(fill = factor(year)), position = "dodge", stat="identity") + ggtitle("Fatalities
  geom_line(aes(x = year, y = nghtfat, color="red"), size = 1.5, group = 1) +
geom_line(aes(x = year, y = wkndfat, color="blue"), size = 1.5, group = 1) +
scale_color_discrete(name = "Rates", labels = c("Night Fatality Rate", "Weekend Fatality Rate")
```



We can see that the fatality rate is loosely decreasing from 1980 to 2004. Night and Weekend fatality rate follow a similar pattern.

Below we analyze the growth in fatality rates by state

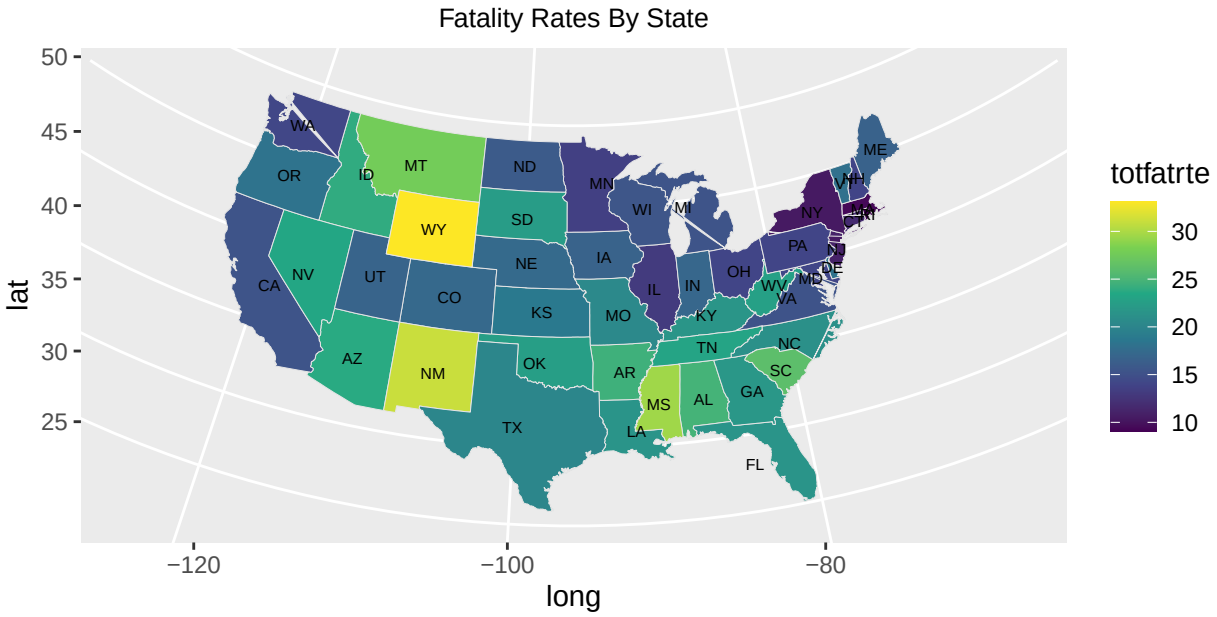
```
us_states <- map_data("state")

traffic.state.perc.name.aggr <- data.with.name %>%
  group_by(name,shortname) %>%
  summarise_at(vars(totfatrte), funs(mean))

data.with.name %>% filter(year %in% c(2004)) %>% dplyr::select(name,bac08,bac10)
```

```
##                name bac08 bac10
## 1           Alabama 1.000 0.000
## 2           Arizona 1.000 0.000
## 3          Arkansas 1.000 0.000
## 4        California 1.000 0.000
## 5          Colorado 0.500 0.500
## 6       Connecticut 1.000 0.000
## 7          Delaware 0.500 0.500
## 8           Florida 1.000 0.000
## 9           Georgia 1.000 0.000
## 10            Idaho 1.000 0.000
## 11         Illinois 1.000 0.000
```

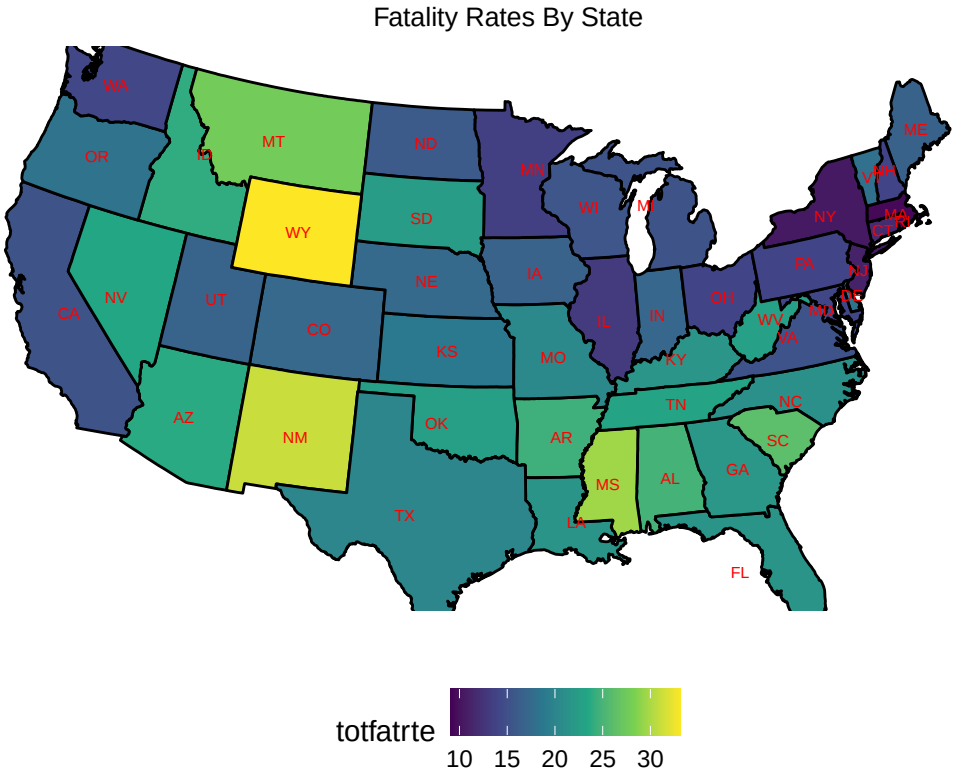
```
## 12        Indiana 1.000 0.000
## 13           Iowa 1.000 0.000
## 14         Kansas 1.000 0.000
## 15       Kentucky 1.000 0.000
## 16      Louisiana 1.000 0.000
## 17          Maine 1.000 0.000
## 18       Maryland 1.000 0.000
## 19  Massachusetts 1.000 0.000
## 20       Michigan 1.000 0.000
## 21      Minnesota 0.000 1.000
## 22    Mississippi 1.000 0.000
## 23       Missouri 1.000 0.000
## 24        Montana 1.000 0.000
## 25       Nebraska 1.000 0.000
## 26         Nevada 1.000 0.000
## 27  New Hampshire 1.000 0.000
## 28     New Jersey 1.000 0.000
## 29     New Mexico 1.000 0.000
## 30       New York 1.000 0.000
## 31 North Carolina 1.000 0.000
## 32   North Dakota 1.000 0.000
## 33           Ohio 1.000 0.000
## 34       Oklahoma 1.000 0.000
## 35         Oregon 1.000 0.000
## 36   Pennsylvania 1.000 0.000
## 37   Rhode Island 1.000 0.000
## 38 South Carolina 1.000 0.000
## 39   South Dakota 1.000 0.000
## 40      Tennessee 1.000 0.000
## 41          Texas 1.000 0.000
## 42           Utah 1.000 0.000
## 43        Vermont 1.000 0.000
## 44       Virginia 1.000 0.000
## 45     Washington 1.000 0.000
## 46  West Virginia 0.667 0.333
## 47      Wisconsin 1.000 0.000
## 48        Wyoming 1.000 0.000
```

```
traffic.state.perc.name.aggr$name <- tolower(traffic.state.perc.name.aggr$name)

state.data = us_states %>%
              group_by(region) %>%
              summarise(lat = mean(c(max(lat), min(lat))),
                        long = mean(c(max(long), min(long)))) %>%
              mutate(state = region) %>%
              left_join(traffic.state.perc.name.aggr, by=c("state"="name"))

us_states %>%
```

```
  left_join(traffic.state.perc.name.aggr, by=c("region"="name")) %>%
  ggplot(aes(x=long,y=lat,group=region, fill=totfatrte))  +
  geom_polygon(color = "gray90", size = 0.1) +
  #coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  coord_map(projection = "albers", lat0 = 45, lat1 = 55) +
  scale_fill_continuous(type = "viridis")+ geom_text(data = state.data, aes(x = long, y = lat,
  #scale_fill_brewer("Oranges")+
  theme(legend.position="bottom",
        axis.line=element_blank(),
        axis.text=element_blank(),
        axis.ticks=element_blank(),
        axis.title=element_blank(),
        panel.background=element_blank(),
        panel.border=element_blank(),
        panel.grid=element_blank())  + theme_gray() +
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + ggtitle("Fatality Rates By State")
```




```
ggplot(data = state.data, aes(map_id = state)) +
  geom_map(aes(fill = totfatrte),  color= "black", map = us_states) +
  expand_limits(x = state.data$long, y = state.data$lat) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 55) + scale_fill_continuous(type = "viridi
  geom_text(data = state.data, aes(x = long, y = lat, label = shortname), size = 2, color = "re
  scale_x_continuous(breaks = NULL) + scale_y_continuous(breaks = NULL) +
  labs(x = "", y = "") + theme(legend.position = "bottom",
                               panel.background = element_blank())  + theme(plot.title = eleme
```



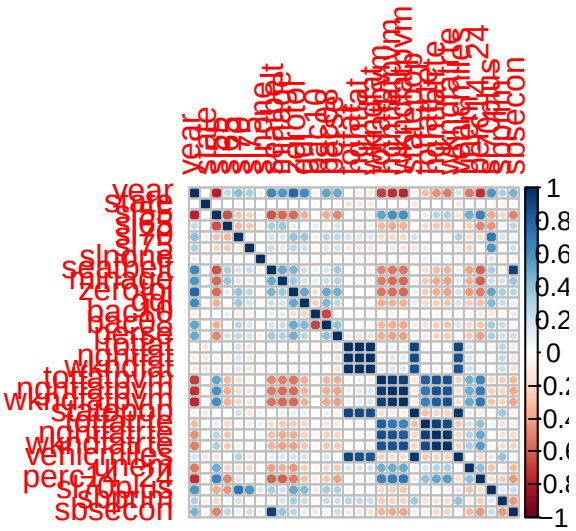
#TODO add GridExtra to so the above are in the same row

We can see that the fatality rates range from ~9 to ~34. Wyoming, New Mexico, Mississippi, Montana, and South Carolina are the states with highest fatality rates while New York, New Jersey and Rhode Island are the states with lowest fatality rates. The pattern shows that the states with more rural roads have higher fatality rates - the geography and road conditions are thus important omitted variables in the dataset. In addition, the fatality split by cause (drunk driving, speeding) by state by year could be an important predictor. Another omitted variable could be the a measure of compliance to the traffic laws - speed limit, seat belt - at the state level.

We also see that North East may be the safest region in road travel.

Below, we analyze the correlation between variables of interest.

```
corrplot(cor(driving.df[1:30]))
```



Percentage 14-24 population is correlated to weekend fatality rate. #TODO Explain in detail with more data

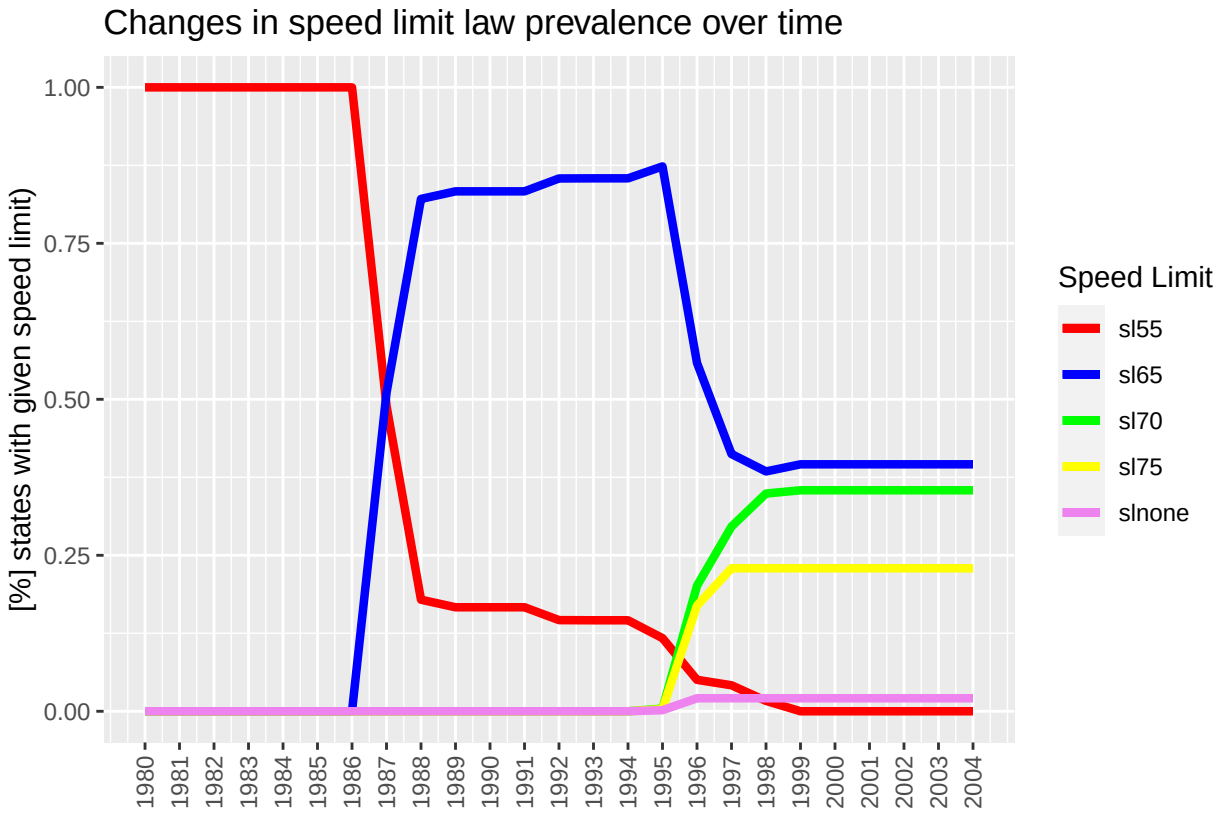
Below, we explore the prevalence of traffic laws over the years. First, we explore the speed limit.

```
# summarize the average statistics for speed limit in a data frame
sl.df <- data %>% group_by(year) %>%
summarise(sl55 = mean(sl55), sl65 = mean(sl65),sl70 = mean(sl70),sl75 = mean(sl75), slnone = me
sl.df
```

```
## # A tibble: 25 x 7
##     year  sl55  sl65  sl70  sl75 slnone totfatrte
##    <int> <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>
##  1  1980 1     0         0     0      0      25.5
##  2  1981 1     0         0     0      0      23.7
##  3  1982 1     0         0     0      0      20.9
##  4  1983 1     0         0     0      0      20.2
##  5  1984 1     0         0     0      0      20.3
##  6  1985 1     0         0     0      0      19.9
##  7  1986 1     0         0     0      0      20.8
##  8  1987 0.490 0.510     0     0      0      20.8
##  9  1988 0.179 0.821     0     0      0      20.9
## 10  1989 0.167 0.833     0     0      0      19.8
## # ... with 15 more rows
```

```
sl.plot <- ggplot(sl.df, aes(x = year)) +
  geom_line(aes(x = year, y = sl55, color = "sl55"), size = 1.5, group = 1)  +
  geom_line(aes(x = year, y = sl65, color = "sl65"), size = 1.5, group = 1) +
  geom_line(aes(x = year, y = sl70, color = "sl70"), size = 1.5, group = 1) +
  geom_line(aes(x = year, y = sl75, color = "sl75"), size = 1.5, group = 1) +
  geom_line(aes(x = year, y = slnone, color = "slnone"), size = 1.5, group = 1) +
  scale_x_continuous(breaks = seq(min(sl.df$year), max(sl.df$year), 1)) + theme(axis.text.x = e
      values = c(
           sl55="red",
```

```
            sl65="blue",
            sl70="green",
            sl75="yellow",
            slnone="violet")) + labs(title = "Changes in speed limit law prevalence over time",
y = "[%] states with given speed limit)",
x = "")

sl.plot
```



We can see that in 1980, all of the states had a speed limit of 55mph - however, the speed limits were more relaxed in the later years. Next we explore the blood alcohol limit changes over the years. It is worth noting that *Montana* with one of the highest fatality rates, did not enforce a speed limit as recently as 2004 as seen below.

