Lab 3: Panel Models

US Traffic Fatalities: 1980 - 2004

U.S. traffic fatalities: 1980-2004

In this lab, we are asking you to answer the following **causal** question:

> **"Do changes in traffic laws affect traffic fatalities?"**

To answer this question, please complete the tasks specified below using the data provided in `data/driving.Rdata`. This data 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 also provided in the dataset.

```r
load(file="./data/driving.RData")

# Description of the data
desc # description
```

```
##        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
```

(30 points, total) Build and Describe the Data

1. (5 points) Load the data and produce useful features. Specifically:

    - Produce a new variable, called `speed_limit` that re-encodes the data that is in `sl55`, `sl65`, `sl70`, `sl75`, and `slnone`;
    - Produce a new variable, called `year_of_observation` that re-encodes the data that is in `d80`, `d81`, ... , `d04`.
    - Produce a new variable for each of the other variables that are one-hot encoded (i.e. `bac*` variable series).
    - Rename these variables to sensible names that are legible to a reader of your analysis. For example, the dependent variable as provided is called, `totfatrte`. Pick something more sensible, like, `total_fatalities_rate`. There are few enough of these variables to change, that you should change them for all the variables in the data. (You will thank yourself later.)

```
# Produce variable speed_limit
data_clean <- data %>%
  mutate(speed_limit = case_when(
    sl55 > 0.5 ~ 55,
    sl65 >= 0.5 ~ 65,
    sl70 >= 0.5 ~ 70,
    sl75 >= 0.5 ~ 75,
    slnone > 0.5 ~ NA_integer_,
    TRUE ~ NA_real_
  ))

# Produce variable year_of_observation
data_clean <- data_clean %>%
  mutate(year_of_observation = case_when(
    d80 == 1 ~ 1980,
    d81 == 1 ~ 1981,
    d82 == 1 ~ 1982,
    d83 == 1 ~ 1983,
    d84 == 1 ~ 1984,
    d85 == 1 ~ 1985,
    d86 == 1 ~ 1986,
    d87 == 1 ~ 1987,
    d88 == 1 ~ 1988,
    d89 == 1 ~ 1989,
    d90 == 1 ~ 1990,
    d91 == 1 ~ 1991,
    d92 == 1 ~ 1992,
    d93 == 1 ~ 1993,
    d94 == 1 ~ 1994,
    d95 == 1 ~ 1995,
    d96 == 1 ~ 1996,
    d97 == 1 ~ 1997,
    d98 == 1 ~ 1998,
    d99 == 1 ~ 1999,
    d00 == 1 ~ 2000,
    d01 == 1 ~ 2001,
    d02 == 1 ~ 2002,
    d03 == 1 ~ 2003,
    d04 == 1 ~ 2004,
    TRUE ~ NA_real_
  ))

# Produce variable for bac*
data_clean <- data_clean %>%
  mutate(bac = case_when(
    bac10 >= 0.5 ~ 0.10,
    bac08 >= 0.5 ~ 0.08,
    TRUE ~ NA_real_
  ))


# Also convert bac* to indicator variables
data_clean <- data_clean %>%
  mutate(bac08 = case_when(
    bac08 >= 0.5 ~ 1,
    bac08 < 0.5 ~ 0,
    TRUE ~ NA_real_
  ))

# Also convert bac* to indicator variables
data_clean <- data_clean %>%
  mutate(bac10 = case_when(
    bac10 >= 0.5 ~ 1,
    bac10 < 0.5 ~ 0,
    TRUE ~ NA_real_
  ))


# Organize gdl
data_clean <- data_clean %>%
  mutate(gdl = case_when(
    gdl >= 0.5 ~ 1,
    gdl < 0.5 ~ 0,
    TRUE ~ NA_real_
  ))


# Organize sl70plus
data_clean <- data_clean %>%
  mutate(sl70plus = case_when(
    sl70plus >= 0.5 ~ 1,
    sl70plus < 0.5 ~ 0,
    TRUE ~ NA_real_
  ))


# Organize per_se_law
data_clean <- data_clean %>%
  mutate(perse = case_when(
    perse >= 0.5 ~ 1,
    perse < 0.5 ~ 0,
    TRUE ~ NA_real_
  ))

# Remove those rows
data_clean <- subset(data_clean, select = -c(sl55, sl65, sl70, sl75, slnone,
                                             d80, d81, d82, d83, d84,
                                             d85, d86, d87, d88, d89,
                                             d90, d91, d92, d93, d94,
                                             d95, d96, d97, d98, d99,
                                             d00, d01, d02, d03, d04
                                             ))


# Rename variables
data_clean <- data_clean %>%
  dplyr::rename(
    total_fatalities_rate = totfatrte,
    night_fatalities_rate = nghtfatrte,
    weekend_fatalities_rate = wkndfatrte,
    vehicle_miles = vehicmiles,
    unemployment_rate = unem,
    percent_14_24 = perc14_24,
    primary_seatbelt_law = sbprim,
    secondary_seatbelt_law = sbsecon,

    min_drinking_age = minage,
    zero_tolerance_law = zerotol,
    graduated_drivers_license_law = gdl,
    per_se_law = perse,

    total_traffic_fatalities = totfat,
    total_nighttime_fatalities = nghtfat,
    total_weekend_fatalities = wkndfat,
    total_fatalities_per_100_million_miles = totfatpvm,

    nighttime_fatalities_per_100_million_miles = nghtfatpvm,
    weekend_fatalities_per_100_million_miles = wkndfatpvm,
    state_population = statepop,
    speed_limit_over_70 = sl70plus,
    vehicmilespc = vehicmilespc,
  )


# Convert state code to state names
state_lookup <- data.frame(
  state_code = c(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, 29, 30, 31, 32, 33, 34, 35, 36, 37,
                 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51),
  state_name = c("AL", "AZ", "AR", "CA", "CO", "CT", "DE", "DC", "FL", "GA", "ID",
                 "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD", "MA", "MI", "MN",
                 "MS", "MO", "MT", "NE", "NV", "NJ", "NM", "NY", "NC", "ND", "OH",
                 "OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX", "UT", "VT", "VA",
                 "WA", "WV", "WI", "WY")

)

data_clean$state <- state_lookup$state_name[match(data_clean$state,
                                                  state_lookup$state_code)]
```

2. (5 points) Provide a description of the basic structure of the dataset. What is this data? How, where, and when is it collected? Is the data generated through a survey or some other method? Is the data that is presented a sample from the population, or is it a *census* that represents the entire population? Minimally, this should include:

   - How is the our dependent variable of interest `total_fatalities_rate` defined?

The data we are working with contains information about various fatality rates such as weekend fatality rate and nighttime fatality rate for each of the 48 continental US states from 1980 to 2003. Additional information includes attributes pertaining to state alcohol laws and demographic data, including alchohol levels considered thresholds of intoxication for different states, unemployment rate, whether primary or secondary seat belt laws were implemented, state population, and minimum drinking age. According to the data description dictionary, the data was collected over the years 1980 to 2004. It was gathered by state employees using a standard format for use by Professor Freeman's research (Freeman 2007).

Due to the various features collected for each state and their values at different years between 1980 and 2004, we have panel data, where the individuals are the 48 continental US states and the time period is between 1980 and 2004.

3. (20 points) Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable `total_fatalities_rate` and the potential explanatory variables. Minimally, this should include:

   - How is the our dependent variable of interest `total_fatalities_rate` defined?

- What is the average of `total_fatalities_rate` in each of the years in the time period covered in this dataset?

As with every EDA this semester, the goal of this EDA is not to document your own process of discovery – save that for an exploration notebook – but instead it is to bring a reader that is new to the data to a full understanding of the important features of your data as quickly as possible. In order to do this, your EDA should include a detailed, orderly narrative description of what you want your reader to know. Do not include any output – tables, plots, or statistics – that you do not intend to write about.

The variable `total_fatalities_rate` is defined as the total traffic fatalities per 100,000 population in the state.

```r
summary(data_clean) #Hidden for brevity
```

```
##       year         state              seatbelt     min_drinking_age
##  Min.   :1980   Length:1200        Min.   :0.000   Min.   :18.0
##  1st Qu.:1986   Class :character   1st Qu.:0.000   1st Qu.:21.0
##  Median :1992   Mode  :character   Median :1.000   Median :21.0
##  Mean   :1992                      Mean   :1.116   Mean   :20.6
##  3rd Qu.:1998                      3rd Qu.:2.000   3rd Qu.:21.0
##  Max.   :2004                      Max.   :2.000   Max.   :21.0
##
##  zero_tolerance_law graduated_drivers_license_law     bac10
##  Min.   :0.0000     Min.   :0.00                  Min.   :0.0000
##  1st Qu.:0.0000     1st Qu.:0.00                  1st Qu.:0.0000
##  Median :0.0000     Median :0.00                  Median :1.0000
##  Mean   :0.4519     Mean   :0.18                  Mean   :0.6392
##  3rd Qu.:1.0000     3rd Qu.:0.00                  3rd Qu.:1.0000
##  Max.   :1.0000     Max.   :1.00                  Max.   :1.0000
##
##      bac08           per_se_law     total_traffic_fatalities
##  Min.   :0.0000   Min.   :0.0000   Min.   :  63.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 310.0
##  Median :0.0000   Median :1.0000   Median : 676.0
##  Mean   :0.2175   Mean   :0.5517   Mean   : 900.7
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1099.5
##  Max.   :1.0000   Max.   :1.0000   Max.   :5504.0
##
##  total_nighttime_fatalities total_weekend_fatalities
##  Min.   :  26.0             Min.   :  10.0
##  1st Qu.: 139.8             1st Qu.:  70.0
##  Median : 316.0             Median : 163.0
##  Mean   : 427.3             Mean   : 222.3
##  3rd Qu.: 518.2             3rd Qu.: 277.0
##  Max.   :2918.0             Max.   :1499.0
##
##  total_fatalities_per_100_million_miles
##  Min.   :0.780
##  1st Qu.:1.577
##  Median :2.020
##  Mean   :2.122
##  3rd Qu.:2.500
##  Max.   :5.700
##
##  nighttime_fatalities_per_100_million_miles
##  Min.   :0.2700
##  1st Qu.:0.6847
##  Median :0.9130
##  Mean   :0.9990
##  3rd Qu.:1.2110
##  Max.   :3.0030
##
##  weekend_fatalities_per_100_million_miles state_population
##  Min.   :0.1140                           Min.   :  453401
##  1st Qu.:0.3410                           1st Qu.: 1641938
##  Median :0.4770                           Median : 3700425
##  Mean   :0.5255                           Mean   : 5329896
##  3rd Qu.:0.6420                           3rd Qu.: 6069563
##  Max.   :1.6750                           Max.   :35894000
##
##  total_fatalities_rate night_fatalities_rate weekend_fatalities_rate
##  Min.   : 6.20         Min.   : 2.660        Min.   : 1.180
##  1st Qu.:14.38         1st Qu.: 6.338        1st Qu.: 3.240
##  Median :18.43         Median : 8.420        Median : 4.390
##  Mean   :18.92         Mean   : 8.796        Mean   : 4.606
##  3rd Qu.:22.77         3rd Qu.:10.650        3rd Qu.: 5.680
##  Max.   :53.32         Max.   :29.600        Max.   :14.430
##
##  vehicle_miles     unemployment_rate percent_14_24   speed_limit_over_70
##  Min.   :  3.703   Min.   : 2.200    Min.   :11.70   Min.   :0.0000
##  1st Qu.: 14.574   1st Qu.: 4.500    1st Qu.:13.90   1st Qu.:0.0000
##  Median : 33.863   Median : 5.600    Median :14.90   Median :0.0000
##  Mean   : 46.323   Mean   : 5.951    Mean   :15.33   Mean   :0.2092
##  3rd Qu.: 58.639   3rd Qu.: 7.000    3rd Qu.:16.60   3rd Qu.:0.0000
##  Max.   :329.600   Max.   :18.000    Max.   :20.30   Max.   :1.0000
##
##  primary_seatbelt_law secondary_seatbelt_law  vehicmilespc    speed_limit
##  Min.   :0.0000       Min.   :0.0000         Min.   : 4372   Min.   :55.00
##  1st Qu.:0.0000       1st Qu.:0.0000         1st Qu.: 7788   1st Qu.:55.00
##  Median :0.0000       Median :0.0000         Median : 9013   Median :65.00
##  Mean   :0.1792       Mean   :0.4683         Mean   : 9129   Mean   :62.93
##  3rd Qu.:0.0000       3rd Qu.:1.0000         3rd Qu.:10327   3rd Qu.:65.00
##  Max.   :1.0000       Max.   :1.0000         Max.   :18390   Max.   :75.00
##                                                              NA's   :9
##  year_of_observation      bac
##  Min.   :1980        Min.   :0.0800
##  1st Qu.:1986        1st Qu.:0.1000
##  Median :1992        Median :0.1000
##  Mean   :1992        Mean   :0.0952
##  3rd Qu.:1998        3rd Qu.:0.1000
##  Max.   :2004        Max.   :0.1000
##                      NA's   :191
```

```
dim(data_clean)
```

```
## [1] 1200   29
```

```
ptraffic_fatalities <- pdata.frame(data_clean,
                                   index=c("state", "year_of_observation"))

# Check for gaps in the time series of each state
ptraffic_fatalities %>%
  is.pconsecutive()
```

```
##   AL   AR   AZ   CA   CO   CT   DC   DE   FL   GA   IA   ID   IL   IN   KS   KY
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   LA   MA   MD   ME   MI   MN   MO   MS   MT   NC   ND   NE   NJ   NM   NV   NY
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   OH   OK   OR   PA   RI   SC   SD   TN   TX   UT   VA   VT   WA   WI   WV   WY
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```r
pdim(ptraffic_fatalities)
```

```
## Balanced Panel: n = 48, T = 25, N = 1200
```

The dataset consists of 1200 observations and 27 variables converted to a balanced panel format. The variable `state` is the individual identifier and a factor variable with 48 levels (one for each of the 48 contiguous federal states of the US). The variable `year` is the time identifier and a factor variable with 25 levels identifying the period when the observations were made.