```
data.with.name  %>%  filter (year %in% c('2004') & slnone == 1) %>% dplyr::select (name)
```

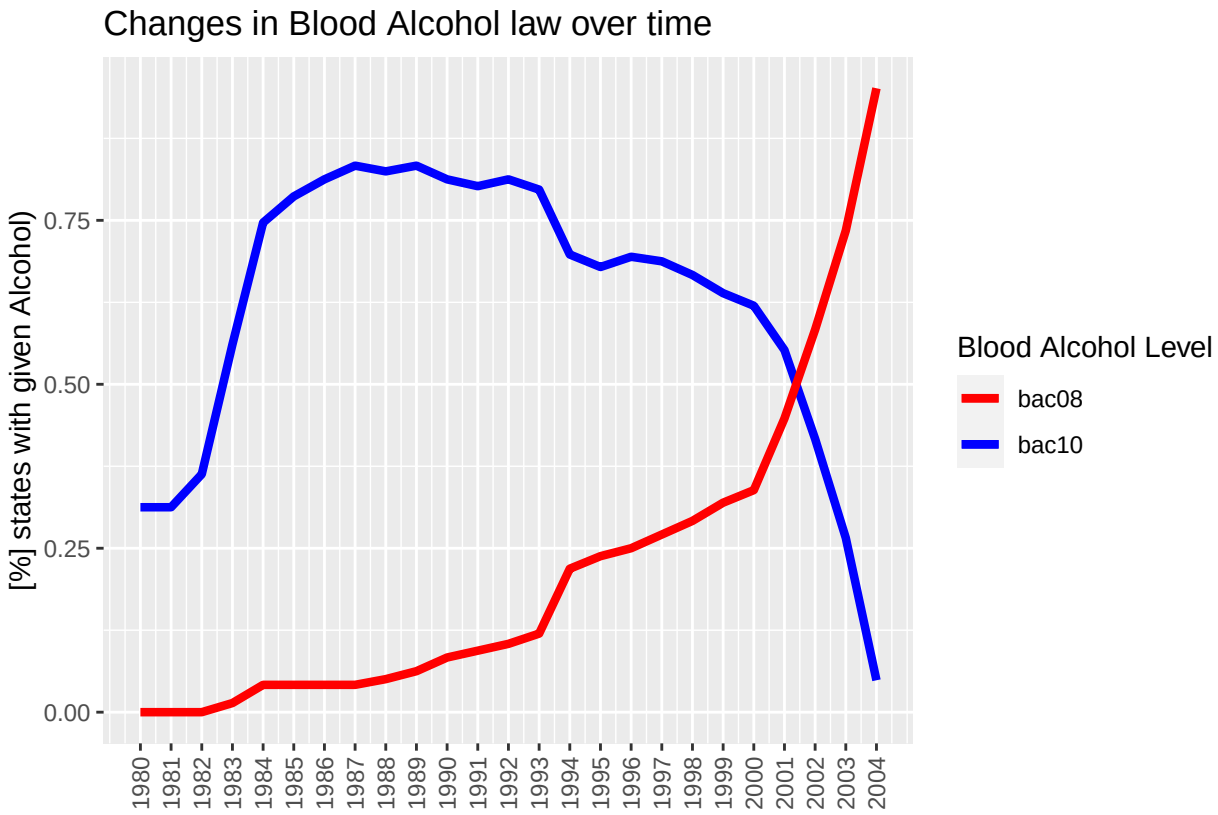
```
##      name
## 1 Montana
```

Next, we look at the permitted alcohol limit while driving.

```
# summarize the average statistics for blood alcohol in a data frame
bac.df <- data %>% group_by(year) %>%
summarise(bac10 = mean(bac10), bac08 = mean(bac08))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
bac.plot <- ggplot(bac.df, aes(x = year)) +
geom_line(aes(y = bac10, color='bac10'), size = 1.5, group = 1) + geom_line(aes(y=bac08, color=
      values = c(
           bac08="red",
           bac10="blue")) + labs(title = "Changes in Blood Alcohol law over time",
y = "[%] states with given Alcohol)",
x = "")

bac.plot
```



Over the years, more states are adopting stricter alcohol limits. This is consistent with the decrease in fatality rates over time that observed before.

```
data.with.name  %>%  filter (year %in% c('2004') & bac08 == 0) %>% dplyr::select (name)
```

```
##        name
## 1 Minnesota
```

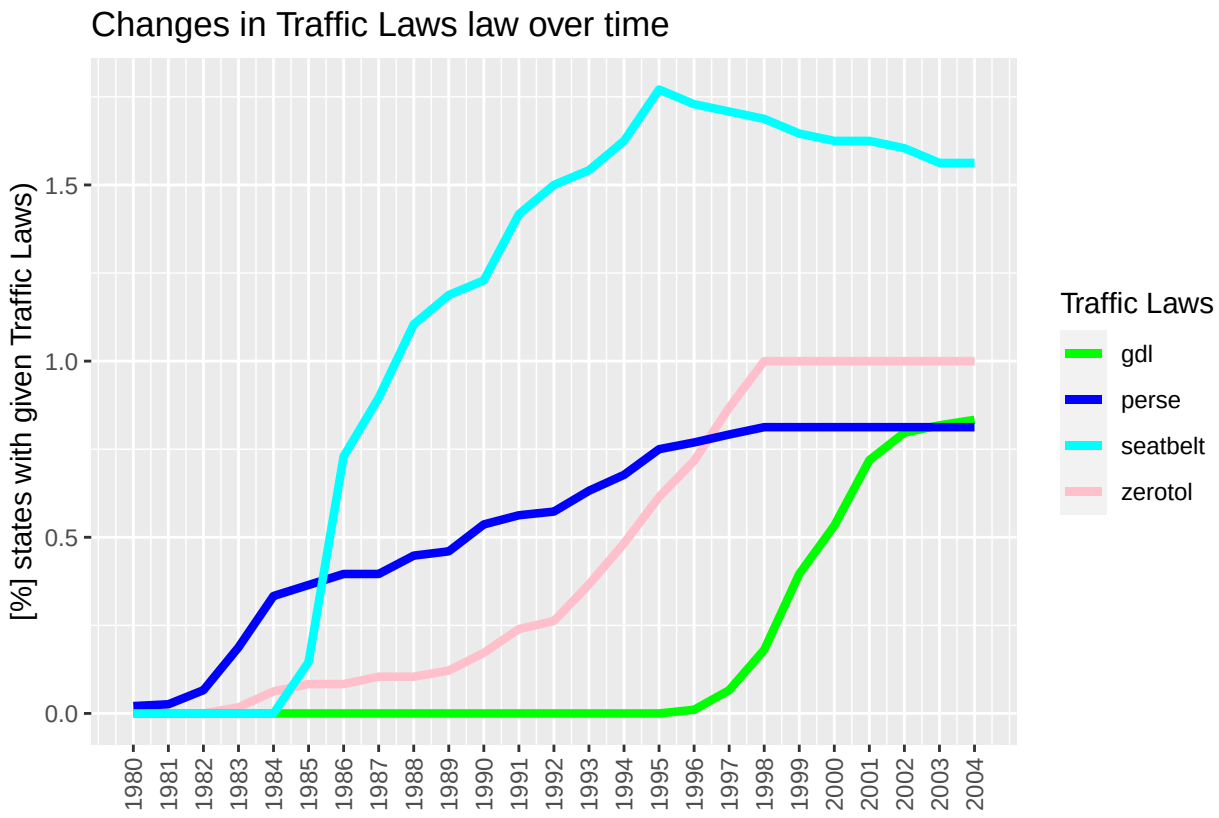
Minnesota is the only US State that still allows alcohol limit to be 0.10 in 2004.

Next, we will examine the general traffic law prevalence across years.

```
#summarize general laws
glaws.df <- data %>% group_by(year) %>%
summarise(zerotol = mean(zerotol), gdl = mean(gdl), perse = mean(perse), seatbelt = mean(seatb
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
glaws.plot <- ggplot(glaws.df, aes(x = year)) +
geom_line(aes(y = zerotol, color='zerotol'), size = 1.5, group = 1) + geom_line(aes(y=gdl, colo
      values = c(
           zerotol="pink",
           gdl="green",
           perse="blue",
           seatbelt="cyan")) + labs(title = "Changes in Traffic Laws law over time",
y = "[%] states with given Traffic Laws)",
x = "")

glaws.plot
```




```
data.with.name  %>%  filter (year %in% c('2004') & bac08 == 0) %>% dplyr::select (name)
```

```
##        name
## 1 Minnesota
```

```
data %>% filter (year %in% c('2004'))%>%
summarise(zerotol = sum(zerotol), gdl = sum(gdl), perse = sum(perse), seatbelt = sum(seatbelt))
```

```
##   zerotol gdl perse seatbelt
## 1      48  40    39       75
```

```
data %>% filter (year %in% c('2004'))%>% group_by(seatbelt) %>%
summarise(n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   seatbelt `n()`
##      <int> <int>
## 1        0     1
## 2        1    19
## 3        2    28
```

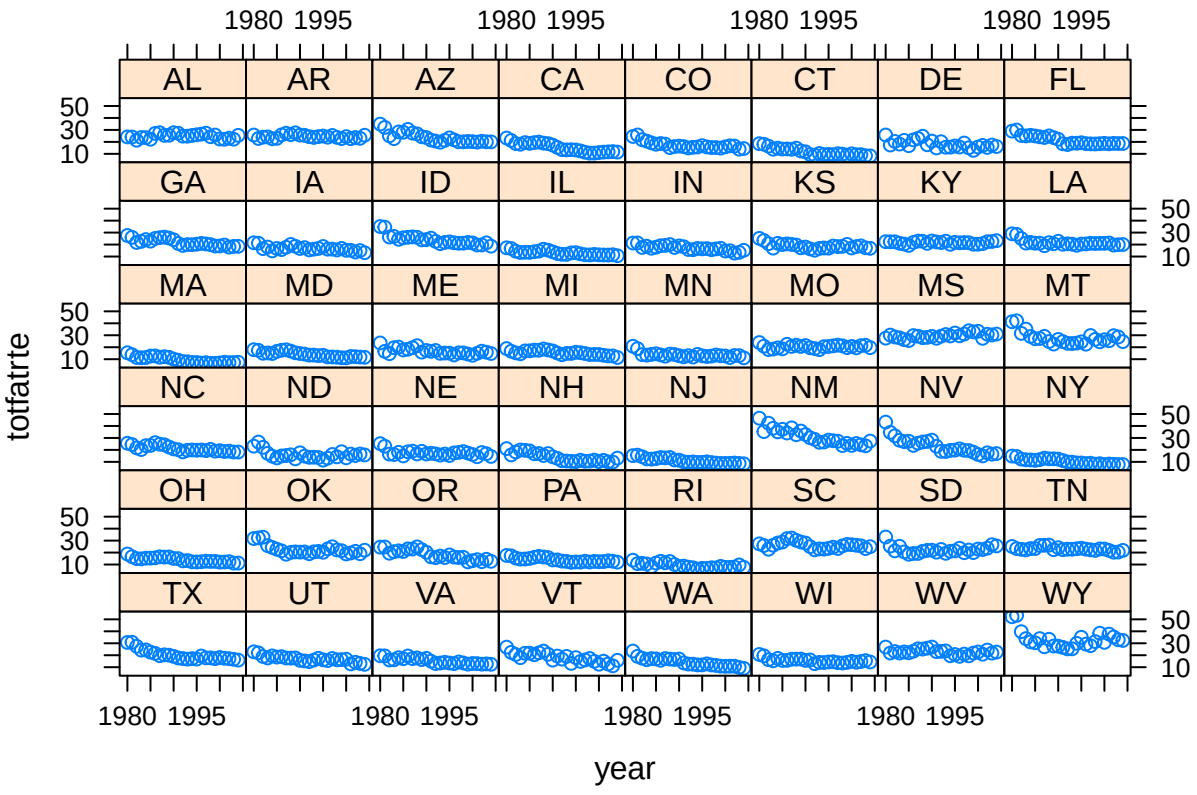
```
data.with.name  %>%  filter (year %in% c('2004') & seatbelt == 0) %>% dplyr::select (name)
```

```
##            name
## 1 New Hampshire
```

Over the years, we see more and more states adopting stricter traffic laws. As of 2004, all 48 states follow Zero Tolerance law, 40 states follow graduated DL law and 39 states follow per se law. Except *New Hampshire*, all states require at least the Primary seatbelt. Overall, it would seem that the decrease in fatality rates over years is in line with the stricter traffic laws requirements.

Below, we examine the fatality trend for individual states over years.

```
xyplot(totfatrte ~ year | shortname, data=data.with.name, as.table=T)
```



The above xyplot confirms that most of the States shows an overall decrease in the traffic fatality rate, except *Mississippi* the rates of which are shown below. One interesting point is that the traffic fatality rate is not dependent on the state area or population - the top 2 states with size and population, Texas and California, are not amoing the top states in traffic fatality rate.

```
data.with.name  %>%  filter (year %in% c('1980','2004') & shortname == 'MS') %>% dplyr::select
```

```
##   year totfatrte
## 1 1980     27.57
## 2 2004     31.00
```

Lets proceed to examine the individual behavior in the panel. First, we'll examine how the traffic fatality rates changed over years for the first 3 States from the top and bottom of the fatality rate.

```
#fatality change by state
traffic.state.aggr <- data.with.name %>%
  group_by(shortname) %>%
  summarise_at(vars(totfat, nghtfat,wkndfat), funs(sum))

traffic.state.perc.aggr <- data.with.name %>%
  group_by(shortname) %>%
  summarise_at(vars(totfatrte,nghtfatrte,wkndfatrte), funs(mean))

top.3.fatalities <- traffic.state.perc.aggr %>%
         filter(rank(desc(totfatrte))<=3) %>% arrange(desc(totfatrte))

bottom.3.fatalities <- traffic.state.perc.aggr %>%
         filter(rank((totfatrte))<=3) %>% arrange((totfatrte))

cbind(top.3.fatalities[,1:2],bottom.3.fatalities[,1:2])
```

```
##   shortname totfatrte shortname totfatrte
## 1        WY   33.1408        RI    9.0900
## 2        NM   31.0608        MA    9.4512
## 3        MS   29.5548        NY   10.4380
```

```
data.top.filtered <- data.with.name %>% filter(shortname %in% c("MS","NM","WY"))
data.bottom.filtered <- data.with.name %>% filter(shortname %in% c("MA","NY","RI"))

data.with.name %>% filter(shortname %in% c("MS","NM","WY","MA","NY","RI") & year == '2004') %>%
```

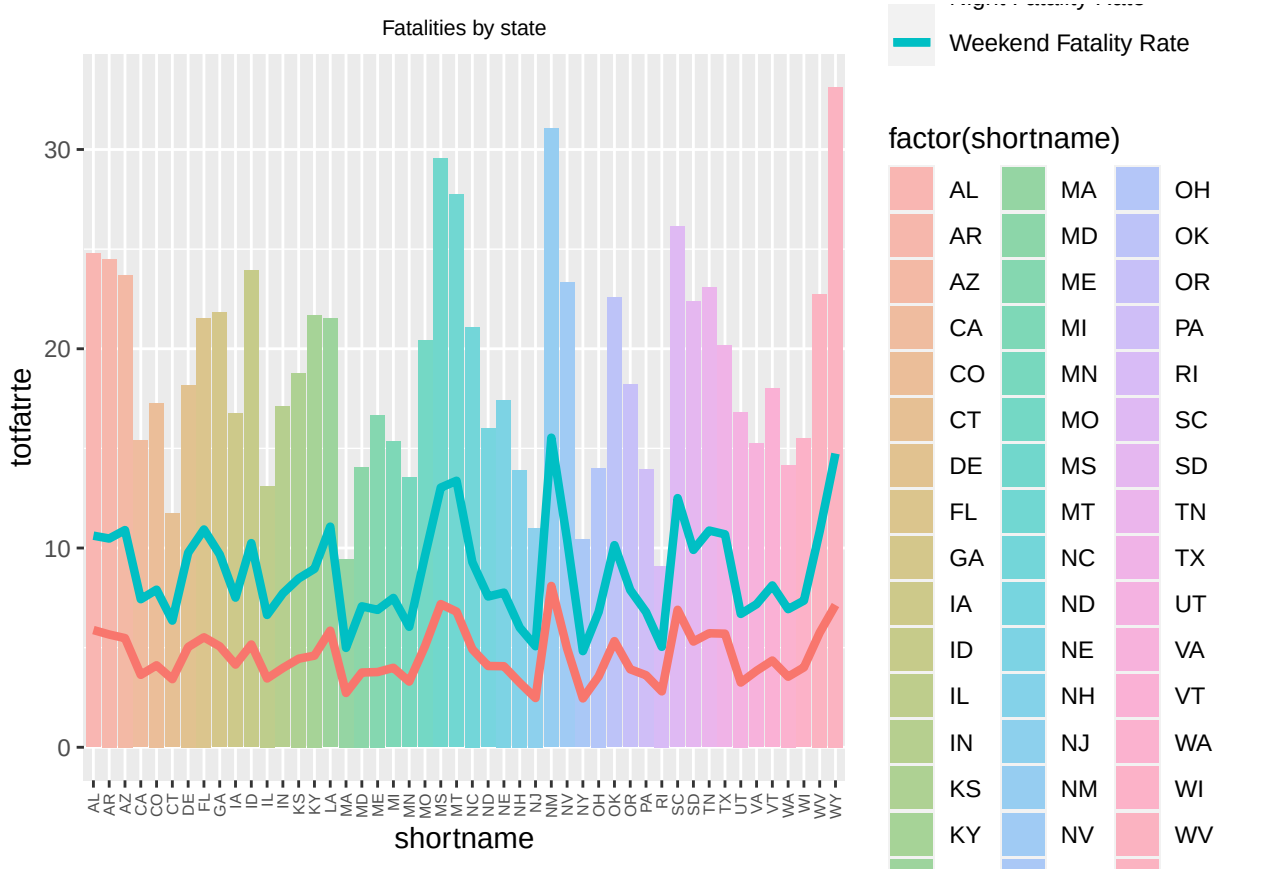
```
##   year          name totfatrte
## 1 2004       Wyoming     32.35
## 2 2004   Mississippi     31.00
## 3 2004    New Mexico     27.38
## 4 2004      New York      7.77
## 5 2004  Rhode Island      7.68
## 6 2004 Massachusetts      7.42
```

```
#data.with.name %>% filter(year == '2004') %>% dplyr::select(year,name,totfatrte) %>% arrange(

data.merged <- union(data.top.filtered,data.bottom.filtered)
```

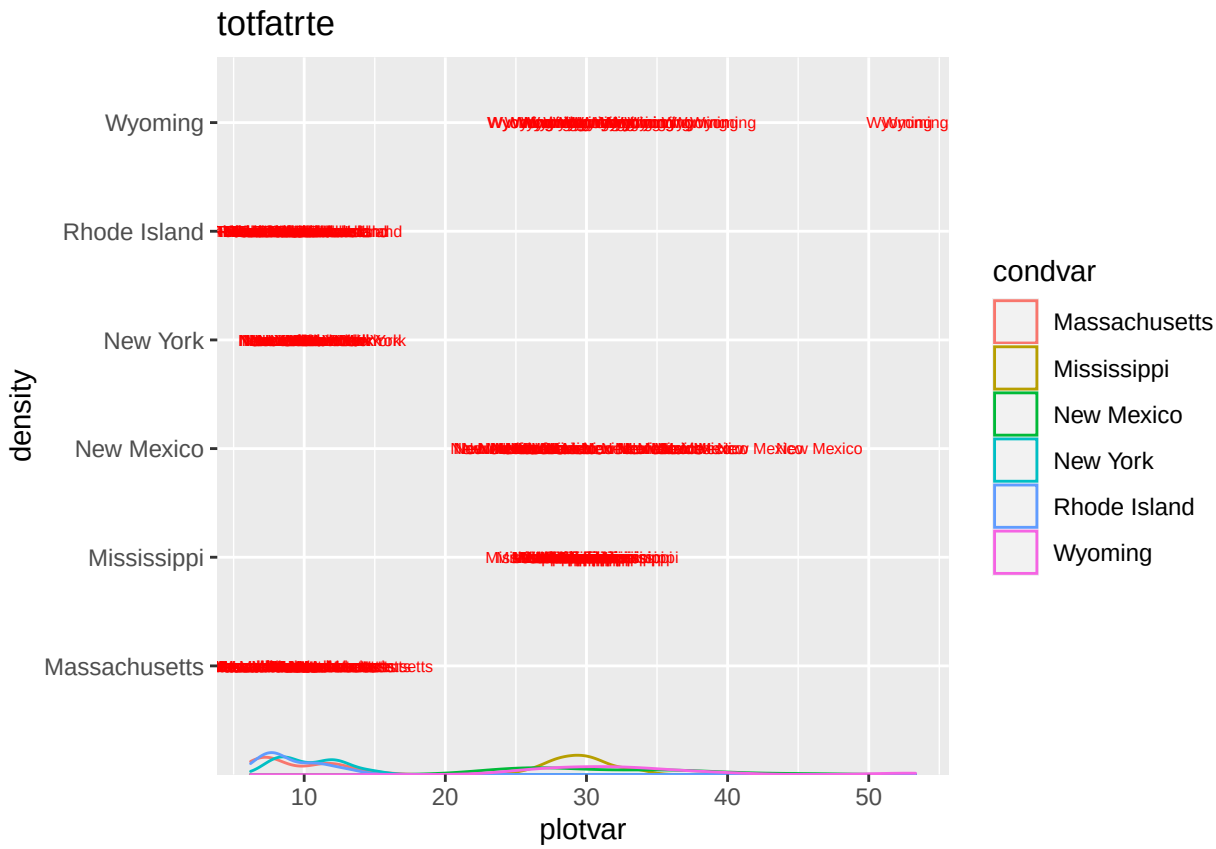
The top 3 are Wyoming, New Mexico and Mississippi. The bottom 3 are Rhode Island, New York and Massachussets. To put this in context, in 2004, in Wyoming, the probability of dying in a motor

vehicle accident is nearly 5 times as high as in Rhode Island, the state with the lowest death rate. Below, we see the average fatality rate for each state across years.

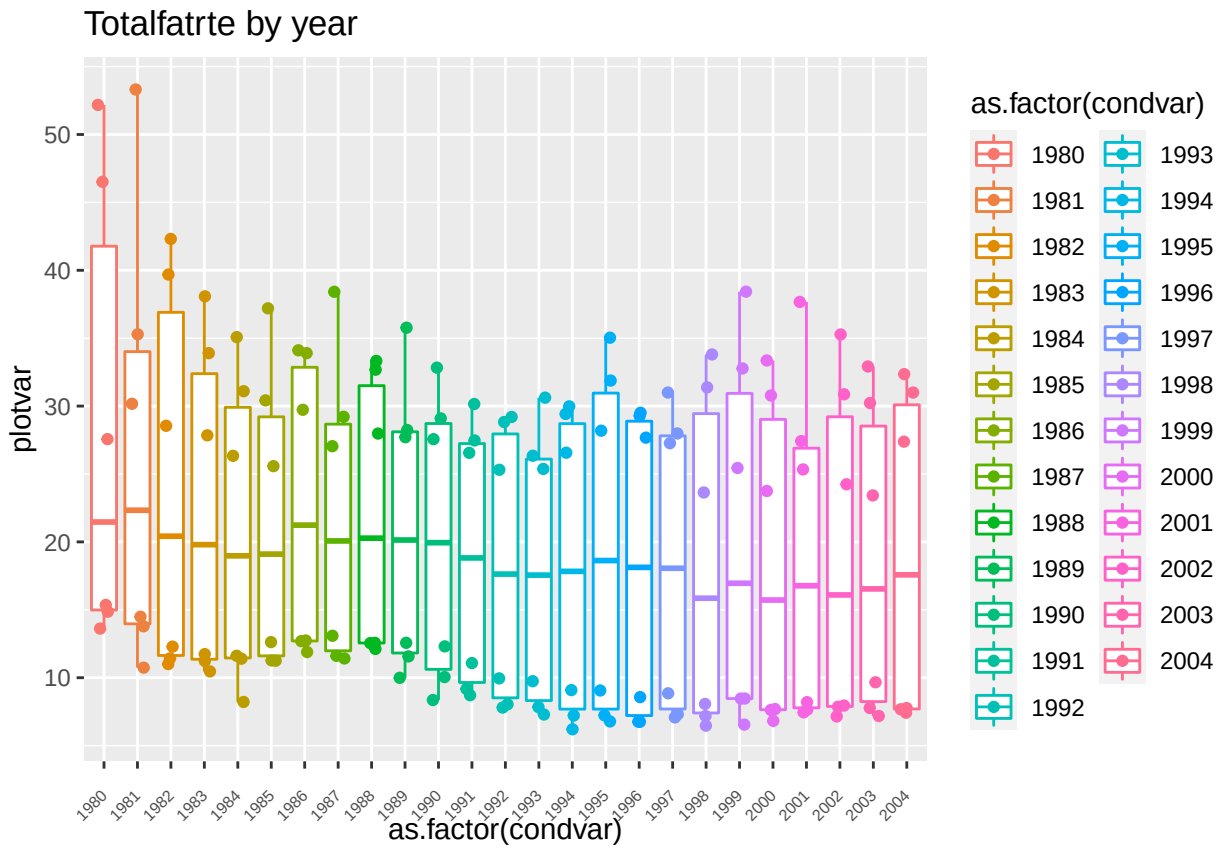
```
ggplot(traffic.state.perc.aggr, aes(shortname, totfatrte)) +
  geom_bar(aes(fill = factor(shortname)), position = "dodge", stat="identity") + ggtitle("Fatal
  scale_fill_hue(c=45,l=80) + theme(plot.title = element_text(size = 8, hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90, size = 6, vjust = 0.5, hjust=1)) +
  geom_line(aes(x = shortname, y = nghtfatrte, color="red"), size = 1.5, group = 1) +
geom_line(aes(x = shortname, y = wkndfatrte, color="blue"), size = 1.5, group = 1) +
  scale_color_discrete(name = "Rates", labels = c("Night Fatality Rate", "Weekend Fatality Rate
```



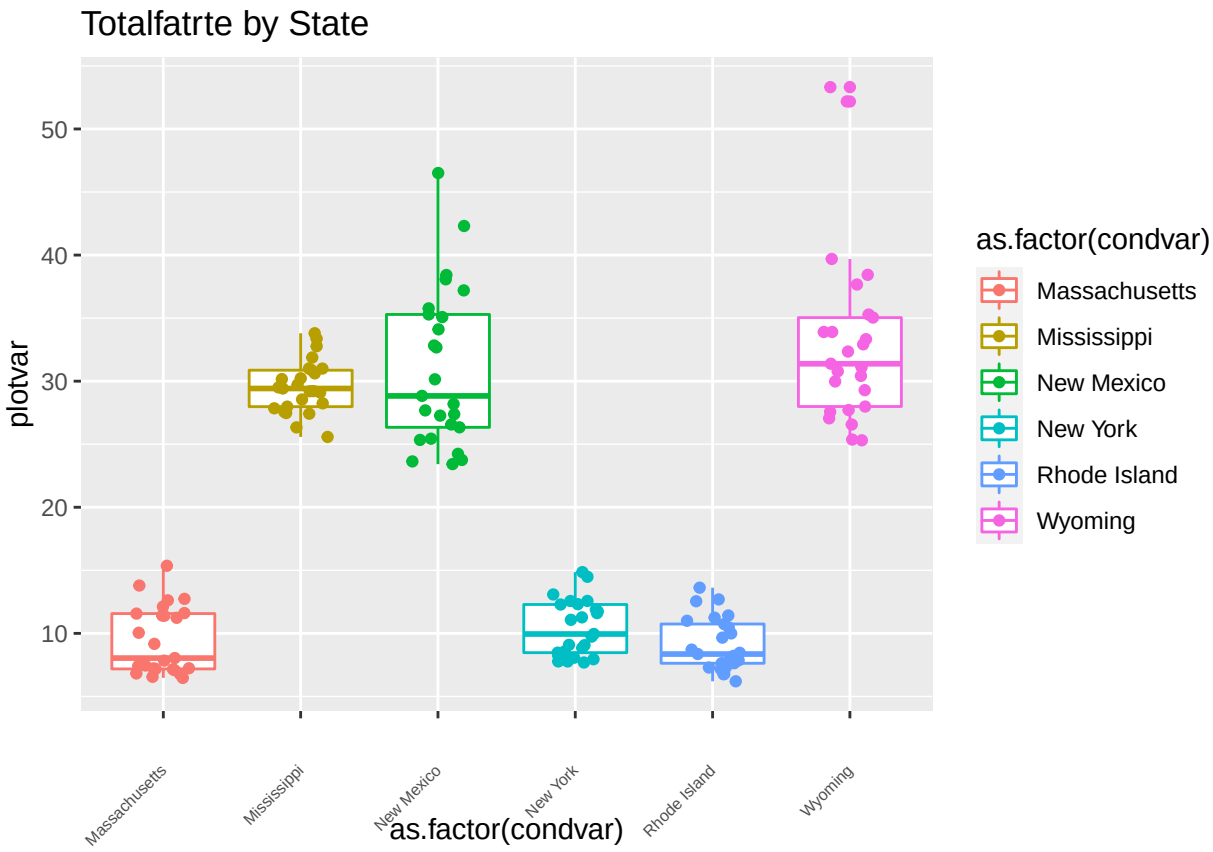

```
# Density
density_plot = function(data, condvar, plotvar, title) {
  ggplot(data, aes(plotvar, group = condvar, color = condvar)) + geom_density() + ggtitle(titl
}

# Conditional Box-plot
conditional_plot = function(data, plotvar, condvar, title) {
  g <- ggplot(data, aes(as.factor(condvar), plotvar, color = as.factor(condvar)))
  g + geom_boxplot() + geom_jitter(width = 0.2) + ggtitle(title) + theme(axis.text.x = element_
}

density_plot(data.merged, data.merged$name, data.merged$totfatrte, "totfatrte")
```



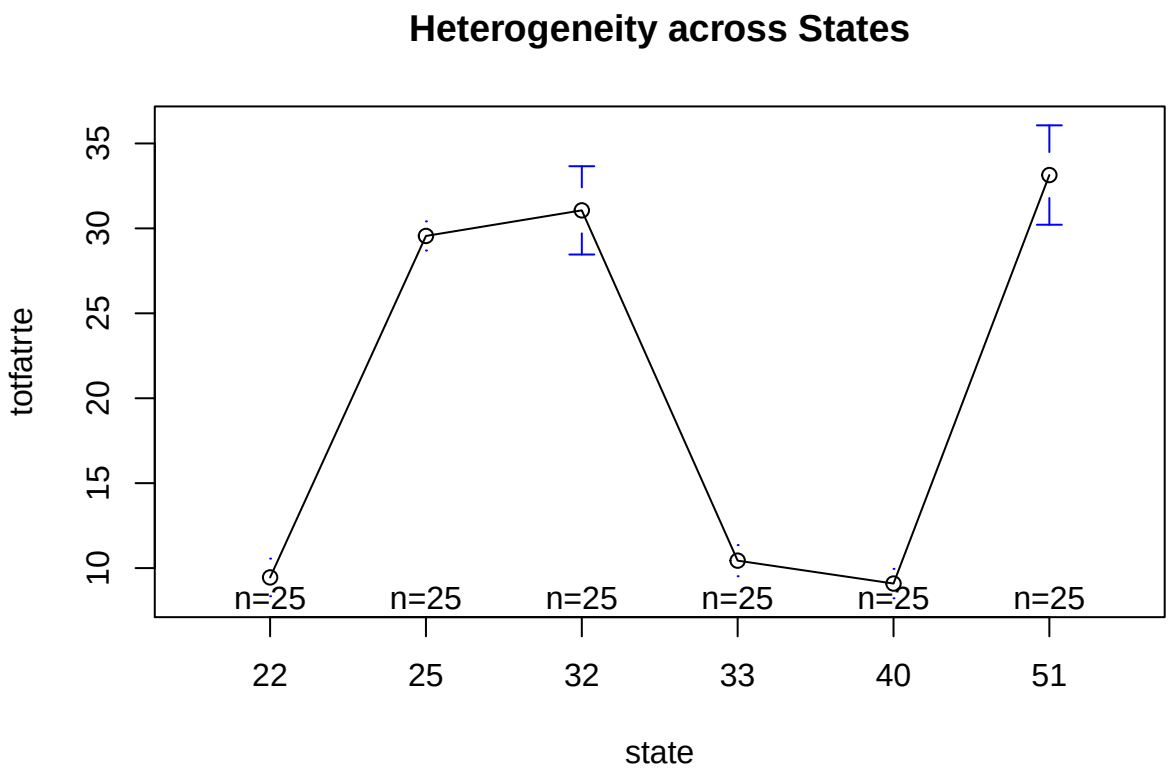