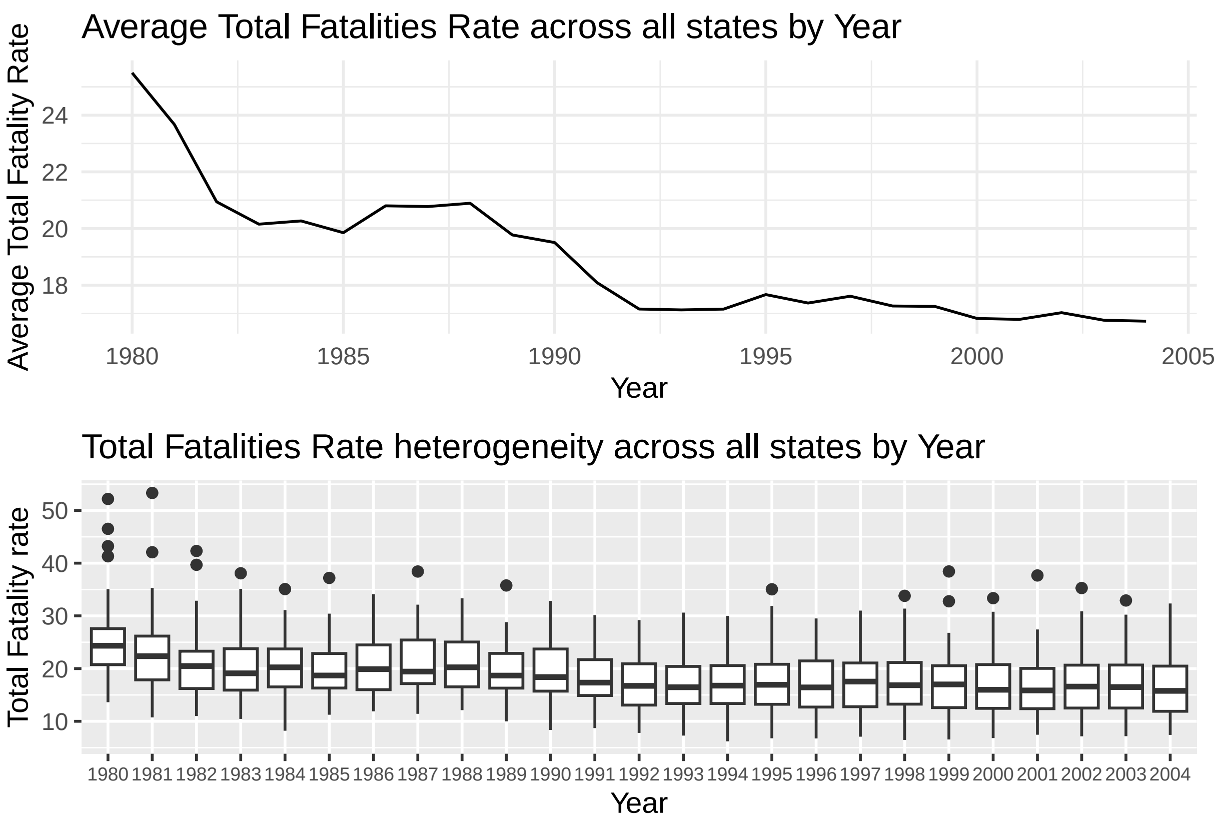
From the summary statistics, we notice that some variables such as vehicmilespc and state population have much higher numerical ranges than do other numerical variables. This may skew modeling results and will be addressed in the EDA and modeling sections.

The average of `total_fatalities_rate` in each of the years is seen below. It is evident that fatality rate has decreased over the years on average.

```r
average_by_year <- aggregate(ptraffic_fatalities$total_fatalities_rate,
                             by = list(ptraffic_fatalities$year), FUN = mean)
names(average_by_year) <- c("year", "average_total_fatalities_rate")
average_by_year
```

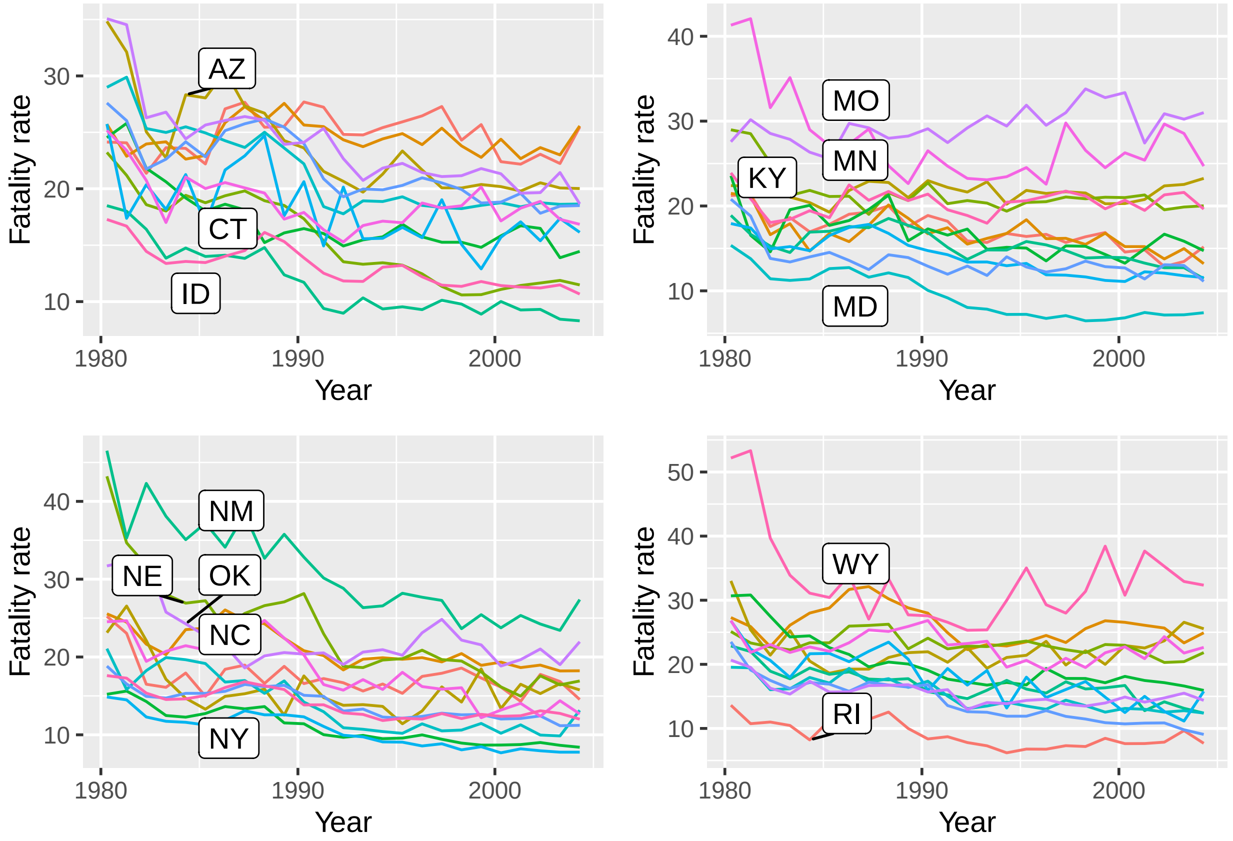
```
##    year average_total_fatalities_rate
## 1  1980                      25.49458
## 2  1981                      23.67021
## 3  1982                      20.94250
## 4  1983                      20.15292
## 5  1984                      20.26750
## 6  1985                      19.85146
## 7  1986                      20.80042
## 8  1987                      20.77479
## 9  1988                      20.89167
## 10 1989                      19.77229
## 11 1990                      19.50521
## 12 1991                      18.09479
## 13 1992                      17.15792
## 14 1993                      17.12771
## 15 1994                      17.15521
## 16 1995                      17.66854
## 17 1996                      17.36938
## 18 1997                      17.61062
## 19 1998                      17.26542
## 20 1999                      17.25042
## 21 2000                      16.82562
## 22 2001                      16.79271
## 23 2002                      17.02958
## 24 2003                      16.76354
## 25 2004                      16.72896
```

```
plot_avg <- ggplot(data = average_by_year,
                   aes(x = year, y = average_total_fatalities_rate)) +
  geom_line() +
  labs(x = "Year", y = "Average Total Fatality Rate") +
  ggtitle("Average Total Fatalities Rate across all states by Year") +
  theme_minimal()

plot_total <- ptraffic_fatalities %>%
  group_by(year) %>%
  ggplot(aes(x = as.character(year), y = total_fatalities_rate)) +
  geom_boxplot() +
  ggtitle("Total Fatalities Rate heterogeneity across all states by Year") +
  labs(x = "Year",  y = "Total Fatality rate") +
  theme(axis.text.x = element_text(size = 7))

plot_avg / plot_total
```



Let's see how our variables of interest has changed over time.

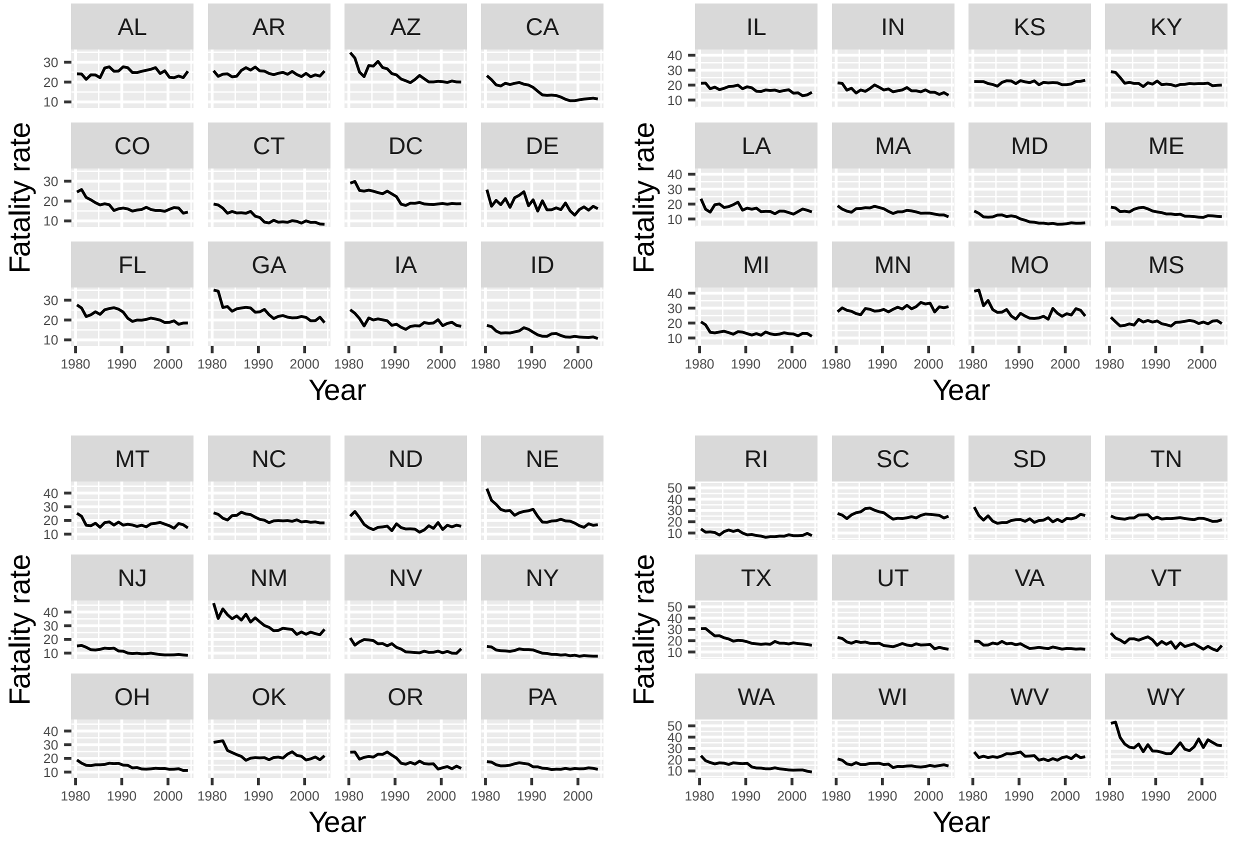
```
# Our variable of interest is total_fatalities_rate

p2<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line(aes(color = state)) +
  labs(x = "Year",  y = "Fatality rate")+
  geom_label_repel(data = filter(ptraffic_fatalities,
                              as.integer(state) <= 12  & year == 1984),
                   aes(label = state), nudge_x = .75,na.rm = TRUE) +
  theme(legend.position = "none")

p3<- ptraffic_fatalities %>%
   filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line(aes(color = state)) +
  labs(x = "Year",  y = "Fatality rate")+
  geom_label_repel(data = filter(ptraffic_fatalities, as.integer(state) > 12 &
                                   as.integer(state) <= 24  & year == 1984),
                   aes(label = state), nudge_x = .75,na.rm = TRUE) +
  theme(legend.position = "none")

p4<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line(aes(color = state)) +
  labs(x = "Year",  y = "Fatality rate")+
  geom_label_repel(data = filter(ptraffic_fatalities,
                                 as.integer(state) > 24 &as.integer(state) <= 36
                                 & year == 1984),
                   aes(label = state), nudge_x = .75,na.rm = TRUE) +
  theme(legend.position = "none")

p5<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line(aes(color = state)) +
  labs(x = "Year",  y = "Fatality rate") +
  geom_label_repel(data = filter(ptraffic_fatalities,
                                 as.integer(state) > 36 & year == 1984),
                   aes(label = state),
                   nudge_x = .75,na.rm = TRUE) +
  theme(legend.position = "none")

grid.arrange(p2,p3,p4, p5, nrow = 2, ncol = 2)
```



Looks like in almost all states, the traffic fatality rate has decreased over the time period. Let's simplify our visualization further.

```
p6<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line() +
  facet_wrap(~ state, nrow = 3) +
  labs(x = "Year",  y = "Fatality rate") +
  theme(legend.position = "none", axis.text.x = element_text(size = 5),
        axis.text.y = element_text(size = 5))
```

```r
p7<-ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line() +
  facet_wrap(~ state, nrow = 3)+
  labs(x = "Year",  y = "Fatality rate") +
  theme(legend.position = "none", axis.text.x = element_text(size = 5),
        axis.text.y = element_text(size = 5))

p8<-ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line() +
  facet_wrap(~ state, nrow = 3)+
  labs(x = "Year",  y = "Fatality rate") +
  theme(legend.position = "none", axis.text.x = element_text(size = 5),
        axis.text.y = element_text(size = 5))

p9<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  ggplot(aes(x = as.Date(as.character(year),"%Y"), y = total_fatalities_rate)) +
  geom_line()+
  facet_wrap(~ state, nrow = 3)+
  labs(x = "Year",  y = "Fatality rate") +
  theme(legend.position = "none", axis.text.x = element_text(size = 5),
        axis.text.y = element_text(size = 5))

grid.arrange(p6,p7,p8, p9, nrow = 2, ncol = 2)
```

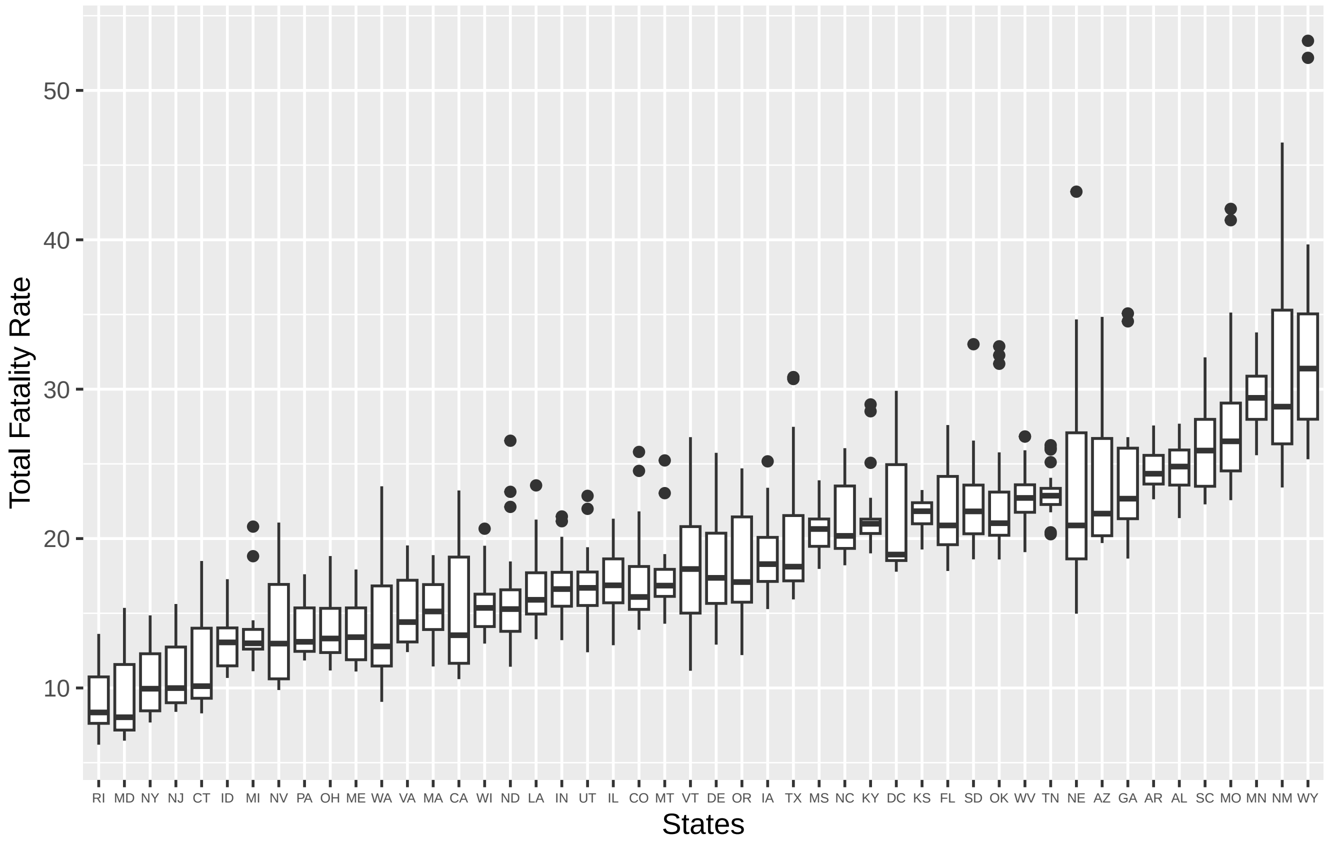


From the plot above, there is a decreasing trend in fatality rate across all states except for `MN`, `MO` and `WY`.