```
# yIndex by year (Heterogeineity across year)
conditional_plot(data.merged, data.merged$totfatrte, data.merged$year, "Totalfatrte by year")
```



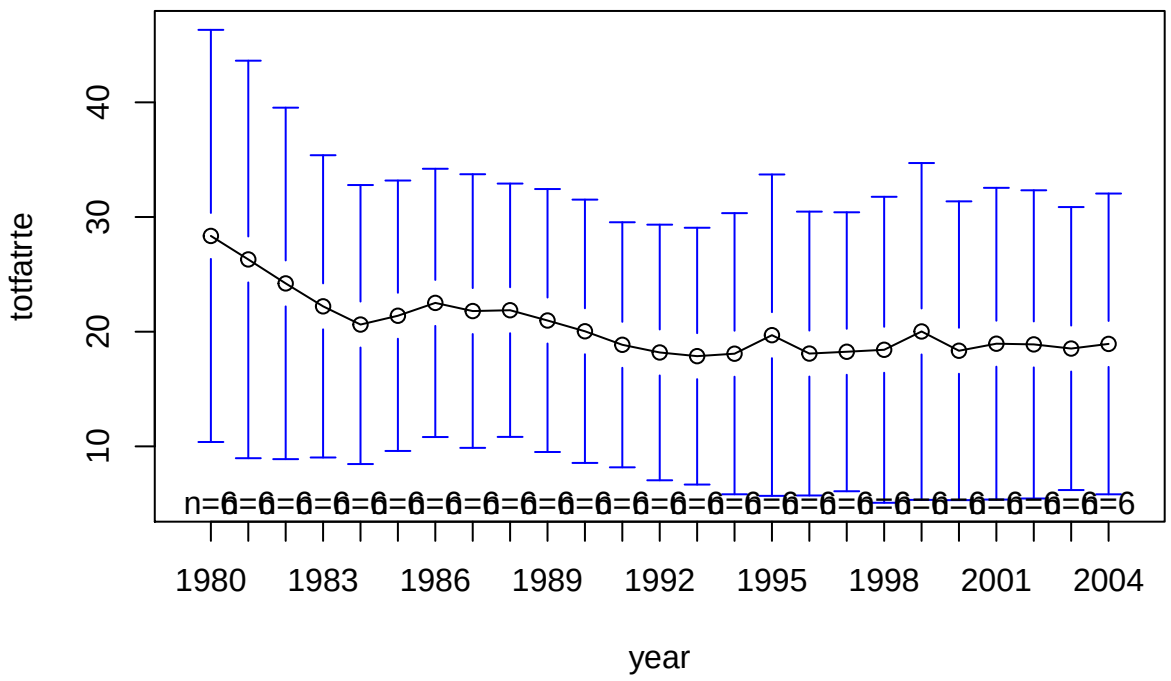

```
# yIndex by country (Heterogeineity across countries)
conditional_plot(data.merged, data.merged$totfatrte, data.merged$name, "Totalfatrte by State")
```




```
#scatterplot(totfatrte ~  shortname|year, boxplots=FALSE, smooth=TRUE, data=data.merged)
#scatterplot(totfatrte ~ df$xIndex | df$country, boxplots=FALSE, xlab="xIndex", ylab="yhat",sm
#abline(lm(data.merged$totfatrte ~ data.merged$shortname),lwd=3, col="blue")

# Heterogeneity across countries
plotmeans(totfatrte ~ state, main="Heterogeneity across States", data=data.merged)
```

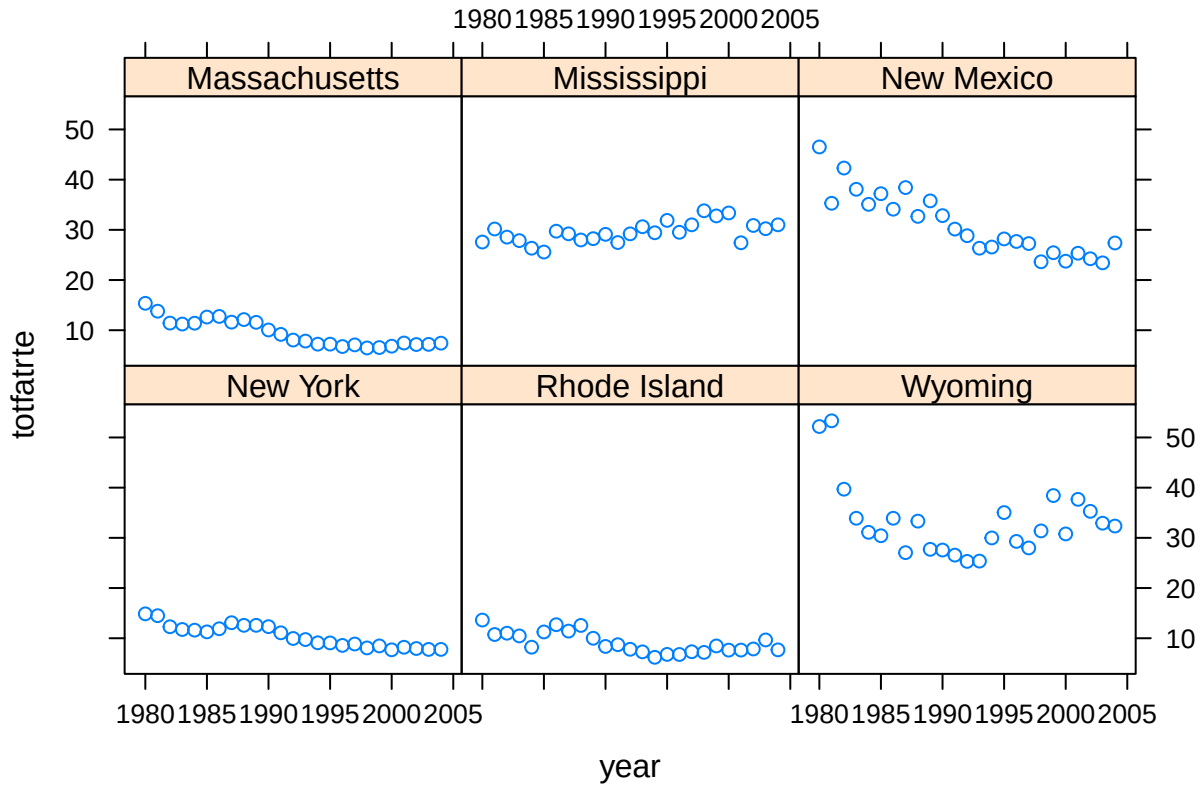



```
plotmeans(totfatrte ~ year, main="Heterogeneity across years", data=data.merged)
```

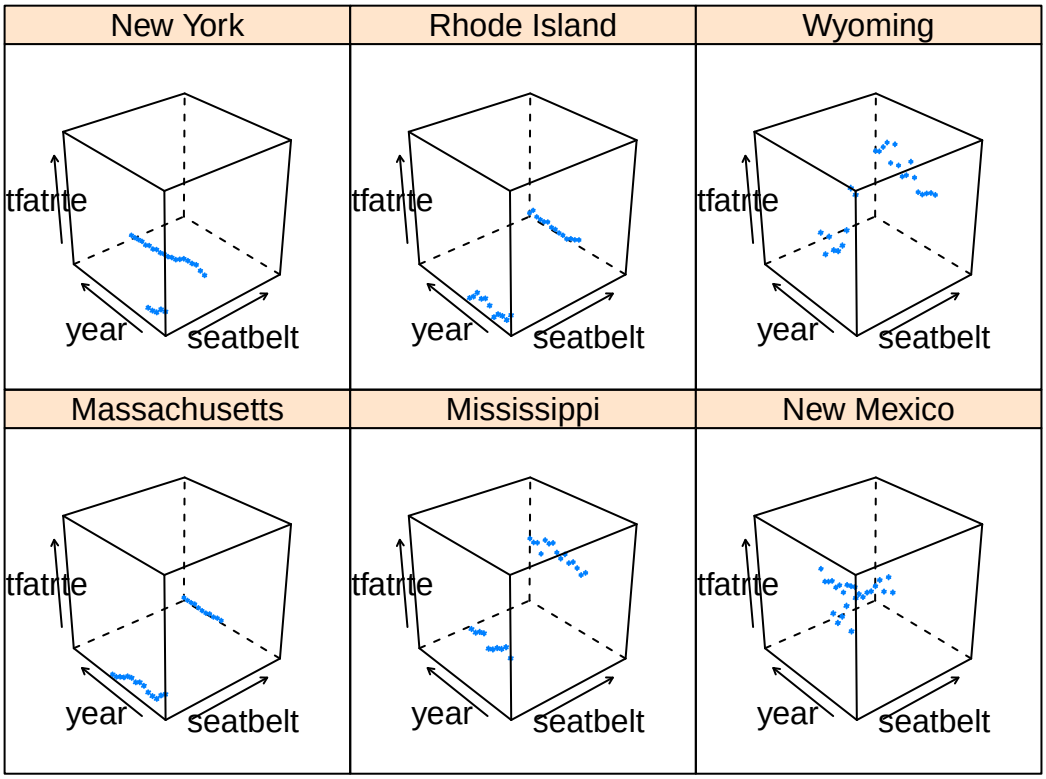



```
#
#coplot(totfatrte ~ year|state, type="l", data=data.merged)
#coplot(totfatrte ~ year|name, type="b", data=data.merged)
```

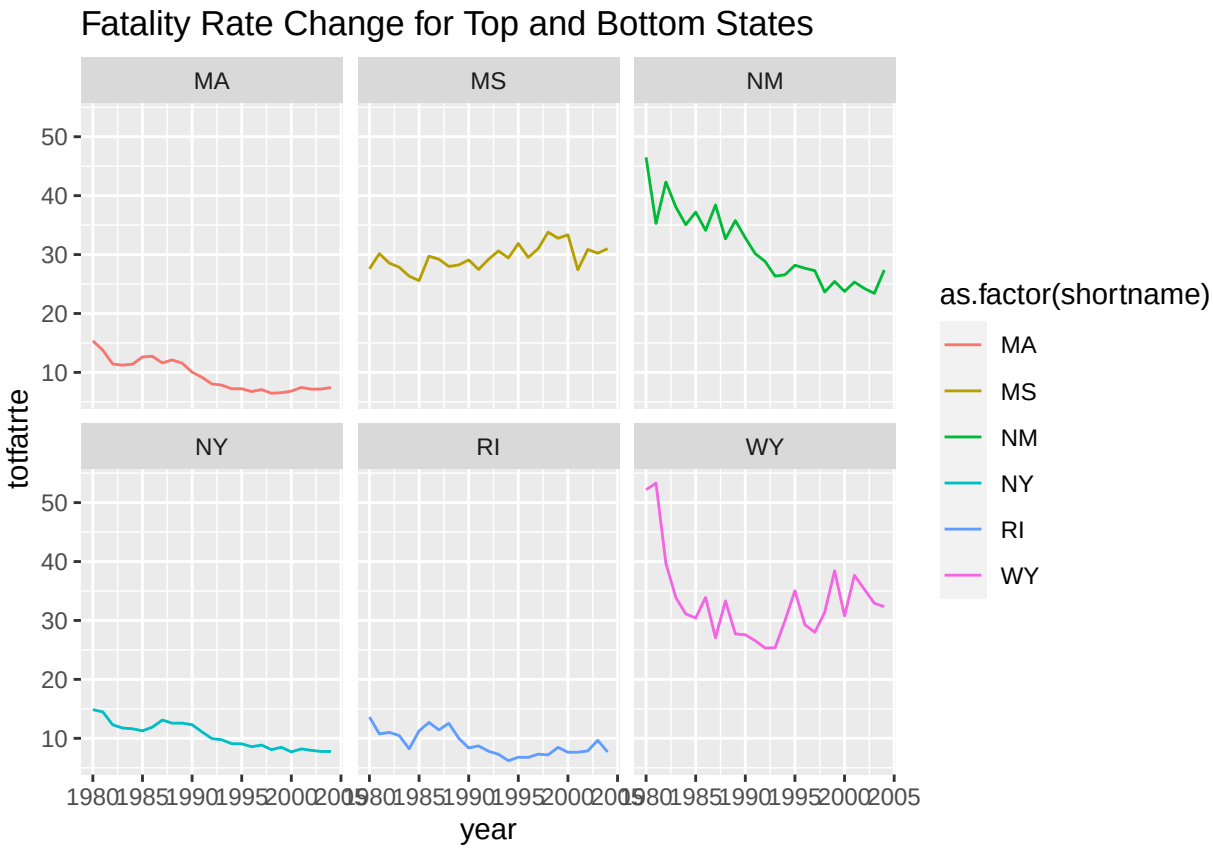
```
xyplot(totfatrte ~ year | name, data=data.merged, as.table=T)
```




```
# check if we need this 3D plot
library(lattice)
cloud(totfatrte ~ seatbelt + year|name, data = data.merged, auto.key = TRUE)
```




```
g <- ggplot(data.merged, aes(year, totfatrte, colour = as.factor(shortname)))
g + geom_line() + ggtitle("Fatality Rate Change for Top and Bottom States") + facet_wrap(~shor
```



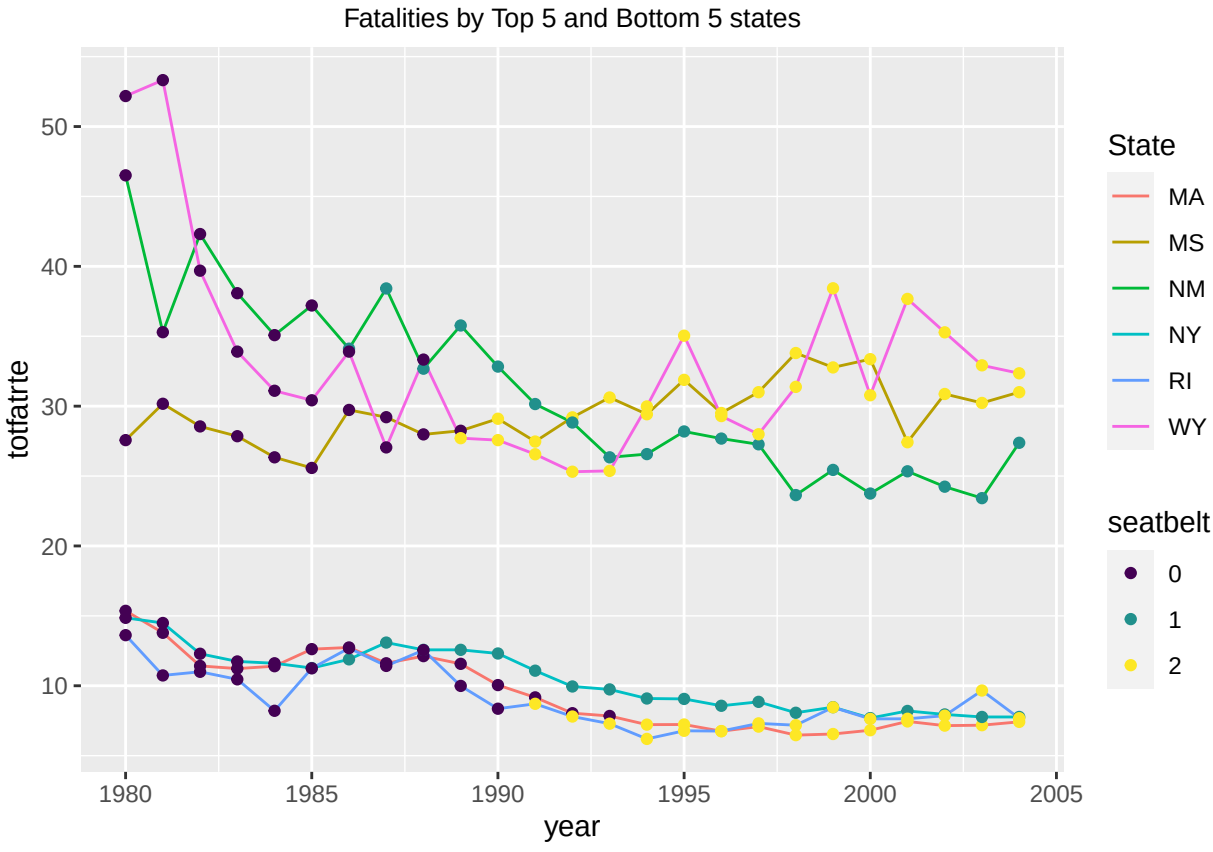
#TODO : The above graphs collectively provide the below information. Some of them are different ways of representing the same info, we need to pick and choose.

We can see that New Mexico and Wyoming has high variance in the data, with NM consistently reducing the traffic fatality rate over years. However, WY reduced the fatality rate from 80's to mid 90's and had a gradual increase after. The bottom 3 states have very low variance across years.

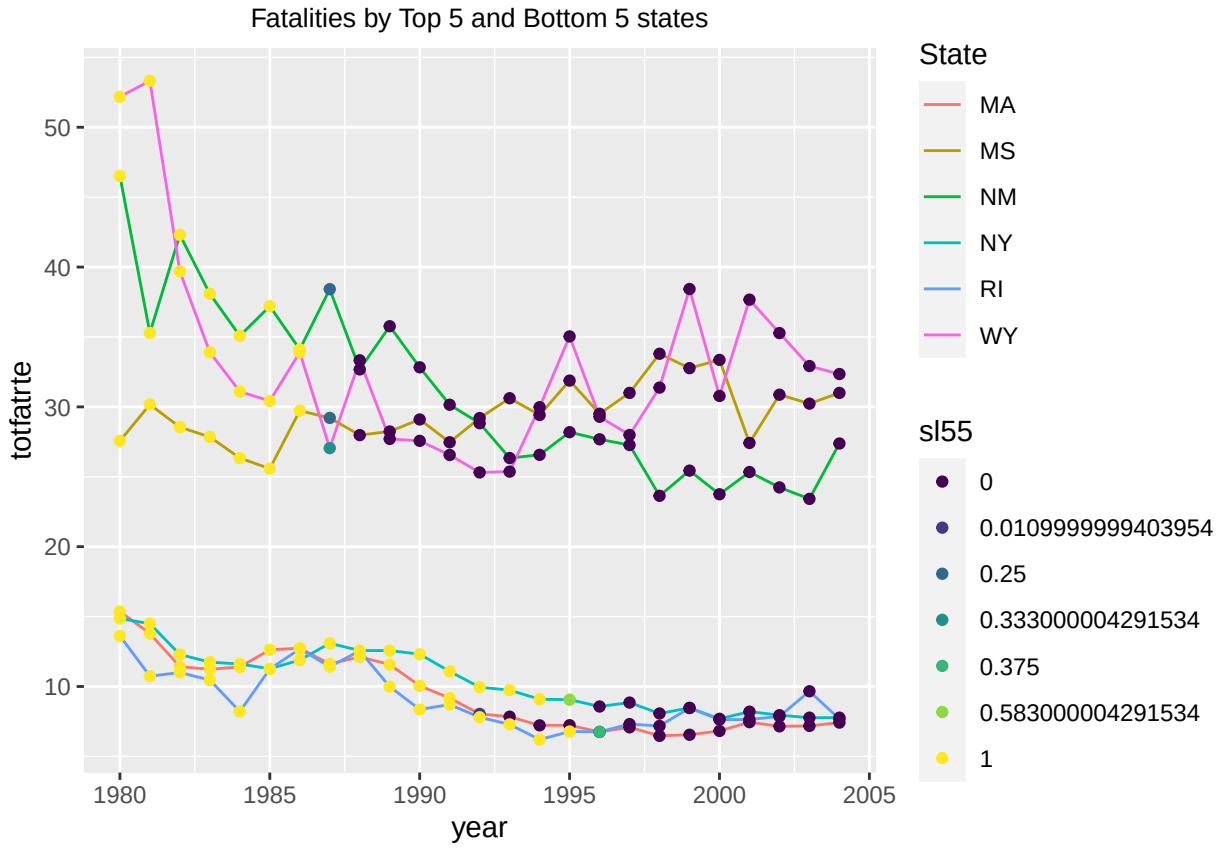
Now we look at how the traffic laws changed over time for the above states. First we create a common function to measure each traffic law.