Let's plot a box plot to show the heterogeneity of traffic fatality rate.

```
# Calculate mean total fatalities rate for each state
mean_fatalities <- ptraffic_fatalities %>%
  dplyr::group_by(state) %>%
  dplyr::summarise(mean_total_fatalities_rate =
                     mean(total_fatalities_rate, na.rm = TRUE))

# Order states by their mean
ordered_states <- mean_fatalities %>%
```

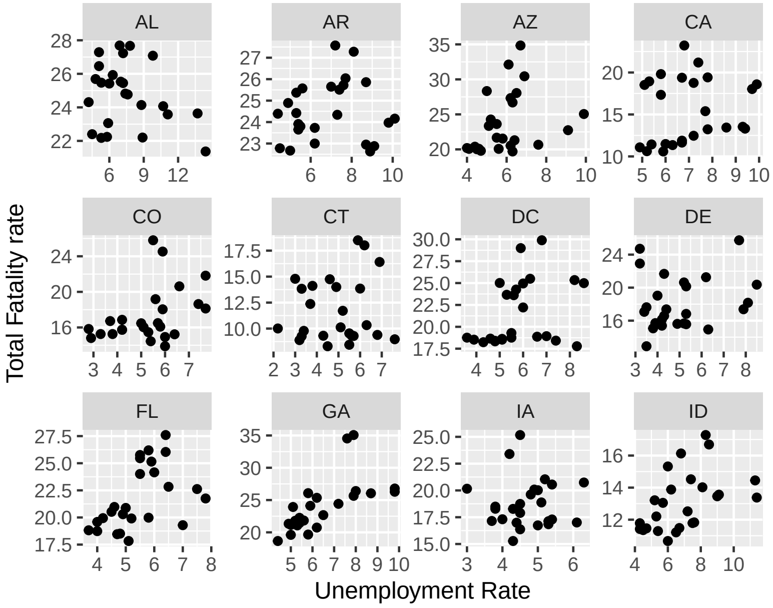
```r
  dplyr::arrange(mean_total_fatalities_rate) %>%
  dplyr::pull(state)

# Reorder
ptraffic_fatalities$state_ordered <- factor(ptraffic_fatalities$state,
                                            levels = ordered_states)

# Box plot to show the heterogeneity of traffic fatality rate.
ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = state_ordered, y = total_fatalities_rate, group=state_ordered)) +
  geom_boxplot() +
  labs(x = "States", y = "Total Fatality Rate") +
  theme(axis.text.x = element_text(size = 5))
```

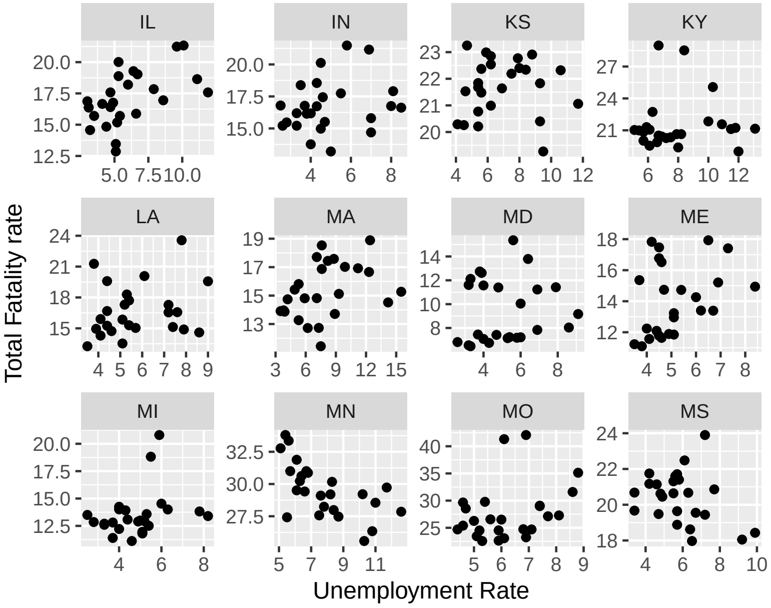


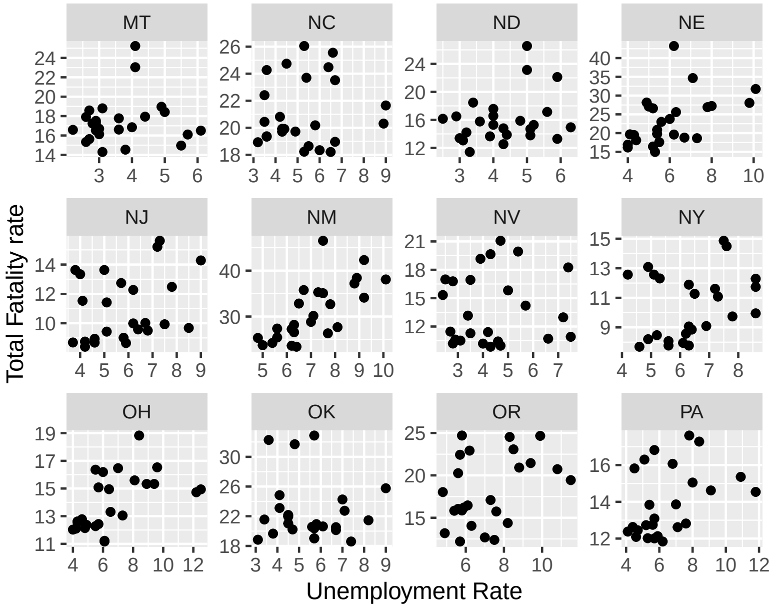
We can see strong differences in traffic fatality rates across states, suggesting that fixed effects are important for controlling for unobserved differences. It looks like state `RI` and `MD` have the lowest fatality rate on average while `NM` and `WY` have the highest fatality rate on average.

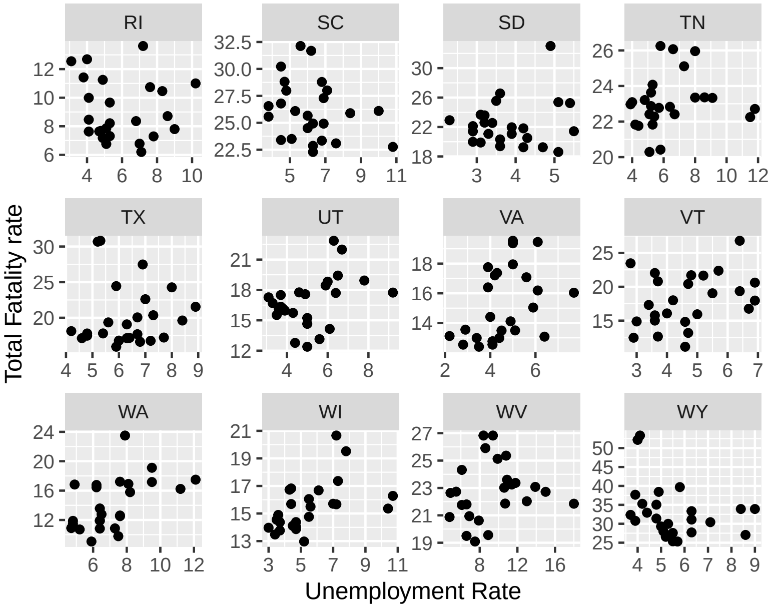
In our exploration, we found that almost all our variables possessed trends over time and correlated (positively or negatively) with our variable of interest. Here, we include a few variables which we think are highly relevant for the analysis of traffic fatality rate. They are `unemployment_rate`, `bac`, `percent_14_24`.

```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  ggplot(aes(x = unemployment_rate, y = total_fatalities_rate)) +
  geom_point() +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "Unemployment Rate",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  ggplot(aes(x = unemployment_rate, y = total_fatalities_rate)) +
  geom_point() +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Unemployment Rate",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  ggplot(aes(x = unemployment_rate, y = total_fatalities_rate)) +
  geom_point() +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Unemployment Rate",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  ggplot(aes(x = unemployment_rate, y = total_fatalities_rate)) +
  geom_point()+
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Unemployment Rate",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
Unemployment Rate

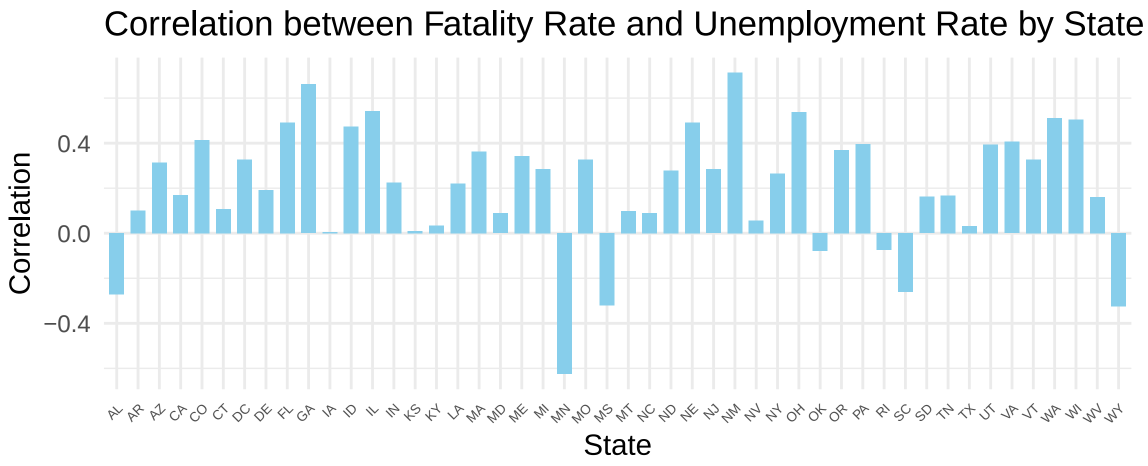

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
Unemployment Rate


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
Unemployment Rate


RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
Unemployment Rate

Although unemployment rate and total fatality rate has decreased over time for all the states, from the plots above, there is some correlation between these two features in a few states. Visually, for example, in the state of `MN`, we see that fatality rate decreases as unemployment rate increases, where as in `GA`, fatality rate increases as unemployment rate increases. See table below:

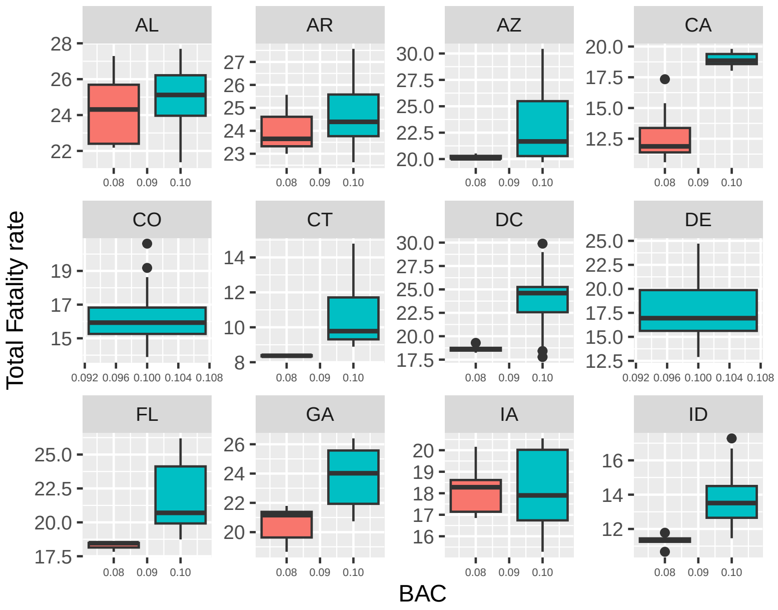
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       unemployment_rate, use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title = "Correlation between Fatality Rate and Unemployment Rate by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

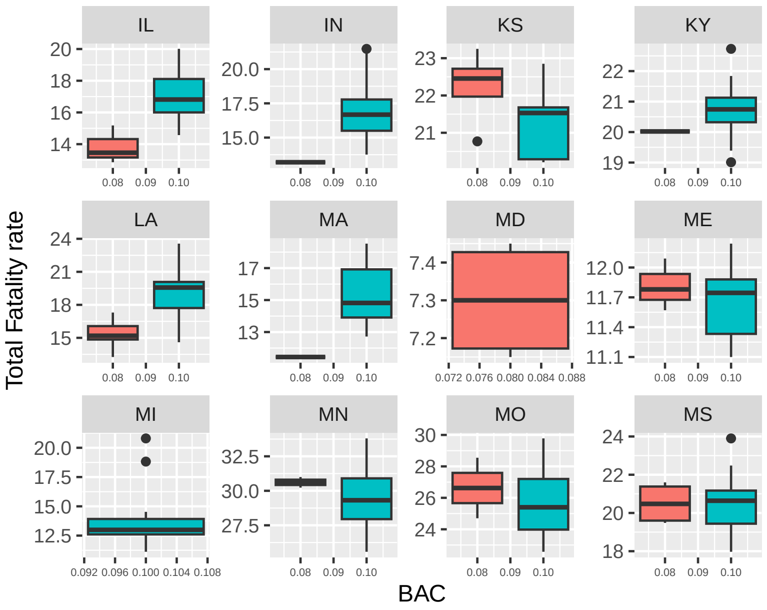


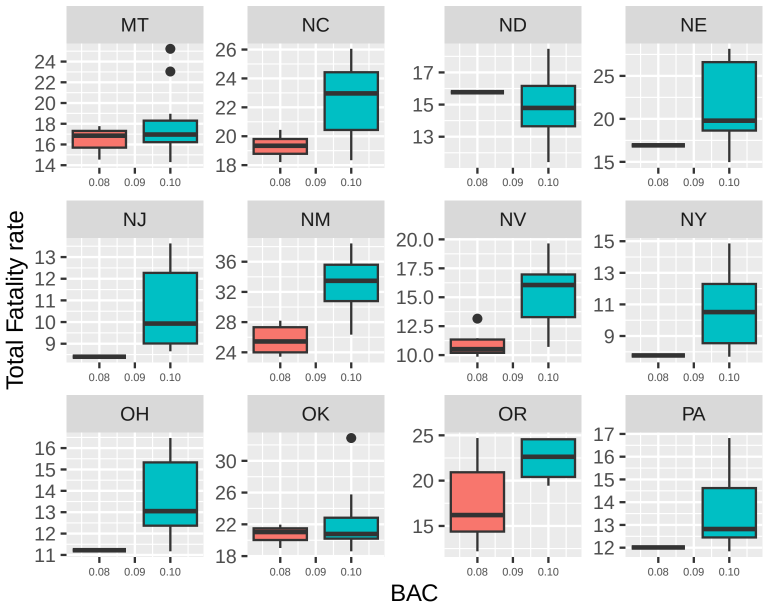
Let's check for BAC.

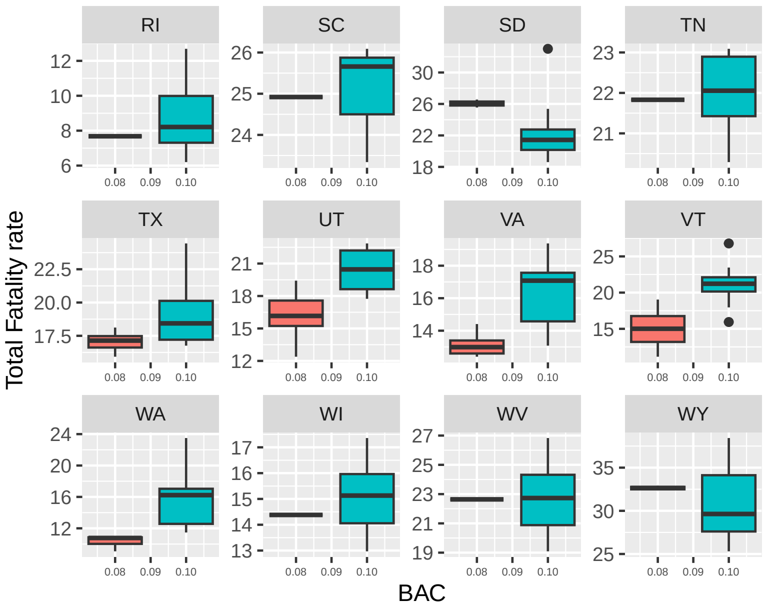
```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = bac, y = total_fatalities_rate, fill = factor(bac))) +
  geom_boxplot() +
```

```r
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "BAC",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = bac, y = total_fatalities_rate, fill = factor(bac))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "BAC",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = bac, y = total_fatalities_rate, fill = factor(bac))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "BAC",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = bac, y = total_fatalities_rate, fill = factor(bac))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "BAC",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


Total Fatality rate
BAC
AL 28 26 24 22 0.08 0.09 0.10
AR 27 26 25 24 23 0.08 0.09 0.10
AZ 30.0 27.5 25.0 22.5 20.0 0.08 0.09 0.10
CA 20.0 17.5 15.0 12.5 0.08 0.09 0.10
CO 19 17 15 0.092 0.096 0.100 0.104 0.108
CT 14 12 10 8 0.08 0.09 0.10
DC 30.0 27.5 25.0 22.5 20.0 17.5 0.08 0.09 0.10
DE 25.0 22.5 20.0 17.5 15.0 12.5 0.092 0.096 0.100 0.104 0.108
FL 25.0 22.5 20.0 17.5 0.08 0.09 0.10
GA 26 24 22 20 0.08 0.09 0.10
IA 20 19 18 17 16 0.08 0.09 0.10
ID 16 14 12 0.08 0.09 0.10

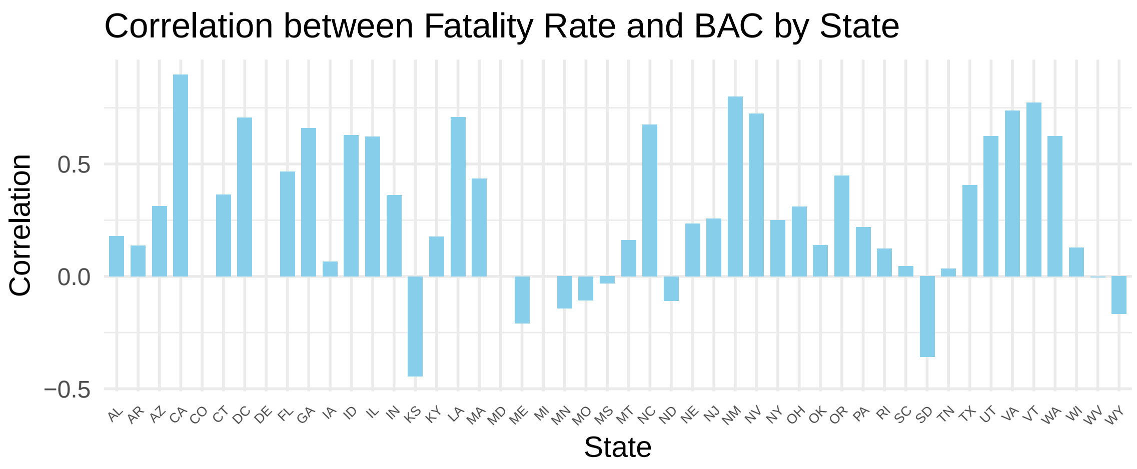

Total Fatality rate
BAC
IL 20 18 16 14 0.08 0.09 0.10
IN 20.0 17.5 15.0 0.08 0.09 0.10
KS 23 22 21 0.08 0.09 0.10
KY 22 21 20 19 0.08 0.09 0.10
LA 24 21 18 15 0.08 0.09 0.10
MA 17 15 13 0.08 0.09 0.10
MD 7.4 7.3 7.2 0.072 0.076 0.080 0.084 0.088
ME 12.0 11.7 11.4 11.1 0.08 0.09 0.10
MI 20.0 17.5 15.0 12.5 0.092 0.096 0.100 0.104 0.108
MN 32.5 30.0 27.5 0.08 0.09 0.10
MO 30 28 26 24 0.08 0.09 0.10
MS 24 22 20 18 0.08 0.09 0.10


Total Fatality rate
BAC
MT 24 22 20 18 16 14 0.08 0.09 0.10
NC 26 24 22 20 18 0.08 0.09 0.10
ND 17 15 13 0.08 0.09 0.10
NE 25 20 15 0.08 0.09 0.10
NJ 13 12 11 10 9 0.08 0.09 0.10
NM 36 32 28 24 0.08 0.09 0.10
NV 20.0 17.5 15.0 12.5 10.0 0.08 0.09 0.10
NY 15 13 11 9 0.08 0.09 0.10
OH 16 15 14 13 12 11 0.08 0.09 0.10
OK 30 26 22 18 0.08 0.09 0.10
OR 25 20 15 0.08 0.09 0.10
PA 17 16 15 14 13 12 0.08 0.09 0.10


Total Fatality rate
BAC
RI 12 10 8 6 0.08 0.09 0.10
SC 26 25 24 0.08 0.09 0.10
SD 30 26 22 18 0.08 0.09 0.10
TN 23 22 21 0.08 0.09 0.10
TX 22.5 20.0 17.5 0.08 0.09 0.10
UT 21 18 15 12 0.08 0.09 0.10
VA 18 16 14 0.08 0.09 0.10
VT 25 20 15 0.08 0.09 0.10
WA 24 20 16 12 0.08 0.09 0.10
WI 17 16 15 14 13 0.08 0.09 0.10
WV 27 25 23 21 19 0.08 0.09 0.10
WY 35 30 25 0.08 0.09 0.10

In the above plot, we can see higher fatality rate for BAC level of 0.10 than 0.08, for almost all states except for states like KS and MN. Below is a correlation plot for better visualization.

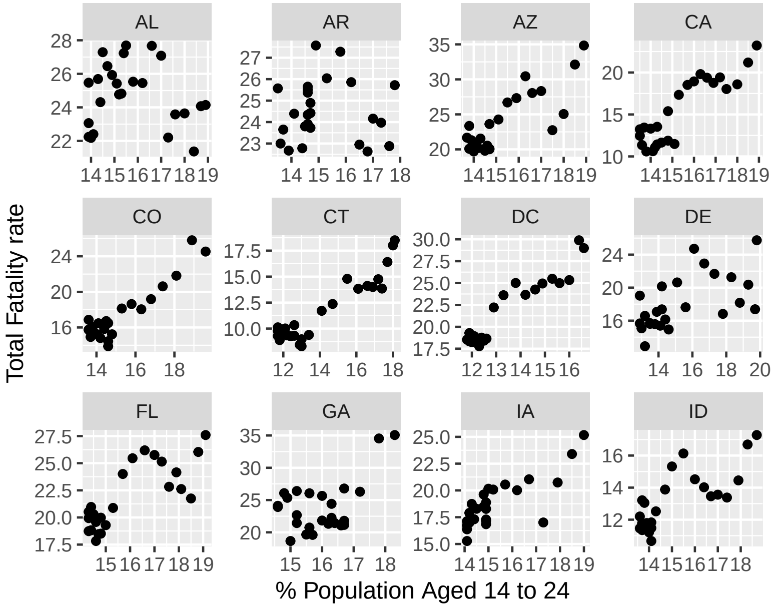
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation =
                       cor(total_fatalities_rate, bac, use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation", title =
         "Correlation between Fatality Rate and BAC by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

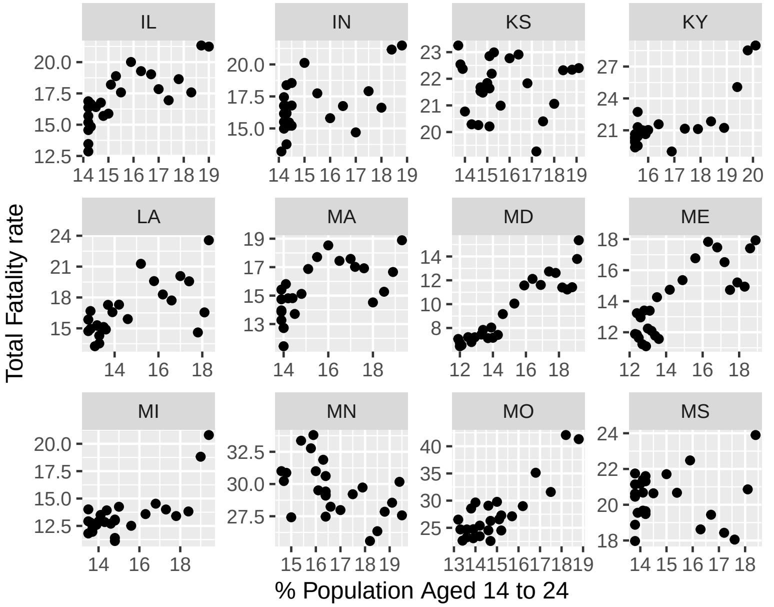


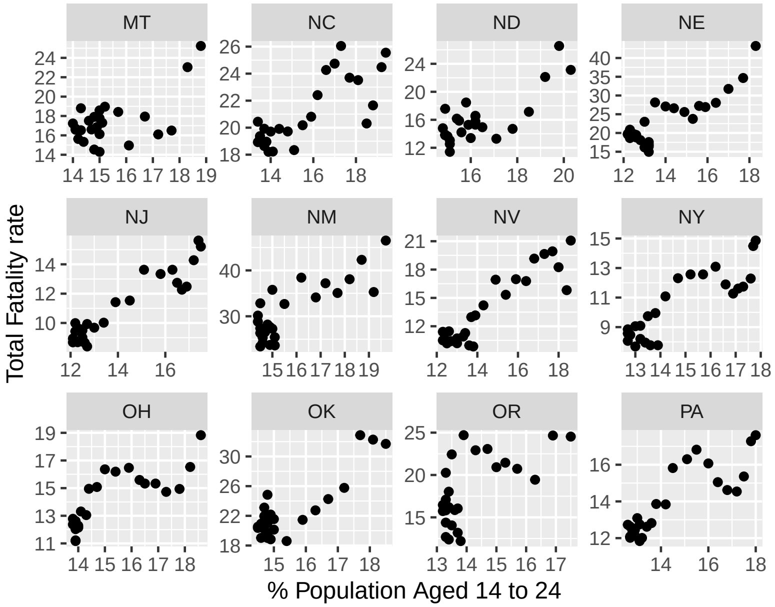
Percentage of population in aged through 14 to 24 seems to show high levels of correlation.