```
eda.states.chart <- function(fieldname) {

top.plot.1 <- ggplot(data.top.filtered, aes(x = year, y = totfatrte, group = shortname, color =
   ggtitle("Fatalities by state - Top 5") +
  geom_line(aes(x = year,y=totfatrte)) + theme_gray() +
  ggnewscale::new_scale_colour() +
  geom_point(aes(colour = factor(get(fieldname)))) + scale_color_discrete(name = fieldname)

bottom.plot.1 <- ggplot(data.bottom.filtered, aes(x = year, y = totfatrte, color = factor(short
  theme(plot.title = element_text(size = 10, hjust = 0.5)) +  scale_color_viridis(discrete = TR
  geom_line(aes(x = year,y=totfatrte)) +
  ggnewscale::new_scale_colour() +
  geom_point(aes(colour = factor(get(fieldname)))) + scale_color_discrete( name = fieldname)

plot.1 <- ggplot(data.merged, aes(x = year, y = totfatrte, group = shortname, color = shortname
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + scale_color_discrete(name = "State
```

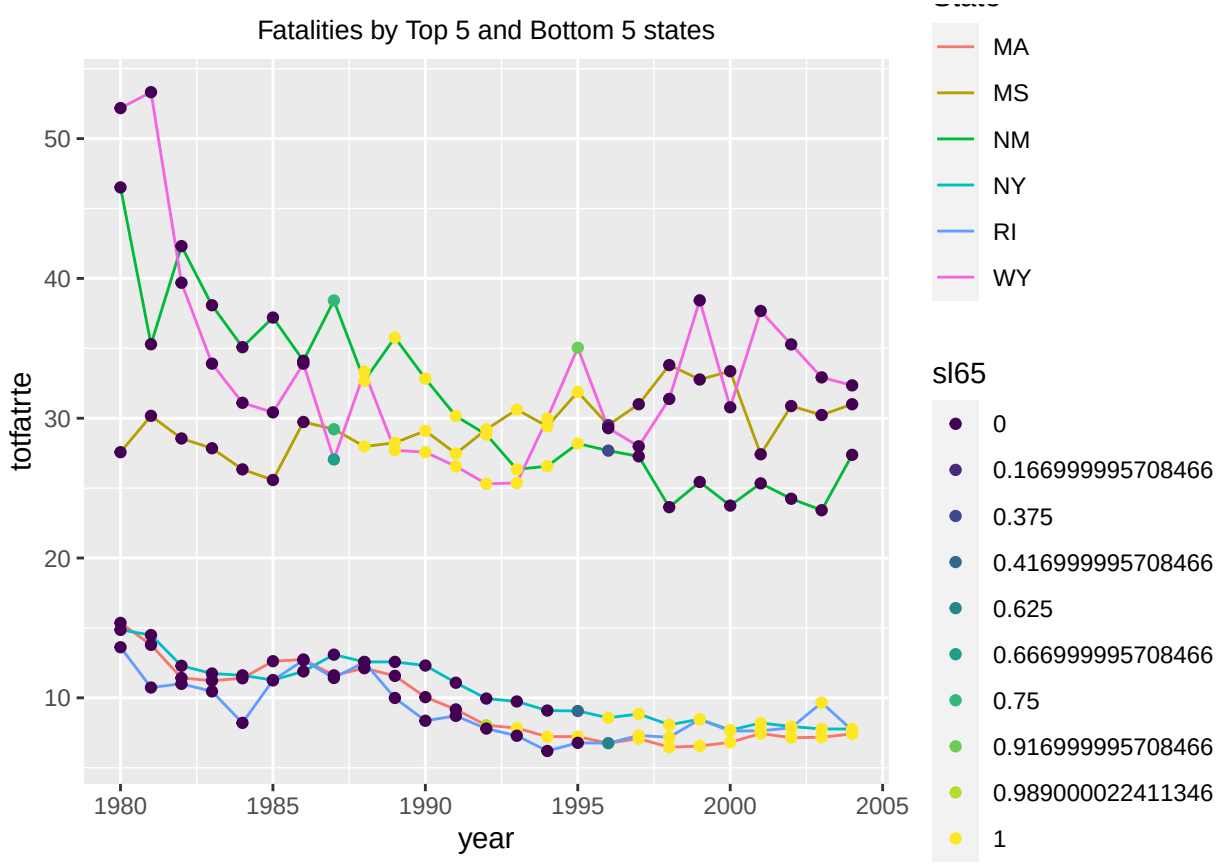
```
  geom_line(aes(x = year,y=totfatrte)) +
  ggnewscale::new_scale_colour() +
  geom_point(aes(colour = factor(get(fieldname)))) + scale_color_viridis(discrete = TRUE, optio

return(plot.1)
#top.plot.1
#bottom.plot.1
#
# grid.arrange(plot.1, top.plot.1, bottom.plot.1,
#                nrow = 1, heights = c(20), ncol = 3)