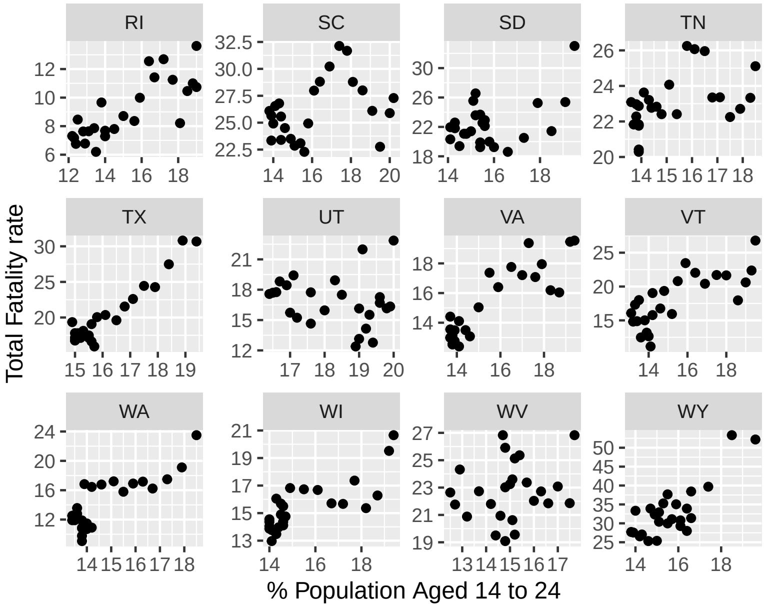
```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  ggplot(aes(x = percent_14_24, y = total_fatalities_rate)) +
  geom_point() +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "% Population Aged 14 to 24",  y = "Total Fatality rate")
```

```r
p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  ggplot(aes(x = percent_14_24, y = total_fatalities_rate)) +
  geom_point() +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "% Population Aged 14 to 24",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  ggplot(aes(x = percent_14_24, y = total_fatalities_rate)) +
  geom_point() +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "% Population Aged 14 to 24",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  ggplot(aes(x = percent_14_24, y = total_fatalities_rate)) +
  geom_point()+
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "% Population Aged 14 to 24",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


Total Fatality rate
% Population Aged 14 to 24
AL 28 26 24 22 14 15 16 17 18 19
AR 27 26 25 24 23 14 15 16 17 18
AZ 35 30 25 20 14 15 16 17 18 19
CA 20 15 10 14 15 16 17 18 19
CO 24 20 16 14 16 18
CT 17.5 15.0 12.5 10.0 12 14 16 18
DC 30.0 27.5 25.0 22.5 20.0 17.5 12 13 14 15 16
DE 24 20 16 14 16 18 20
FL 27.5 25.0 22.5 20.0 17.5 15 16 17 18 19
GA 35 30 25 20 15 16 17 18
IA 25.0 22.5 20.0 17.5 15.0 14 15 16 17 18 19
ID 16 14 12 14 15 16 17 18


Total Fatality rate
% Population Aged 14 to 24
IL 20.0 17.5 15.0 12.5 14 15 16 17 18 19
IN 20.0 17.5 15.0 14 15 16 17 18 19
KS 23 22 21 20 14 15 16 17 18 19
KY 27 24 21 16 17 18 19 20
LA 24 21 18 15 14 16 18
MA 19 17 15 13 14 16 18
MD 14 12 10 8 12 14 16 18
ME 18 16 14 12 12 14 16 18
MI 20.0 17.5 15.0 12.5 14 16 18
MN 32.5 30.0 27.5 15 16 17 18 19
MO 40 35 30 25 13 14 15 16 17 18 19
MS 24 22 20 18 14 15 16 17 18

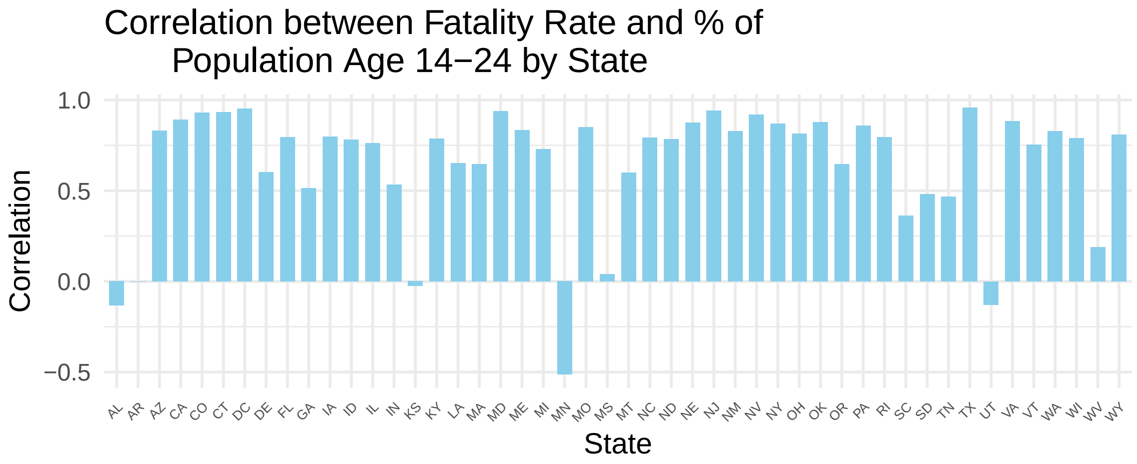

Total Fatality rate
% Population Aged 14 to 24
MT 24 22 20 18 16 14 14 15 16 17 18 19
NC 26 24 22 20 18 14 16 18
ND 24 20 16 12 16 18 20
NE 40 35 30 25 20 15 12 14 16 18
NJ 14 12 10 12 14 16
NM 40 30 15 16 17 18 19
NV 21 18 15 12 12 14 16 18
NY 15 13 11 9 13 14 15 16 17 18
OH 19 17 15 13 11 14 15 16 17 18
OK 30 26 22 18 15 16 17 18
OR 25 20 15 13 14 15 16 17
PA 16 14 12 14 16 18


Total Fatality rate
% Population Aged 14 to 24
RI 12 10 8 6 12 14 16 18
SC 32.5 30.0 27.5 25.0 22.5 14 16 18 20
SD 30 26 22 18 14 16 18
TN 26 24 22 20 14 15 16 17 18
TX 30 25 20 15 16 17 18 19
UT 21 18 15 12 17 18 19 20
VA 18 16 14 14 16 18
VT 25 20 15 14 16 18
WA 24 20 16 12 14 15 16 17 18
WI 21 19 17 15 13 14 16 18
WV 27 25 23 21 19 13 14 15 16 17
WY 50 45 40 35 30 25 14 16 18

We can see from the above scatter plots that the correlations between % population aged 14 to 24 and total fatality rate are very strong. Most of the correlations are positive except for those in states such as KS and MN.

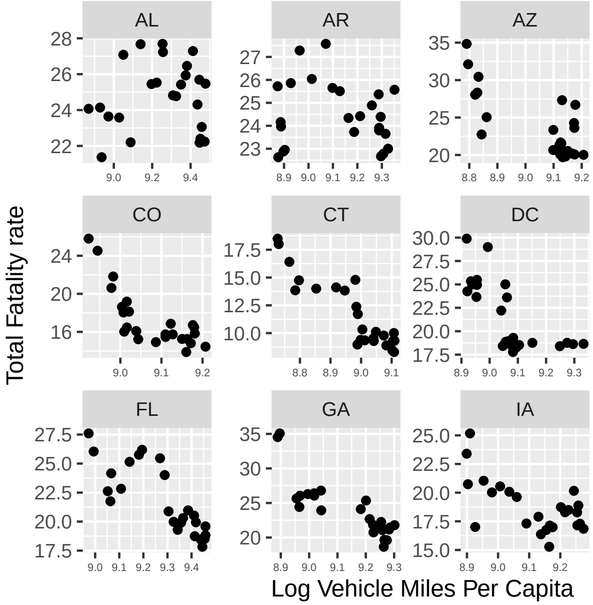
Below is a correlation plot for better visualization.

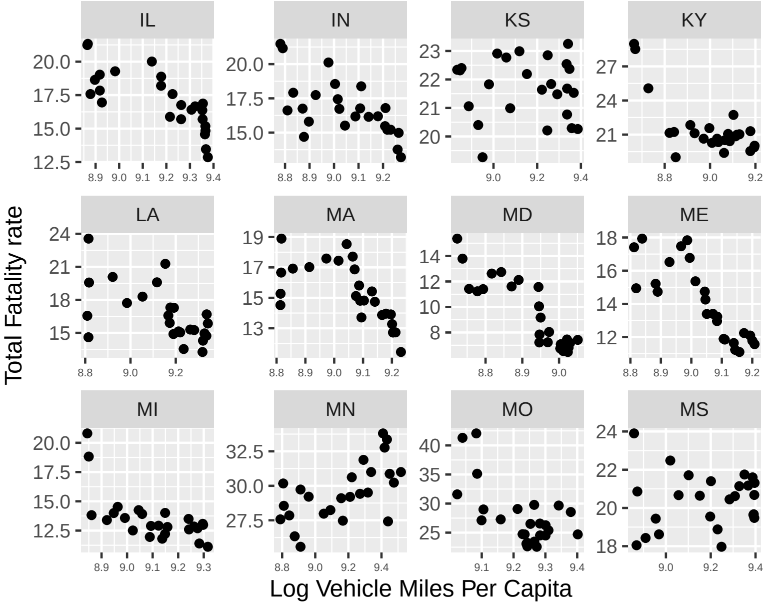
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       percent_14_24, use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title = "Correlation between Fatality Rate and % of
       Population Age 14-24 by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

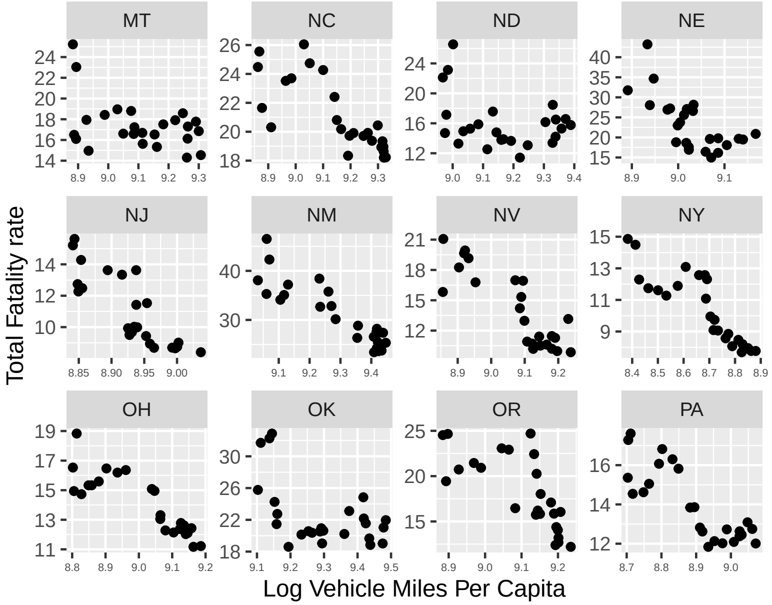


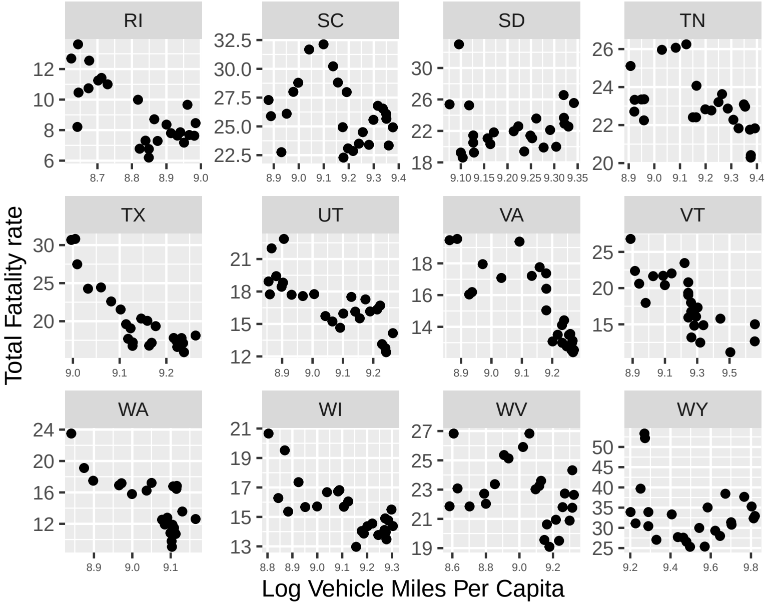
Let's check for vehicle miles traveled, in billions. As mentioned before, this variable has a much higher numerical range compared to other variables. Hence, we apply a log transformation to make the range smaller and more similar to those of other variables.

```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  ggplot(aes(x = log(vehicmilespc), y = total_fatalities_rate)) +
  geom_point() +
  theme(axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "Log Vehicle Miles Per Capita",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  ggplot(aes(x = log(vehicmilespc), y = total_fatalities_rate)) +
  geom_point() +
  theme(axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Log Vehicle Miles Per Capita",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  ggplot(aes(x = log(vehicmilespc), y = total_fatalities_rate)) +
  geom_point() +
  theme(axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Log Vehicle Miles Per Capita",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  ggplot(aes(x = log(vehicmilespc), y = total_fatalities_rate)) +
  geom_point()+
  theme(axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Log Vehicle Miles Per Capita",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
Log Vehicle Miles Per Capita

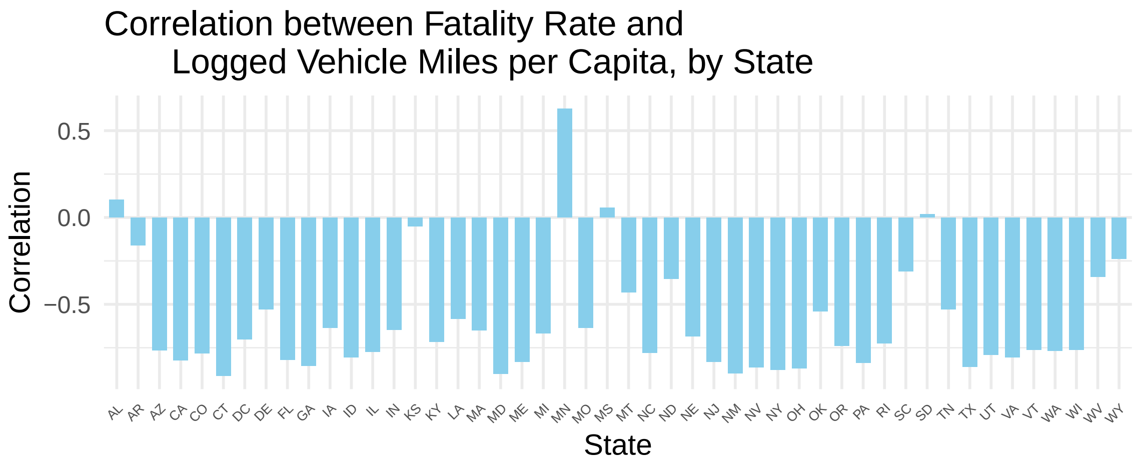

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
Log Vehicle Miles Per Capita


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
Log Vehicle Miles Per Capita


RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
Log Vehicle Miles Per Capita

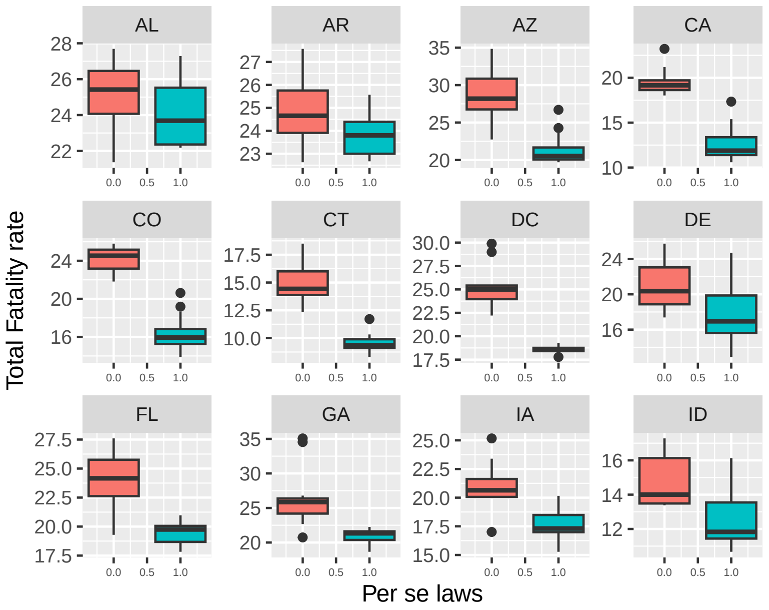
Clearly, almost all states are negatively correlated or have slight correlation with the log vehicle miles traveled, except for states such as MN which has high positive correlation.