}

eda.states.chart("seatbelt")
```



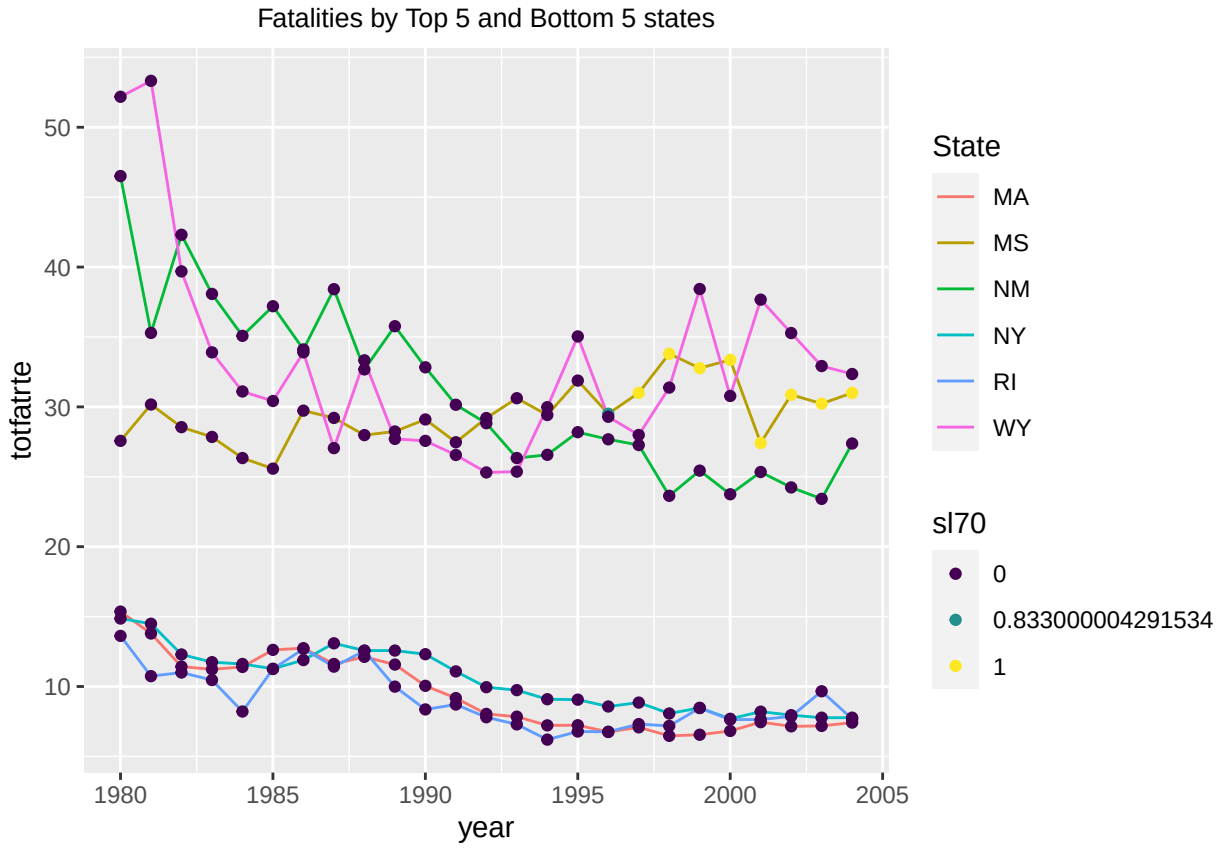

```
eda.states.chart("sl55")
```



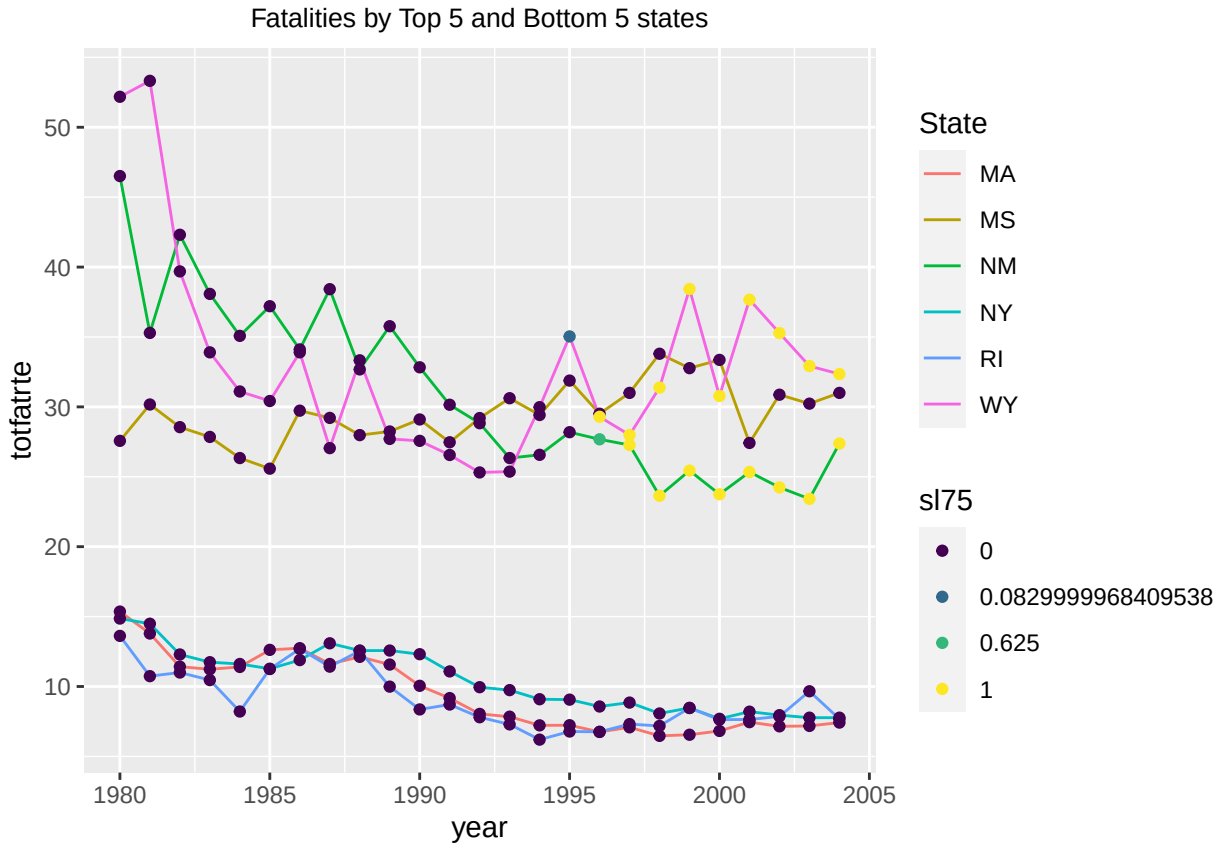

```
eda.states.chart("sl65")
```



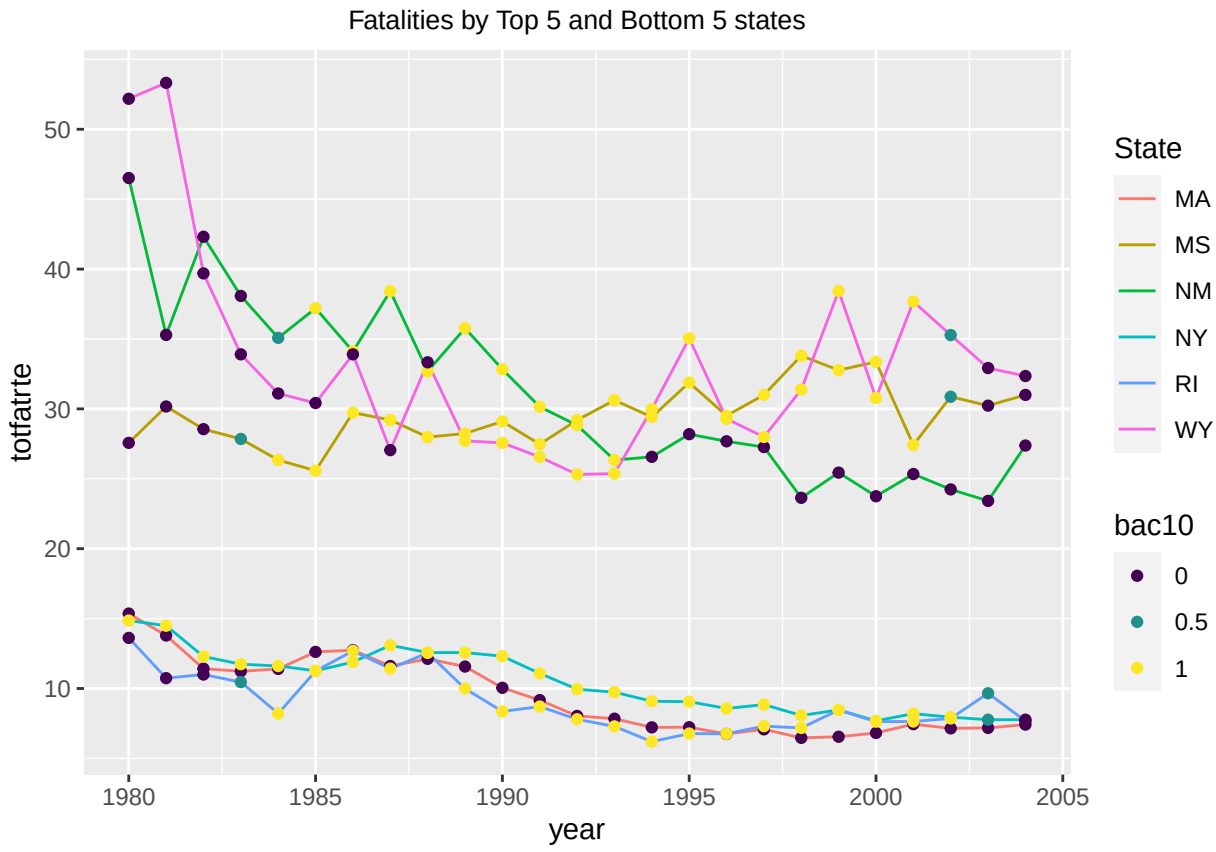

```
eda.states.chart("sl70")
```



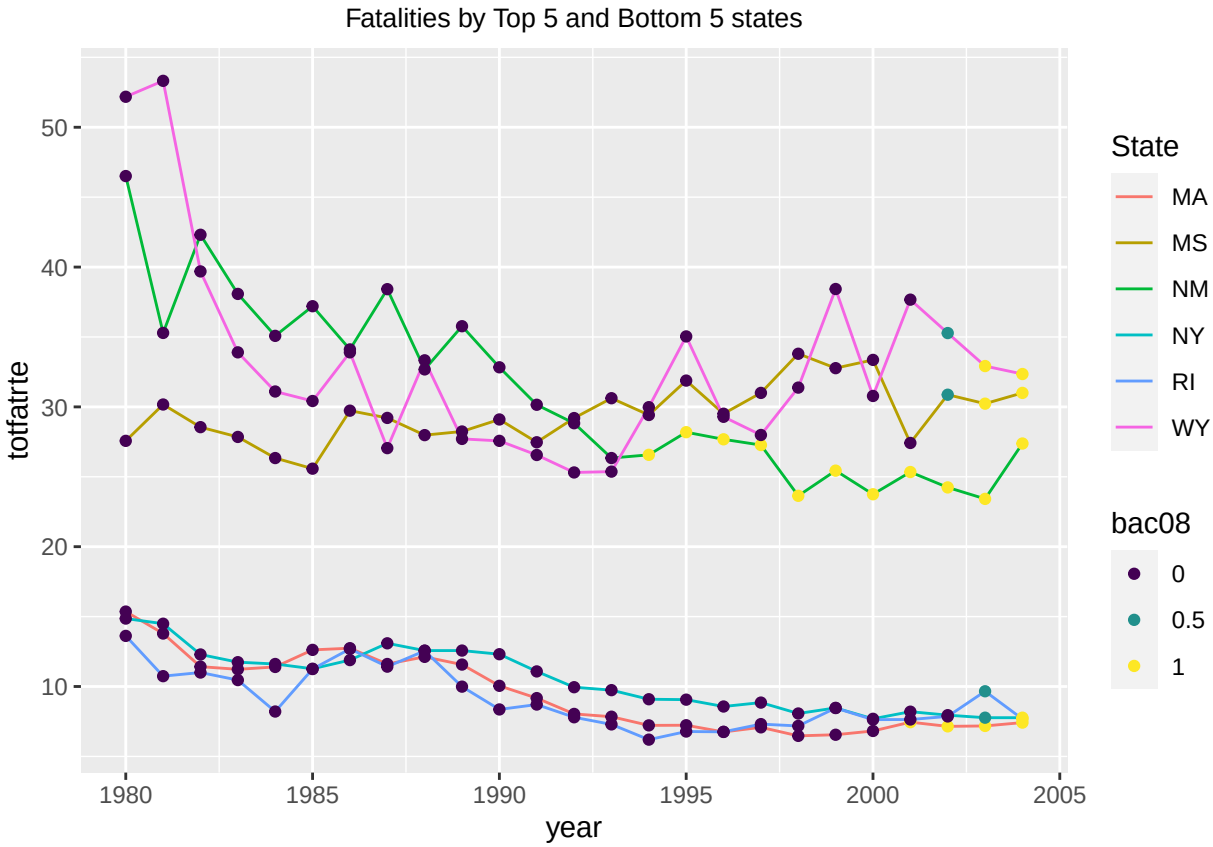

```
eda.states.chart("sl75")
```




```
eda.states.chart("bac10")
```




```
eda.states.chart("bac08")
```




```
# grid.arrange(plot.1, top.plot.1, bottom.plot.1,
#              nrow = 3, heights = c(20,20,20), ncol = 1)
```

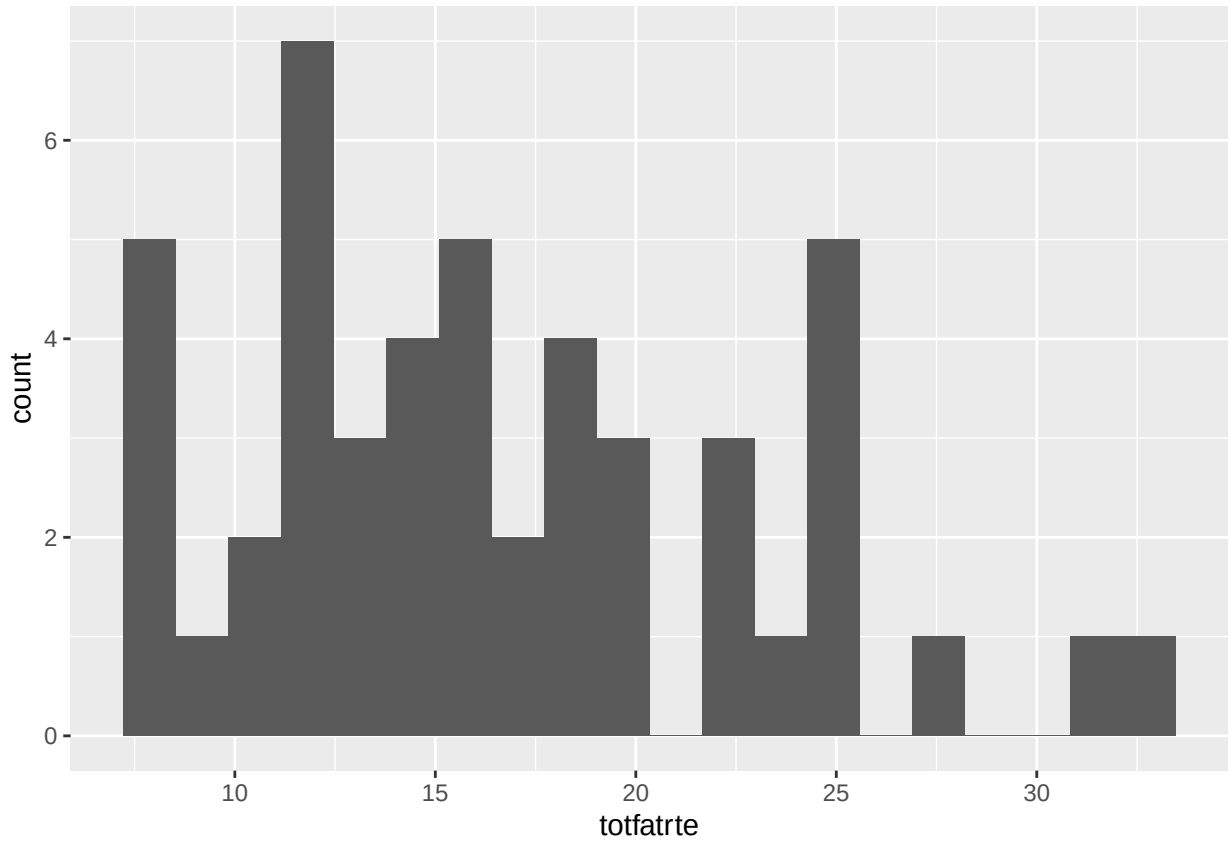
The top 2 states WY and NM have a speed limit of 75mph while Mississippi has a speed limit of 70 mph. In contrast, all the bottom 3 states, NY, MA and RI, have a speed limit of 65 mph. It can also be seen that WY see an upward trend in fatalities from the time they changed the speed limit from 65mph to 75 mph, in 1995.

All 6 states require secondary seatbelt, and limit the maximum permitted blood alcohol at 0.08 while driving, as of 2004.

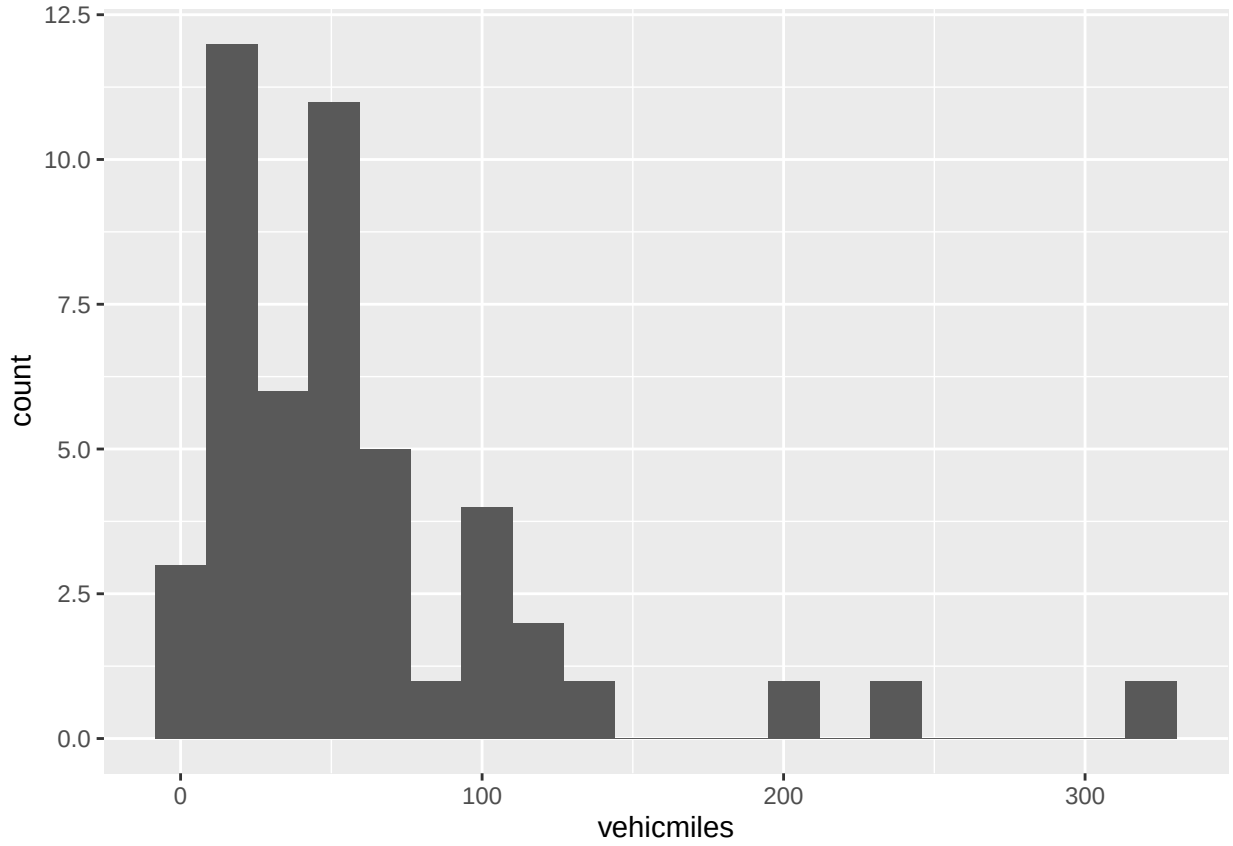
Now we proceed to examine individual cross sections for the first and the last years in the dataset - 1980 and 2004.

```
driving_2004.df <- driving.df %>% filter(year == 2004)
driving_1980.df <- driving.df %>% filter(year == 1980)
```

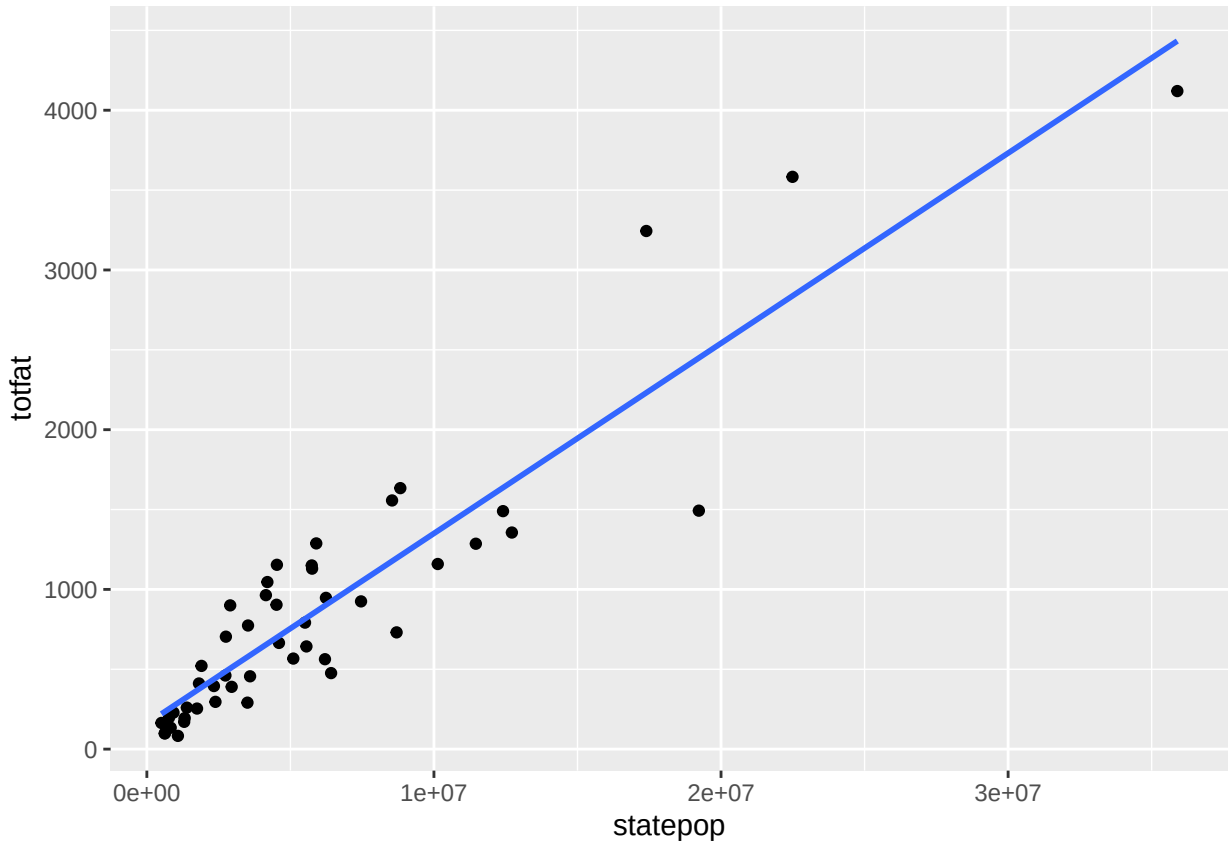
```
driving_2004.df %>%
  ggplot(aes(x = totfatrte)) +
  geom_histogram(bins = 20)
```



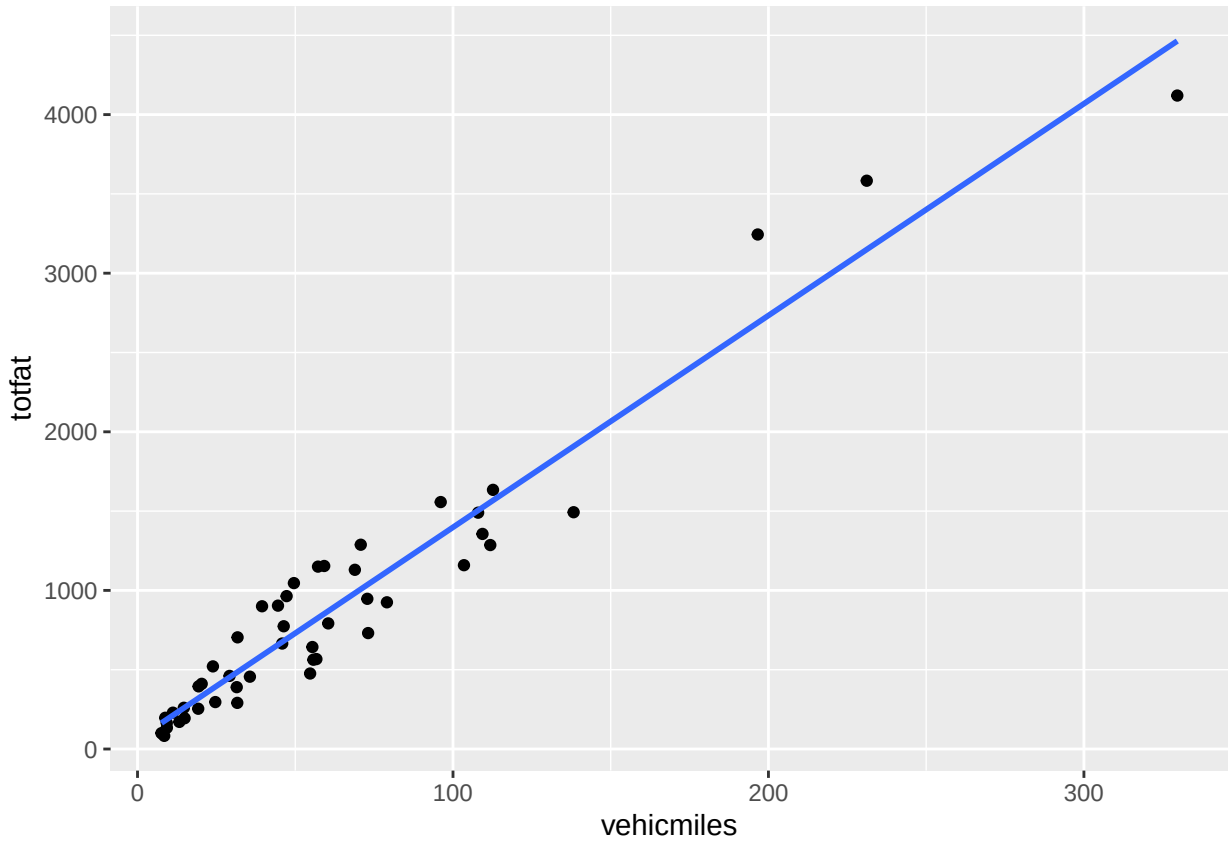