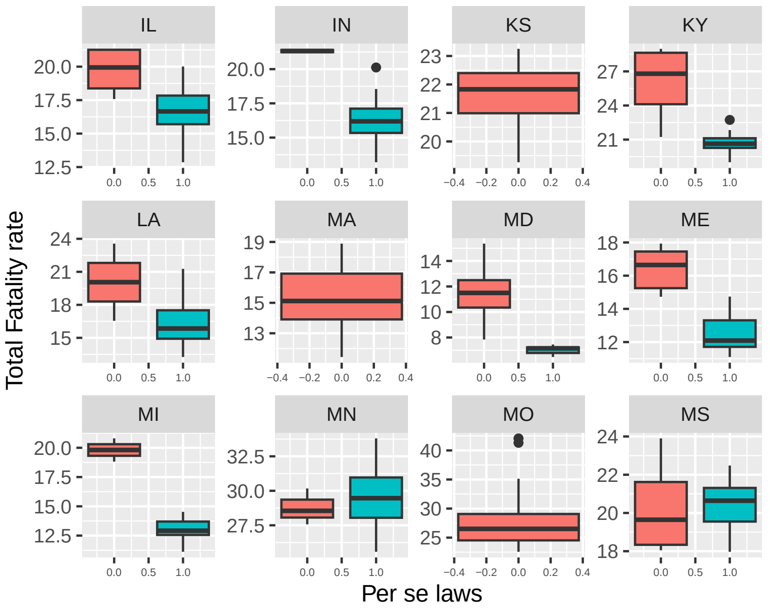
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       log(vehicmilespc), use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation", title =
         "Correlation between Fatality Rate and
       Logged Vehicle Miles per Capita, by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

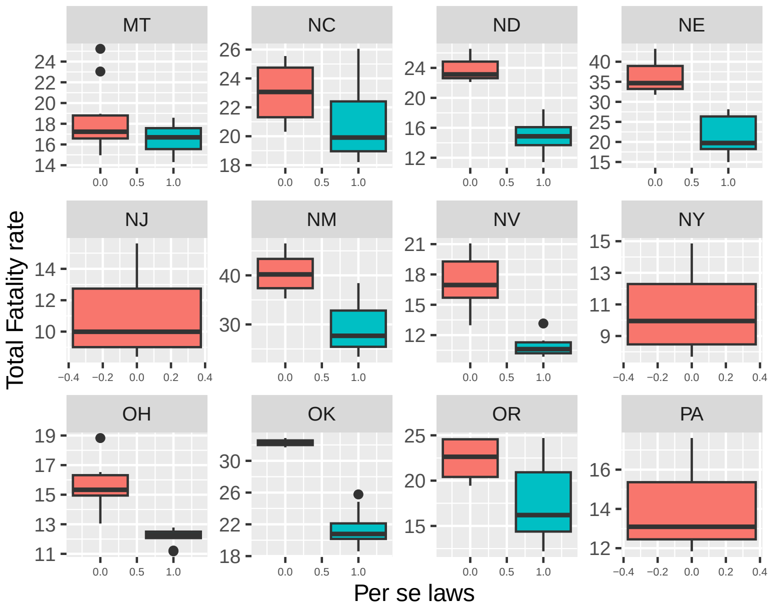


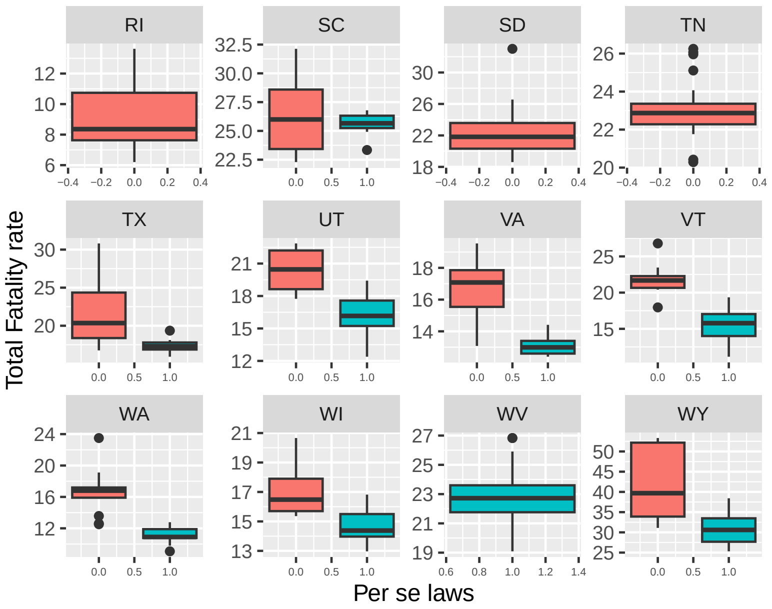
Let's check for per se laws.

```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = per_se_law, y = total_fatalities_rate,
             fill = factor(per_se_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "Per se laws",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = per_se_law, y = total_fatalities_rate,
             fill = factor(per_se_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Per se laws",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = per_se_law, y = total_fatalities_rate,
             fill = factor(per_se_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Per se laws",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = per_se_law, y = total_fatalities_rate,
             fill = factor(per_se_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "Per se laws",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
Per se laws

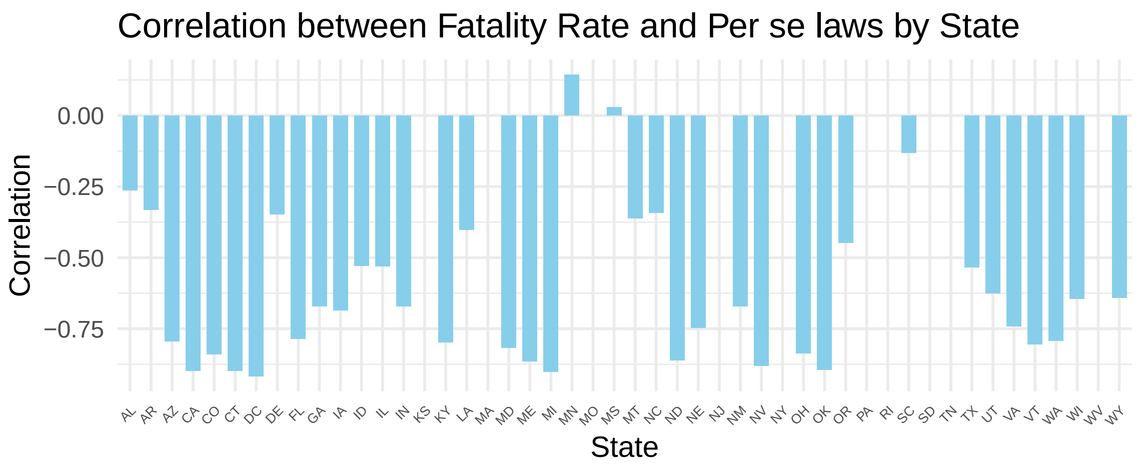

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
Per se laws


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
Per se laws


RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
Per se laws

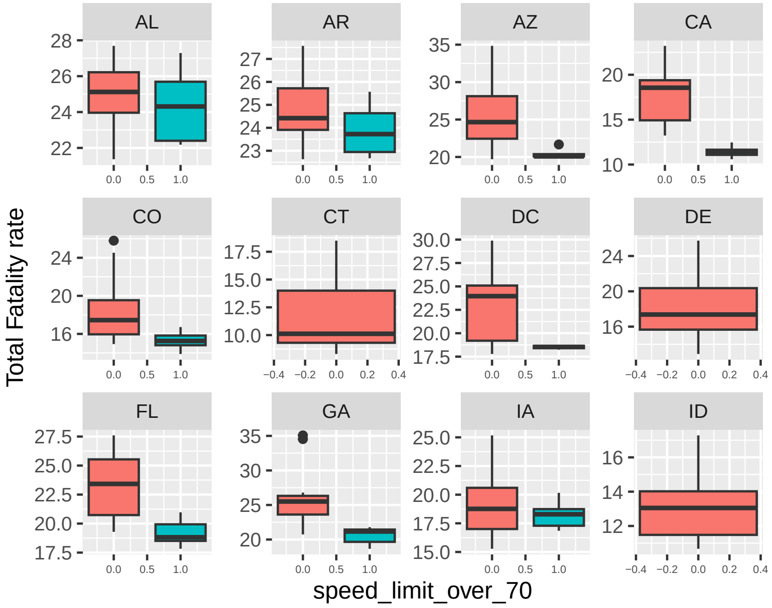
In the above plot, we can see higher fatality rates where there are no per se laws in a state for almost all states except for states like SC and MN. Below is a correlation plot for better visualization.

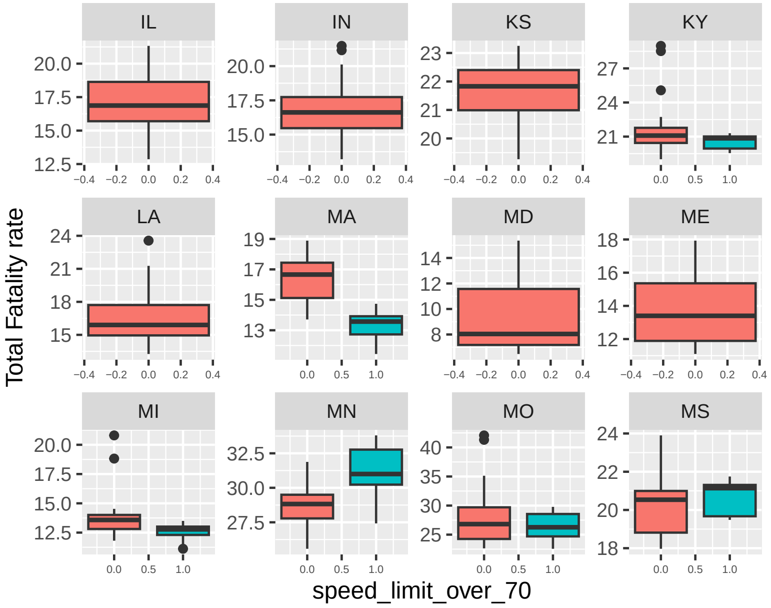
```r
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       per_se_law, use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title = "Correlation between Fatality Rate and Per se laws by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

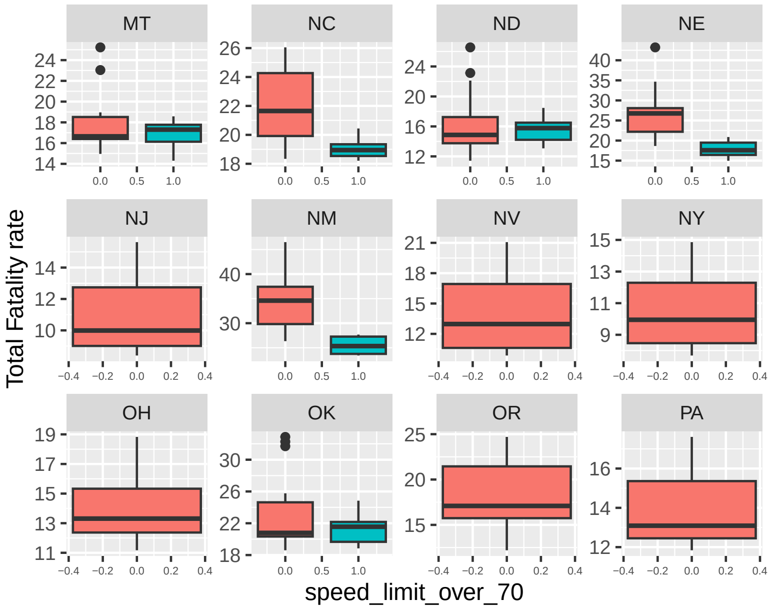


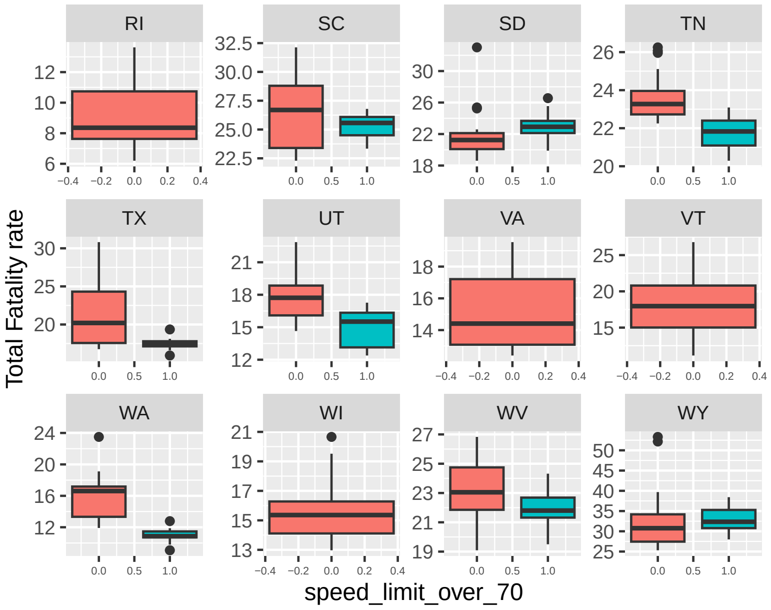
We see that for most states, the inclusion of per se laws is associated with lower fatality rates than those of states without per se laws. Among the states with positive correlation between per se laws and total fatality rates is MN.

```r
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit_over_70, y = total_fatalities_rate,
             fill = factor(speed_limit_over_70))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "speed_limit_over_70",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit_over_70, y = total_fatalities_rate,
             fill = factor(speed_limit_over_70))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "speed_limit_over_70",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit_over_70, y = total_fatalities_rate,
             fill = factor(speed_limit_over_70))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "speed_limit_over_70",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit_over_70, y = total_fatalities_rate,
             fill = factor(speed_limit_over_70))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "speed_limit_over_70",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
speed_limit_over_70

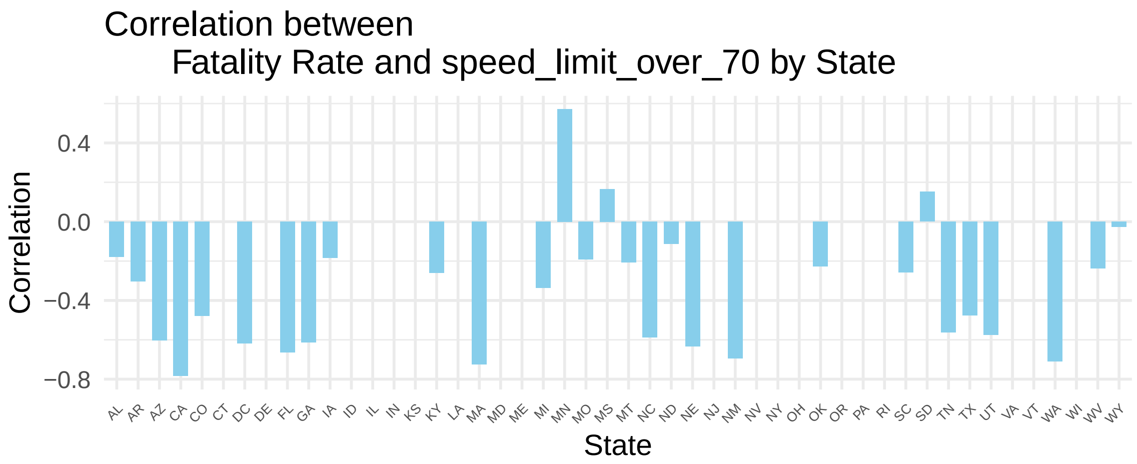

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
speed_limit_over_70


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
speed_limit_over_70

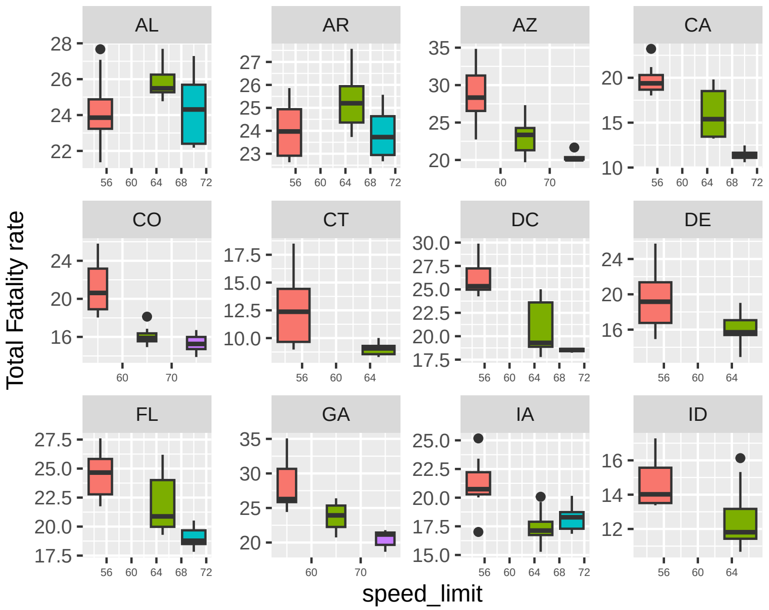

RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
speed_limit_over_70

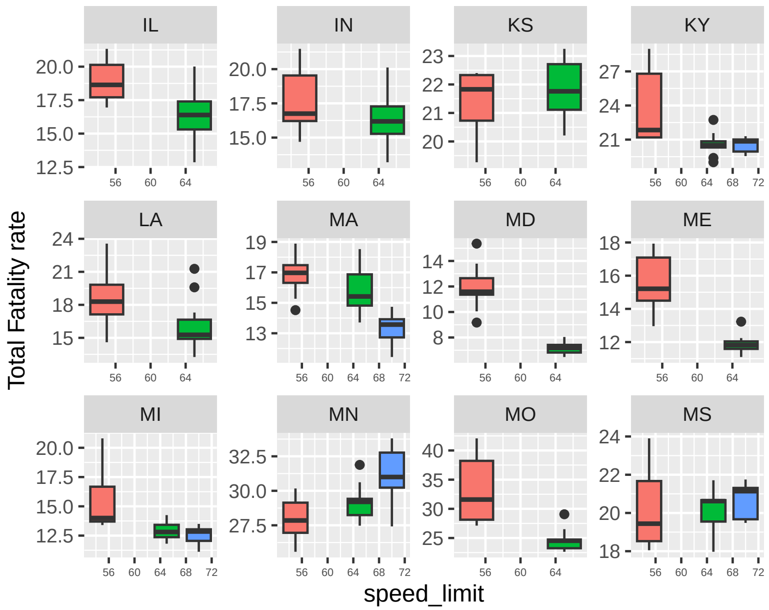
Again, MN seems to have positive and opposite correlation of other states for speed limit over 70 mph. In some states, we see lower total fatality rates when there is a speed limit over 70. Below is a correlation plot for better visualization.

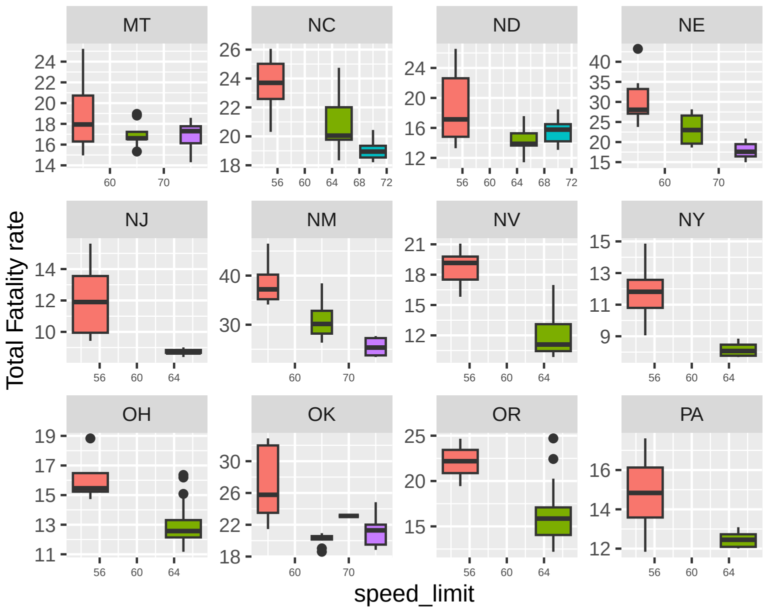
```r
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       speed_limit_over_70,
                                       use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title = "Correlation between
       Fatality Rate and speed_limit_over_70 by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

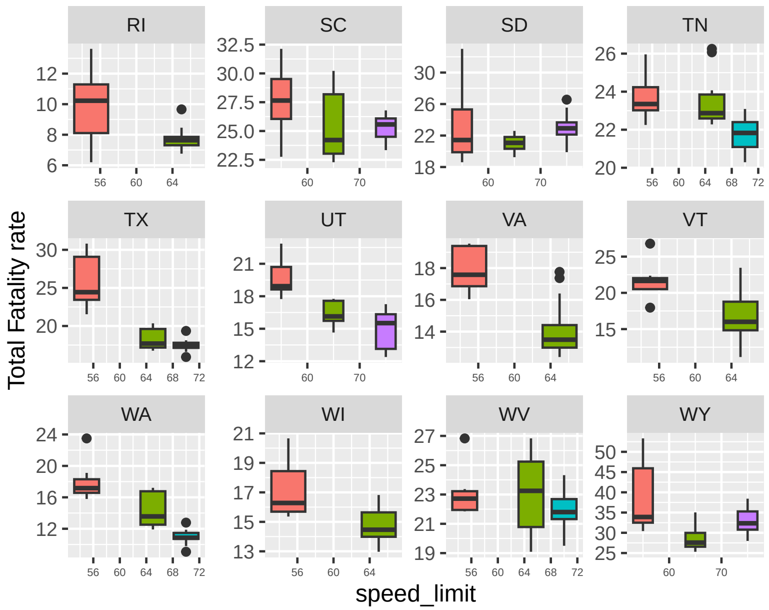


```r
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit, y = total_fatalities_rate,
             fill = factor(speed_limit))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "speed_limit",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit, y = total_fatalities_rate,
             fill = factor(speed_limit))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "speed_limit",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit, y = total_fatalities_rate,
             fill = factor(speed_limit))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "speed_limit",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = speed_limit, y = total_fatalities_rate,
             fill = factor(speed_limit))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "speed_limit",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
speed_limit
56
60
64
68
72
70

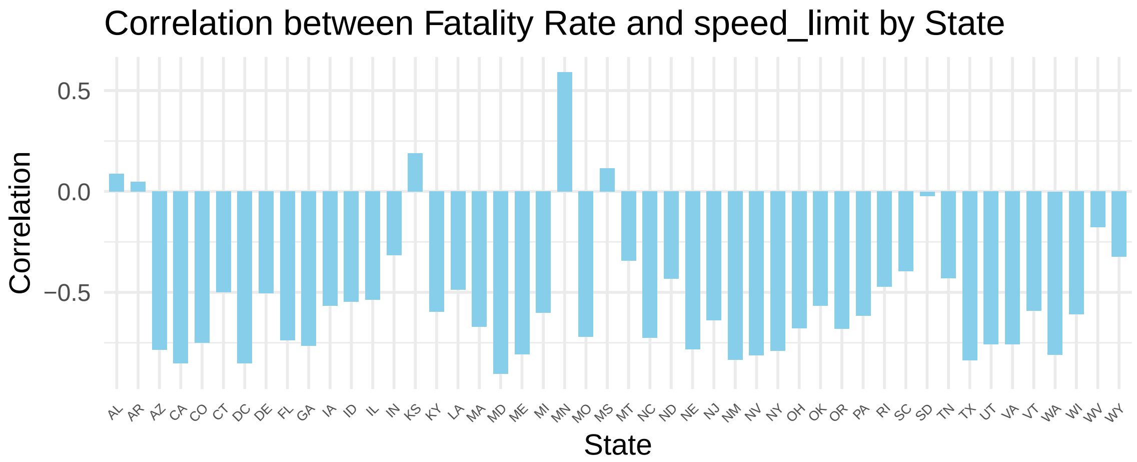

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
speed_limit
56
60
64
68
72


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
speed_limit
56
60
64
68
72
70


RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
speed_limit
56
60
64
68
72
70

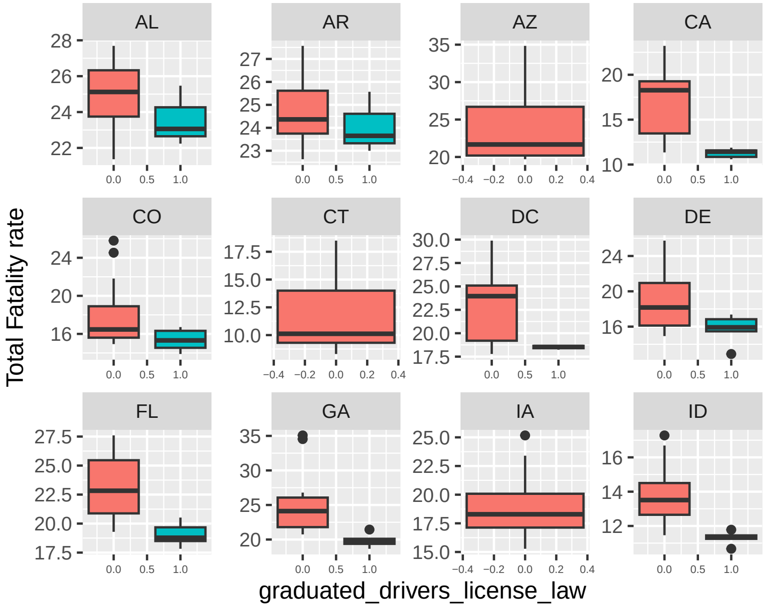
We see that for some states, the total fatality rates when the speed limit is over 75 are generally lower than rates when the speed limit is 55, 65, or 70. GA and MN seem to have opposite correlations than those of other states for speed limits over 70 mph. In some states, we see lower total fatality rates when there is a speed limit over 70. Below is a correlation plot for better visualization.

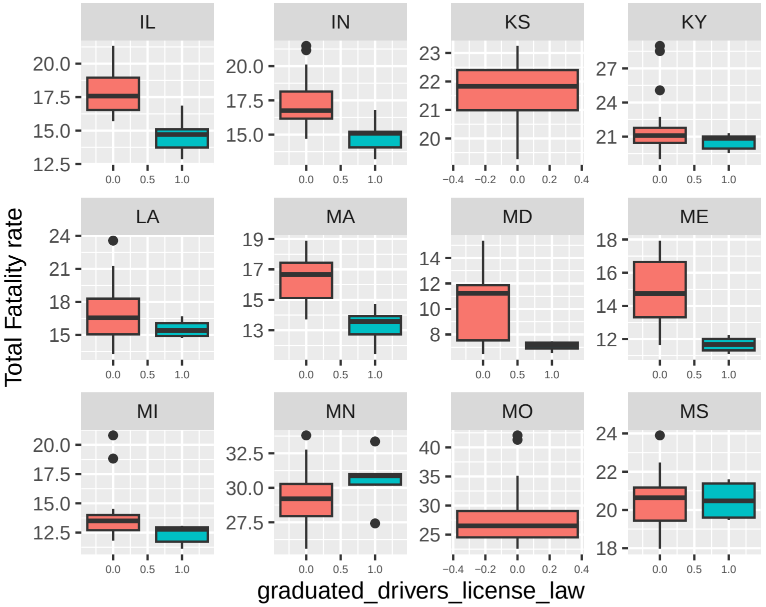
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate, speed_limit,
                                       use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title = "Correlation between Fatality Rate and speed_limit by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

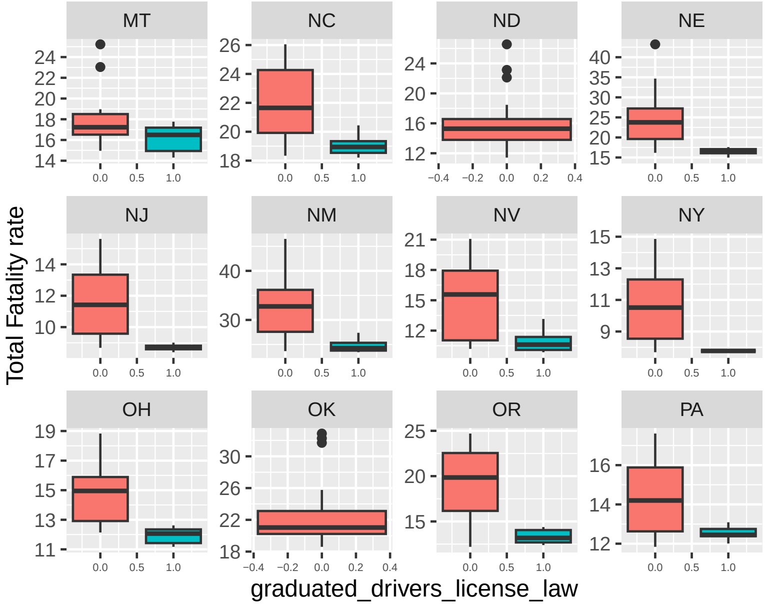


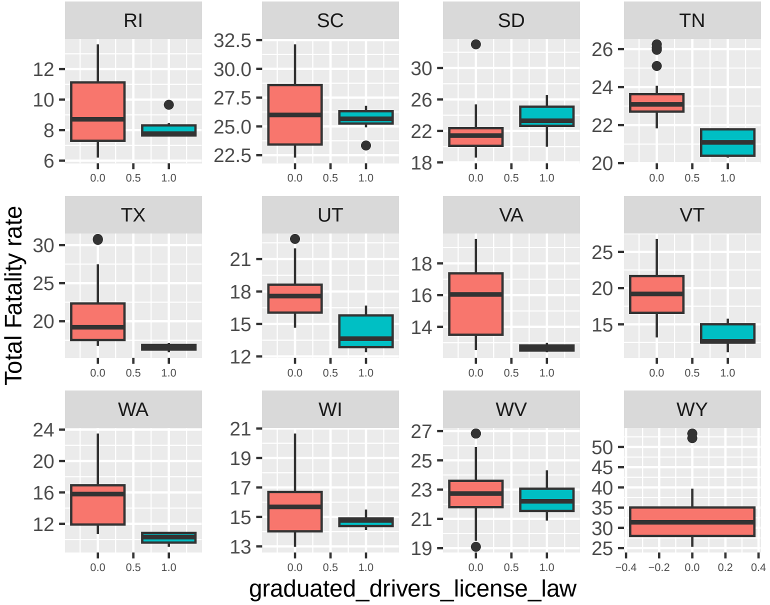
```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = graduated_drivers_license_law, y = total_fatalities_rate,
             fill = factor(graduated_drivers_license_law))) +
  geom_boxplot() +
```

```r
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "graduated_drivers_license_law",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = graduated_drivers_license_law, y = total_fatalities_rate,
             fill = factor(graduated_drivers_license_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "graduated_drivers_license_law",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = graduated_drivers_license_law, y = total_fatalities_rate,
             fill = factor(graduated_drivers_license_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "graduated_drivers_license_law",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = graduated_drivers_license_law, y = total_fatalities_rate,
             fill = factor(graduated_drivers_license_law))) +
  geom_boxplot() +
  theme(legend.position = "none", axis.text.x = element_text(size = 5)) +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "graduated_drivers_license_law",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


Total Fatality rate
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
graduated_drivers_license_law