```
driving_2004.df %>%
  ggplot(aes(x = vehicmiles)) +
  geom_histogram(bins = 20)
```



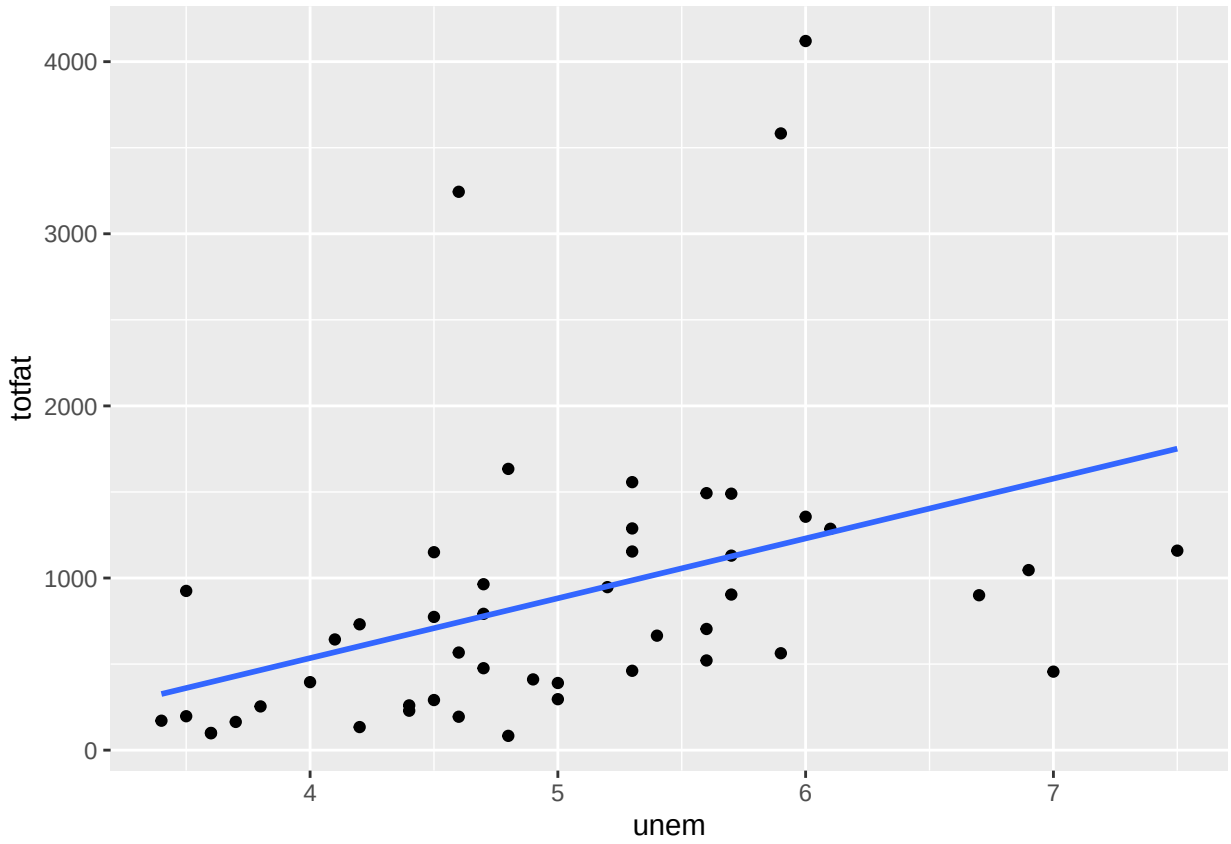

```
driving_2004.df %>%
  ggplot(aes(statepop, totfat)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



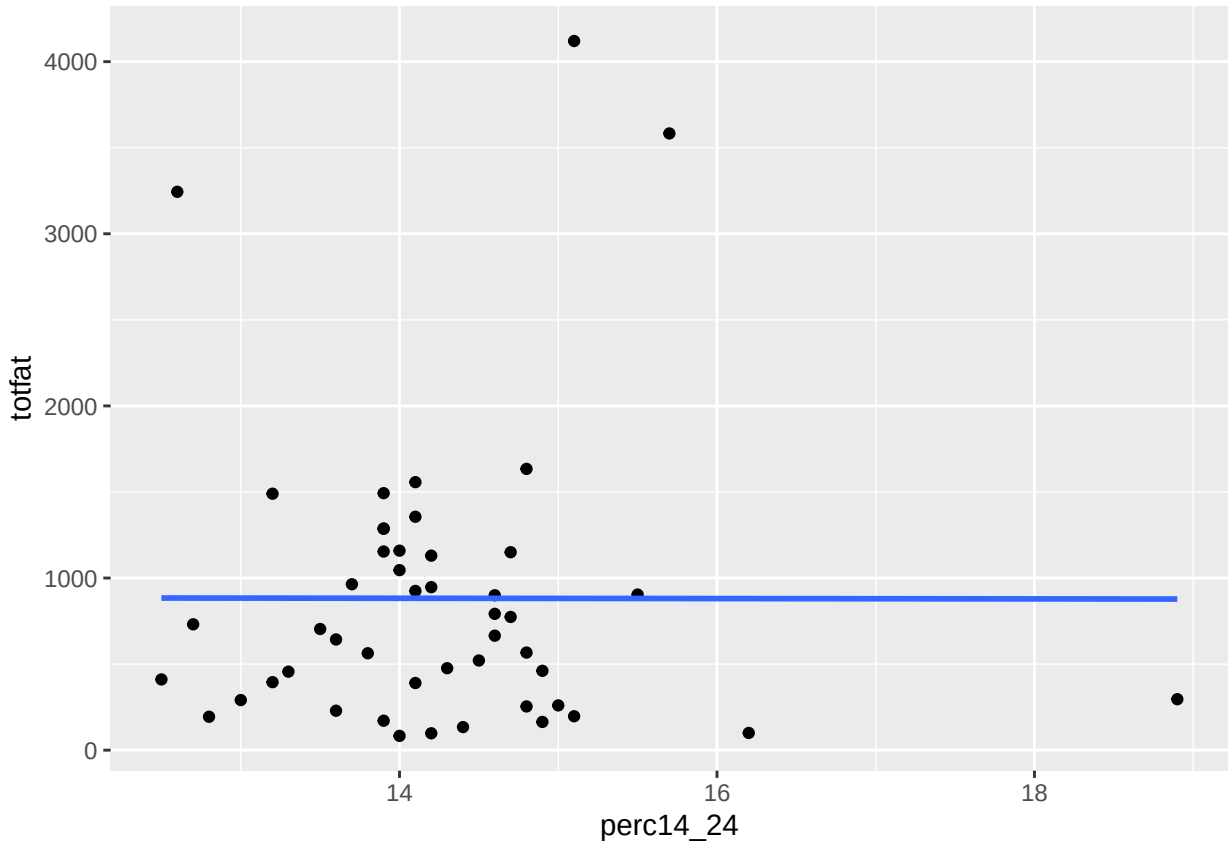

```
driving_2004.df %>%
  ggplot(aes(vehicmiles, totfat)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



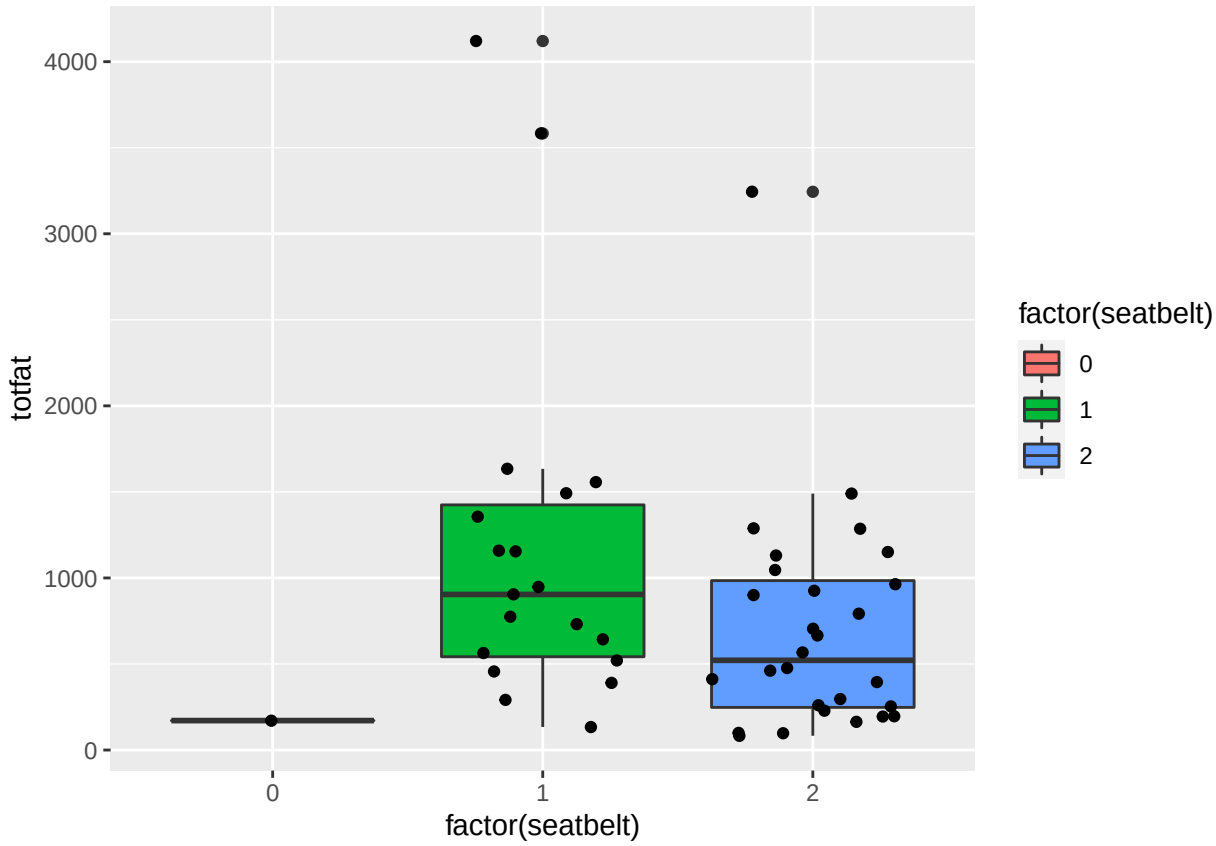

```
driving_2004.df %>%
  ggplot(aes(unem, totfat)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



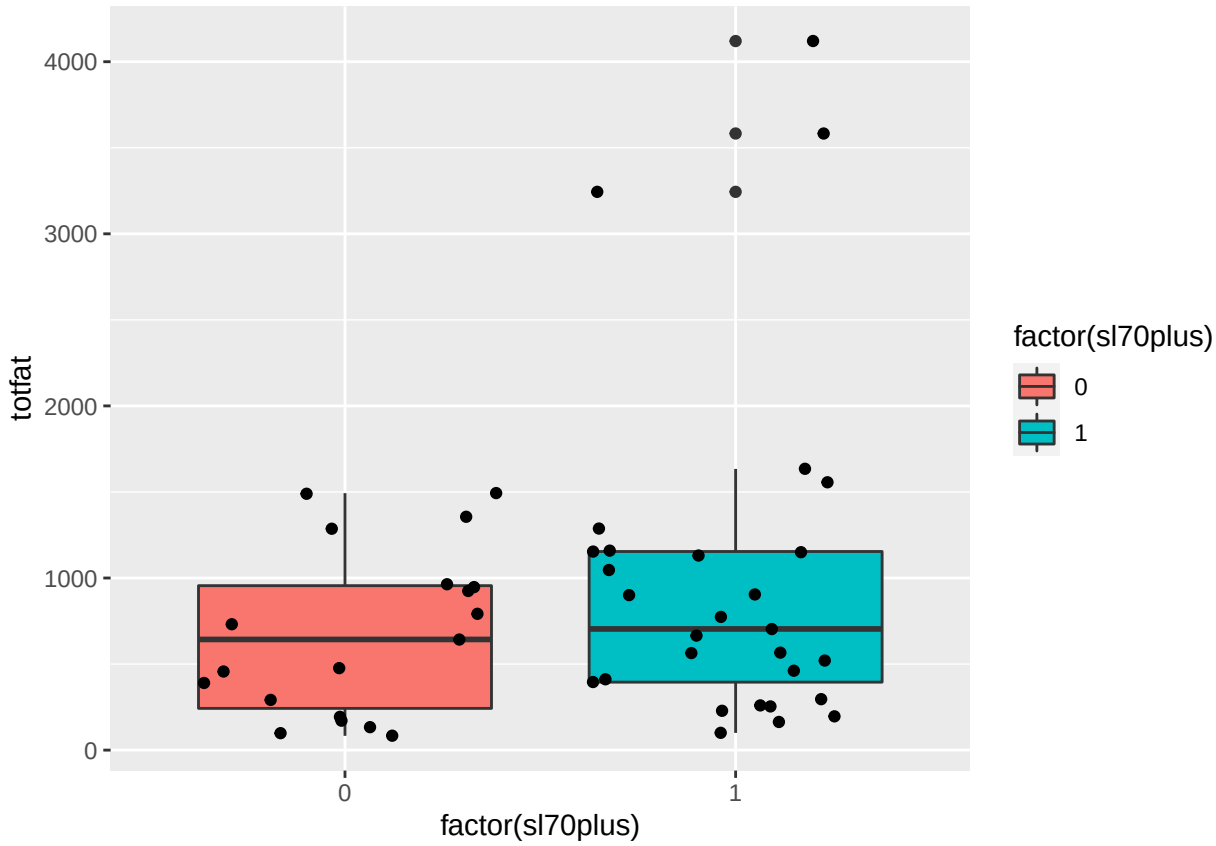

```
driving_2004.df %>%
  ggplot(aes(perc14_24, totfat)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



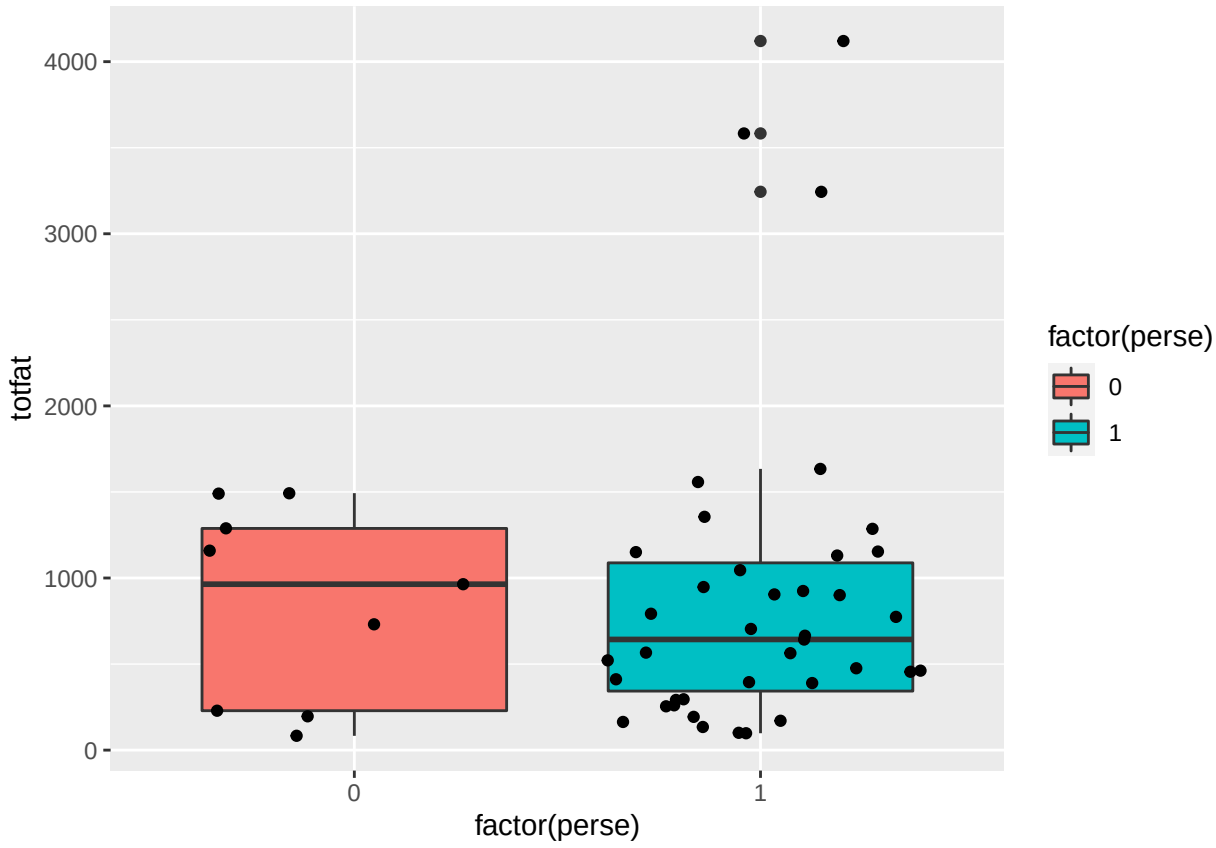

```
driving_2004.df %>%
  ggplot(aes(x = factor(seatbelt), y = totfat)) +
  geom_boxplot(aes(fill = factor(seatbelt))) +
  geom_jitter()
```



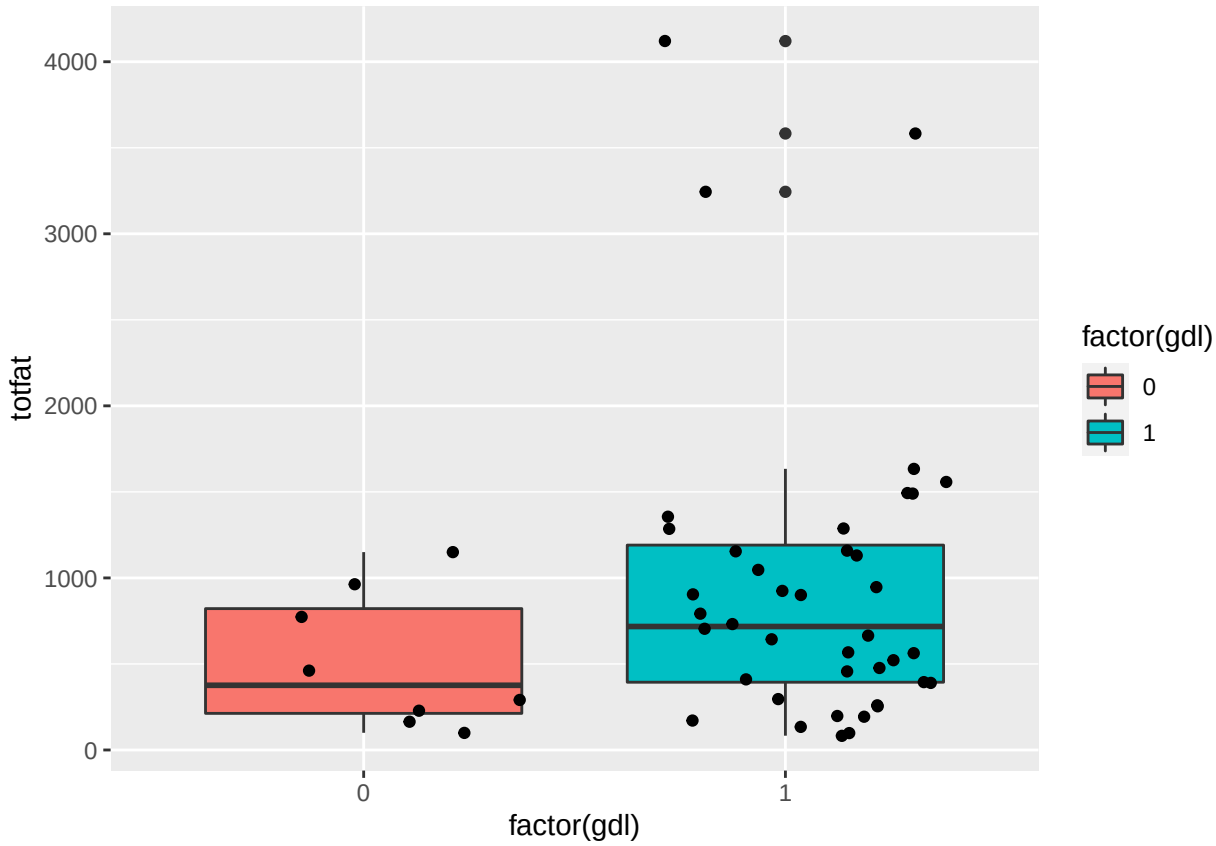

```
driving_2004.df %>%
  ggplot(aes(x = factor(sl70plus), y = totfat)) +
  geom_boxplot(aes(fill = factor(sl70plus))) +
  geom_jitter()
```



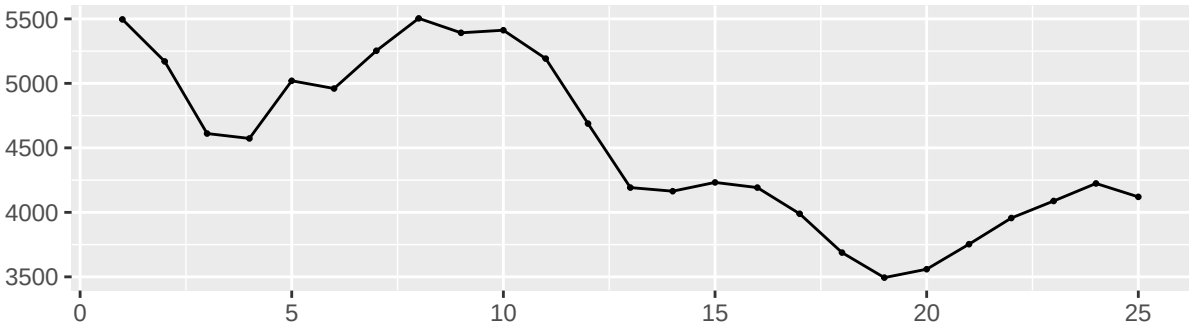

```
driving_2004.df %>%
  ggplot(aes(x = factor(perse), y = totfat)) +
  geom_boxplot(aes(fill = factor(perse))) +
  geom_jitter()
```

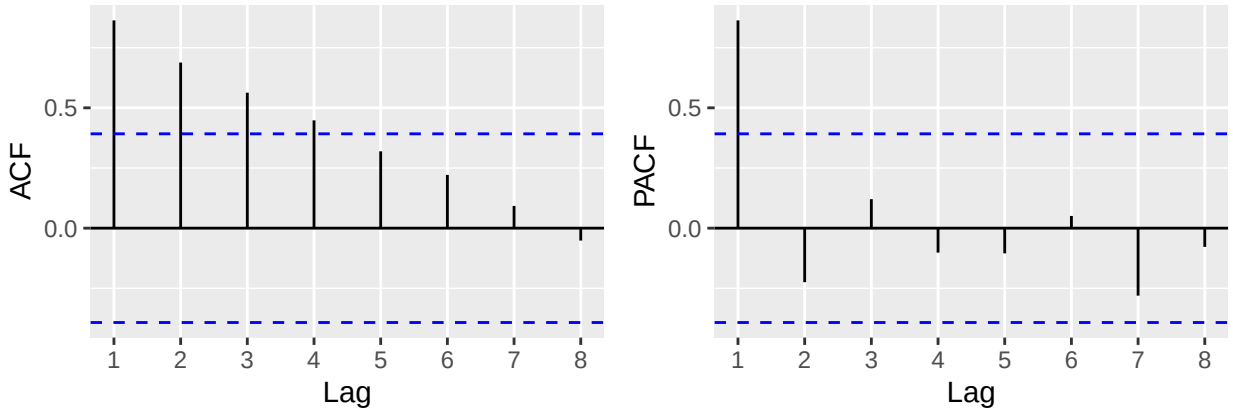



```
driving_2004.df %>%
  ggplot(aes(x = factor(gdl), y = totfat)) +
  geom_boxplot(aes(fill = factor(gdl))) +
  geom_jitter()
```

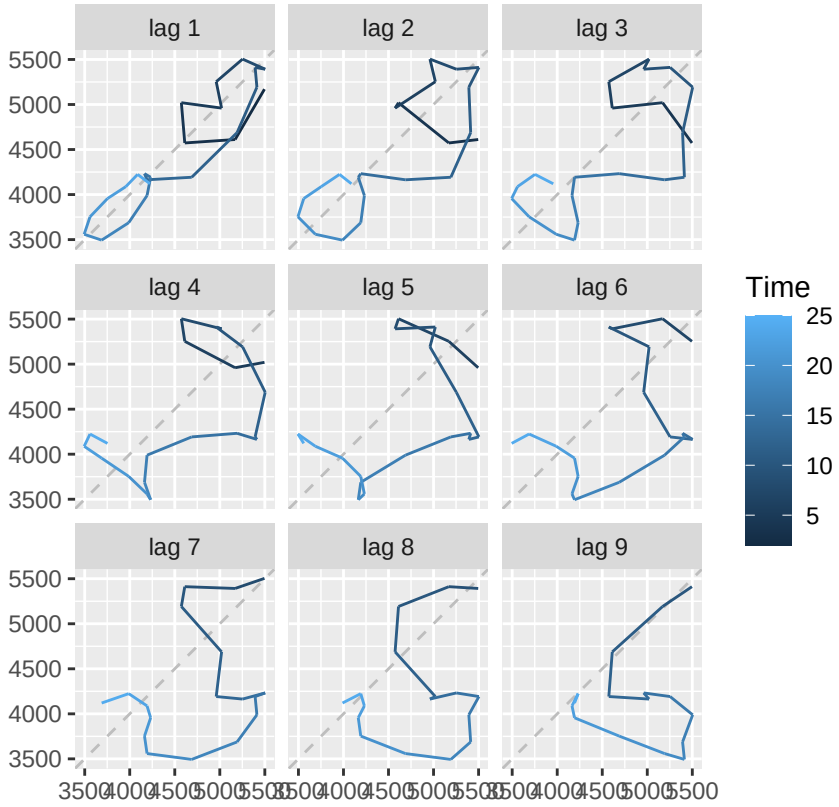



```
#Quick EDA on Time Series Data for State 1
driving_5.tsibble <- driving.df %>% filter(state == 5) %>% as_tsibble(index = year)

driving_5.tsibble$totfat %>% ggtsdisplay()
```

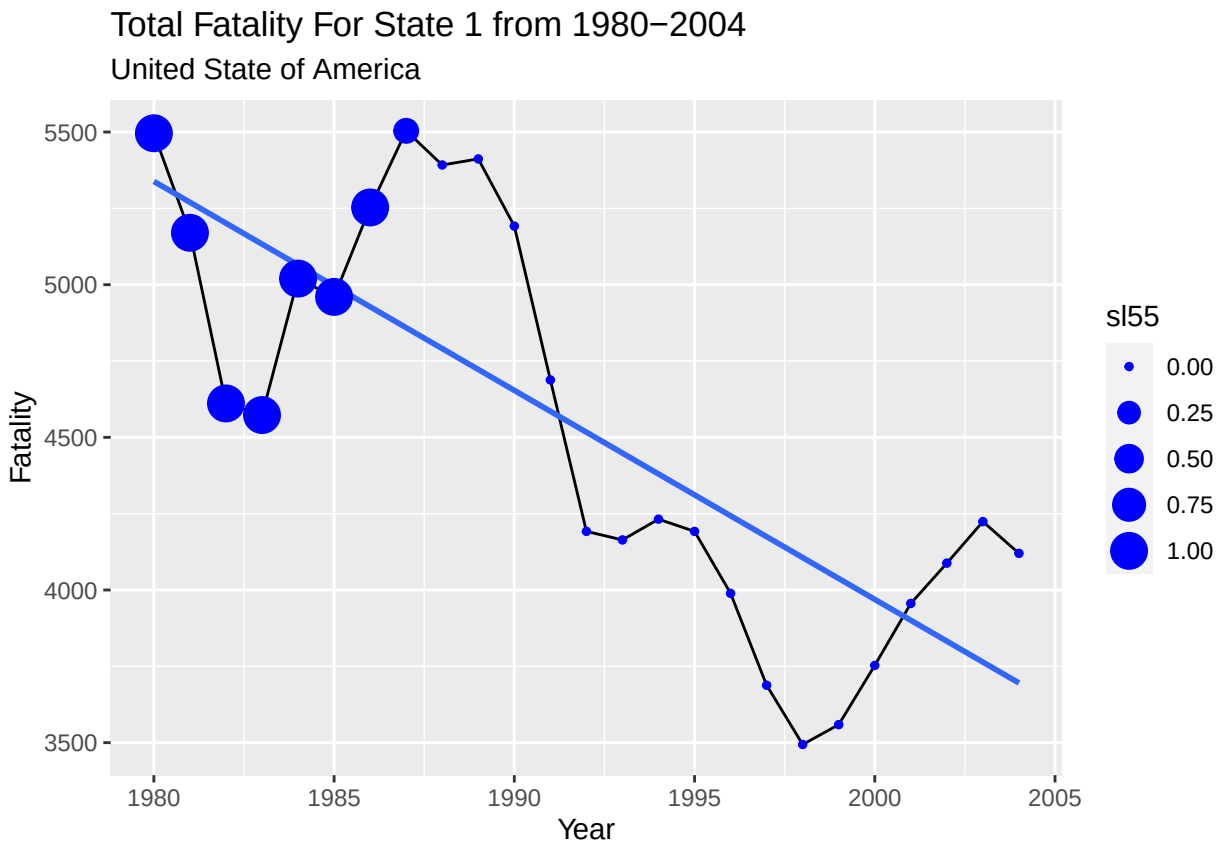





```
driving_5.tsibble$totfat %>% gglagplot()
```



```r
driving_5.tsibble %>%
  ggplot(aes(year, totfat)) +
  geom_line() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  geom_point(aes(x = year, y = totfat, size = sl55), color = "blue") +
  labs(
    title = "Total Fatality For State 1 from 1980-2004",
    subtitle = "United State of America",
    y = "Fatality",
    x = "Year"
    )
```



#TODO rearrange. Describe about all graphs. Move histograms to appendix.

2. (15%) How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.

3. (15%) Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables*. Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the

coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

4. (15%) Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?

5. (10%) Would you perfer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

6. (10%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by $1,000$. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

7. (5%) If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?

#TODO in EDA mention about omitted variables - for example advances in medical field, ER procedures. Also compliance to traffic laws - seatbelt, drunk driving. Also Road condition. For example Wyoming and SC have large rural roads.

#TODO in conclusion The researchers at the Ann Arbor based institute found that while road fatalities were lower than other causes of death, they were still considerable. Michael Sivak and Brandon Schoettle, authors of the report "Mortality from road crashes in the individual U.S. states: A comparison with leading causes of death," found that nationwide, death from road crashes per 100,000 population was 10.9, compared to 34.4 for Alzheimer's, 43.7 for stroke, 48.2 for lung disease, 185.4 from cancer and 197.2 from heart disease. They used 2015 data from the National Highway Traffic Safety Administration and the Centers for Disease Control and Prevention Traffic fatalities from crashes in the United States are lower than other leading causes of death, but still substantial. The Northeast may be the safest region in the country on the roads; Wyoming is the most dangerous state.

correlation