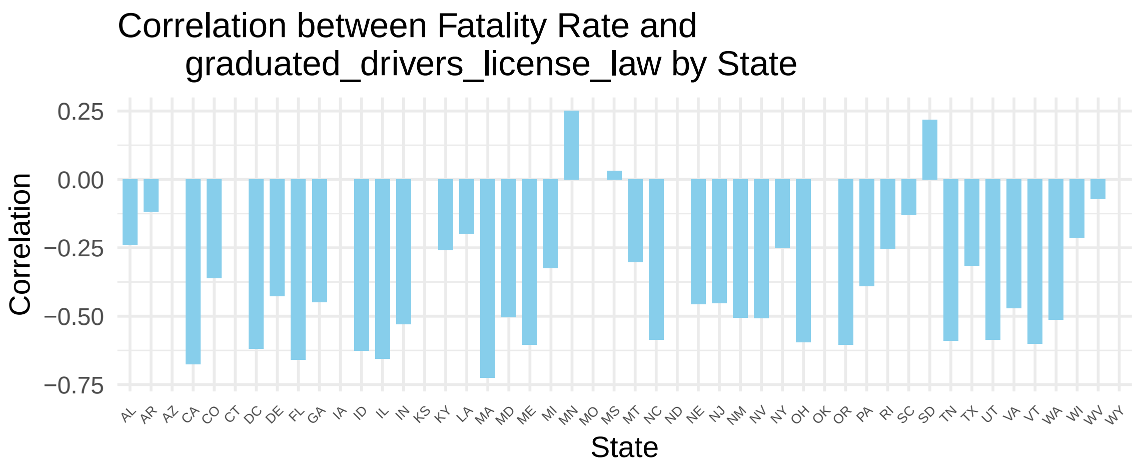

Total Fatality rate
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
graduated_drivers_license_law


Total Fatality rate
MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
graduated_drivers_license_law

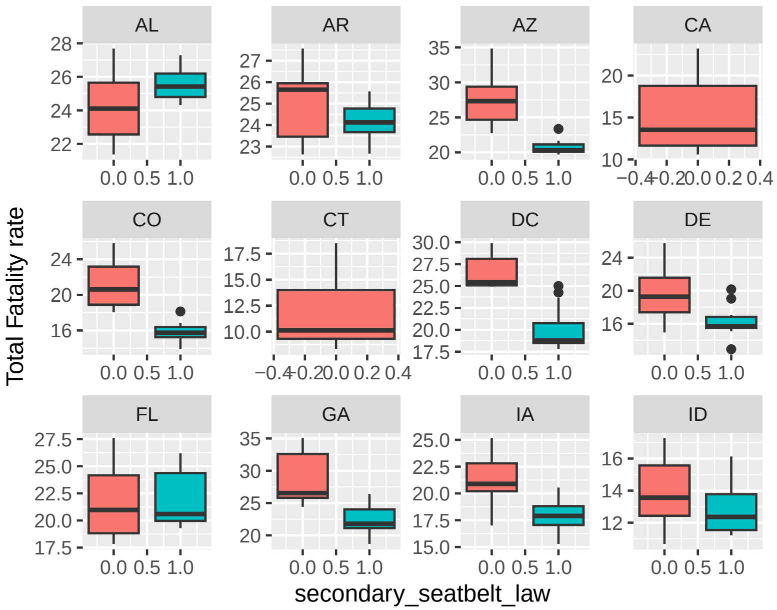

Total Fatality rate
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
graduated_drivers_license_law

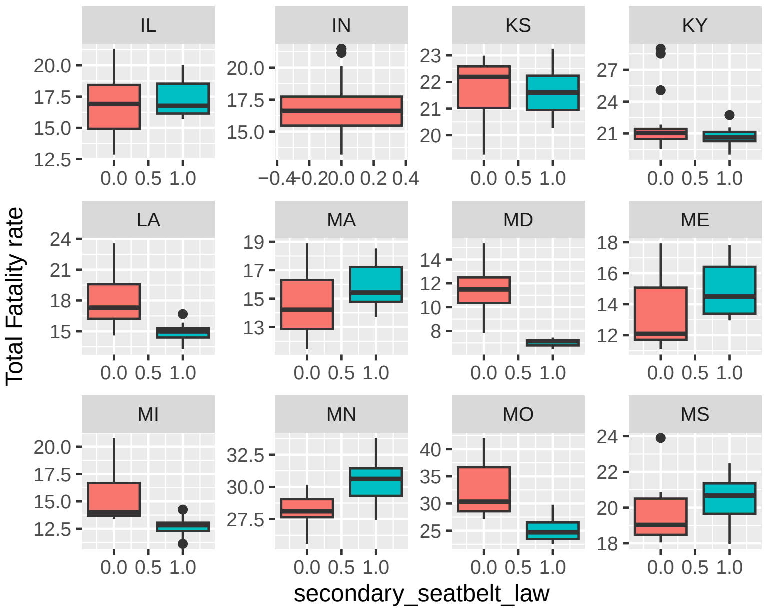
Graduated drivers license laws introduce a ‘tiered’ system for young adults earning their license. First, prospective young drivers get a learner’s permit, allowing them to drive with adult supervision. They then get an intermediate license, which restricts giving others rides without having a fully licensed adult in the car. Finally, a full license is given after certain criteria are met. Age restrictions and criteria vary by state, but this is the general structure of a graduated drivers license law. The motivation is to have new drivers gain some experience driving by themselves before being able to drive others in social situations. From our EDA, we see a negative correlation between fatality rate and the presence of graduated drivers license laws, giving some credence to this idea.

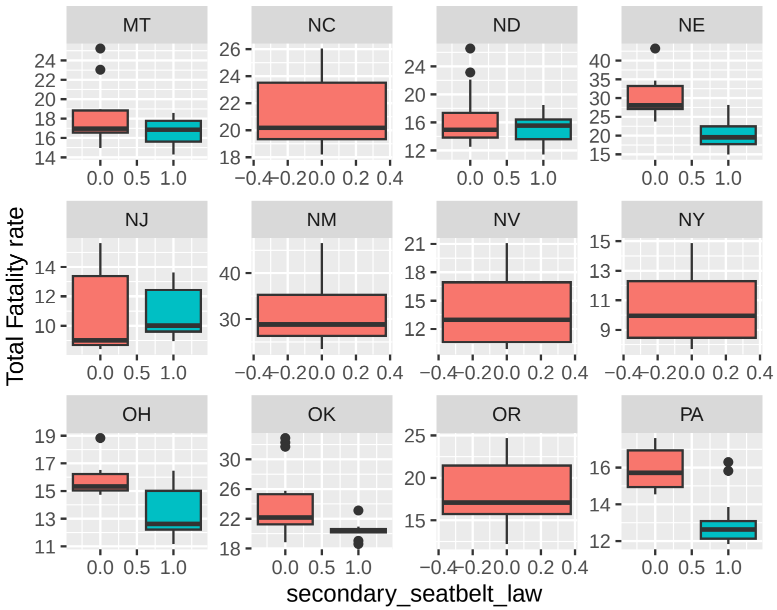
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       graduated_drivers_license_law,
                                       use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title = "Correlation between Fatality Rate and
       graduated_drivers_license_law by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

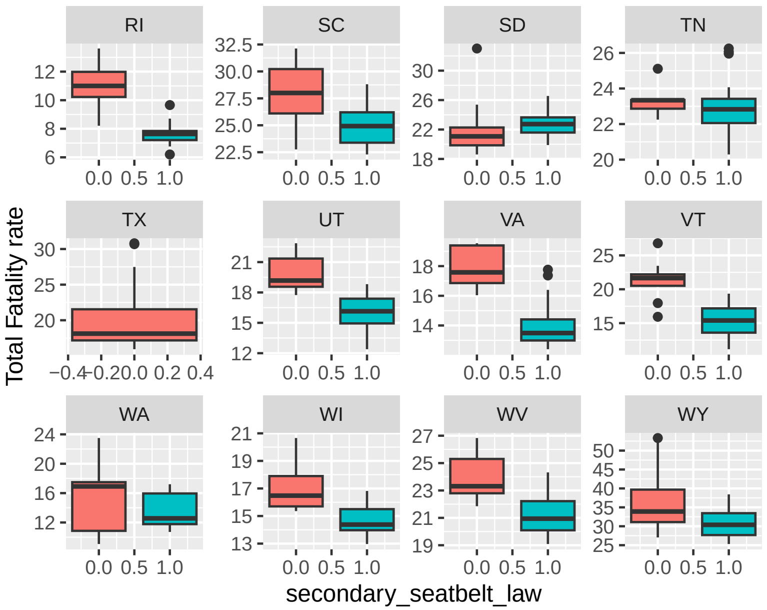


```r
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = secondary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(secondary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "secondary_seatbelt_law",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = secondary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(secondary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "secondary_seatbelt_law",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = secondary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(secondary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "secondary_seatbelt_law",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = secondary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(secondary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "secondary_seatbelt_law",  y = "Total Fatality rate")

grid.arrange(p18,p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
secondary_seatbelt_law
0.0 0.5 1.0
−0.4−0.2 0.0 0.2 0.4

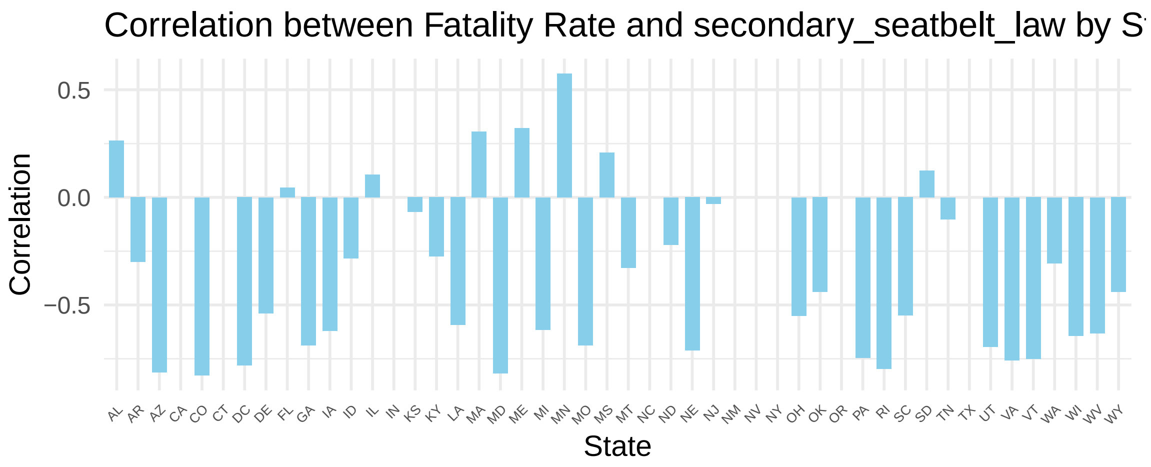

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
secondary_seatbelt_law
0.0 0.5 1.0
−0.4−0.2 0.0 0.2 0.4


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
secondary_seatbelt_law
0.0 0.5 1.0
−0.4−0.2 0.0 0.2 0.4

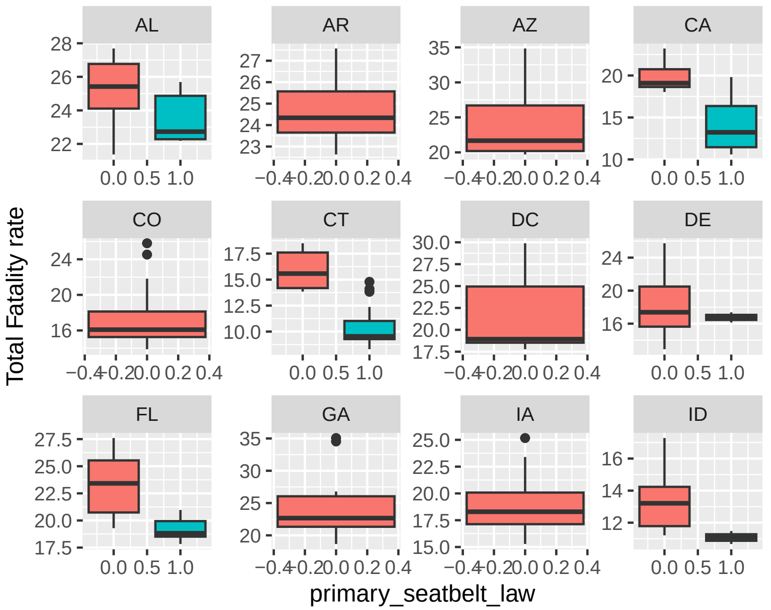

RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
secondary_seatbelt_law
0.0 0.5 1.0
−0.4−0.2 0.0 0.2 0.4

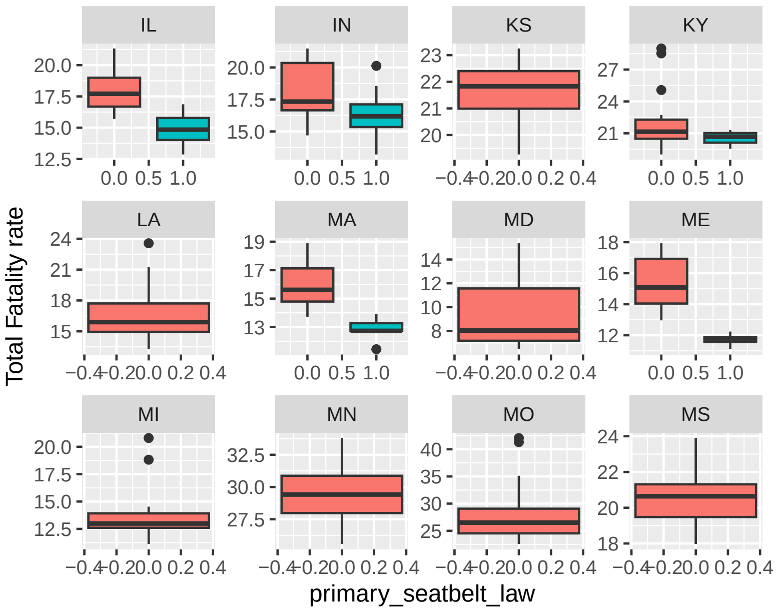
There are two types of seatbelt laws: primary and secondary. A primary seatbelt law allows officers to pull people over and cite them for simply not wearing a seatbelt. A secondary seatbelt law allows a citation only after pulling someone over for another reason. We can see the presence of correlation between the fatality rate and the presence of secondary seatbelt laws. The magnitude of correlation is higher for primary seatbelt laws.

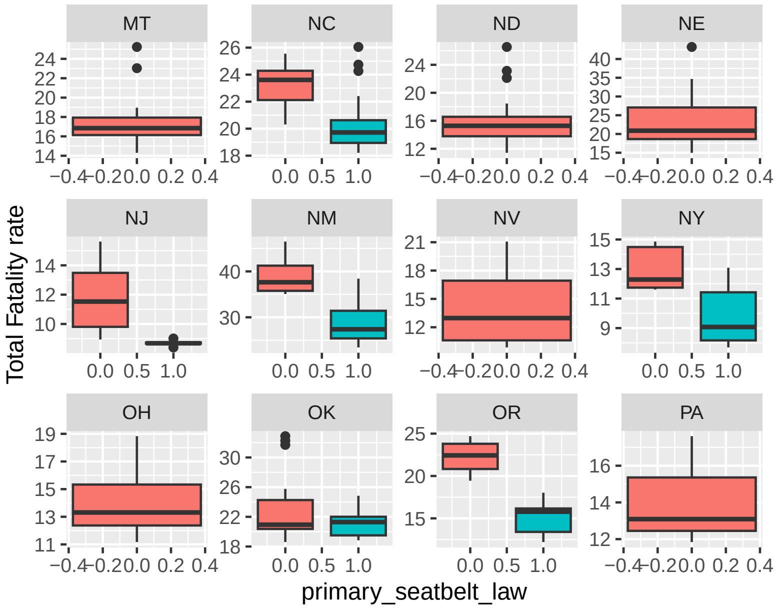
```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       secondary_seatbelt_law,
                                       use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title =
         "Correlation between Fatality Rate and secondary_seatbelt_law by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```

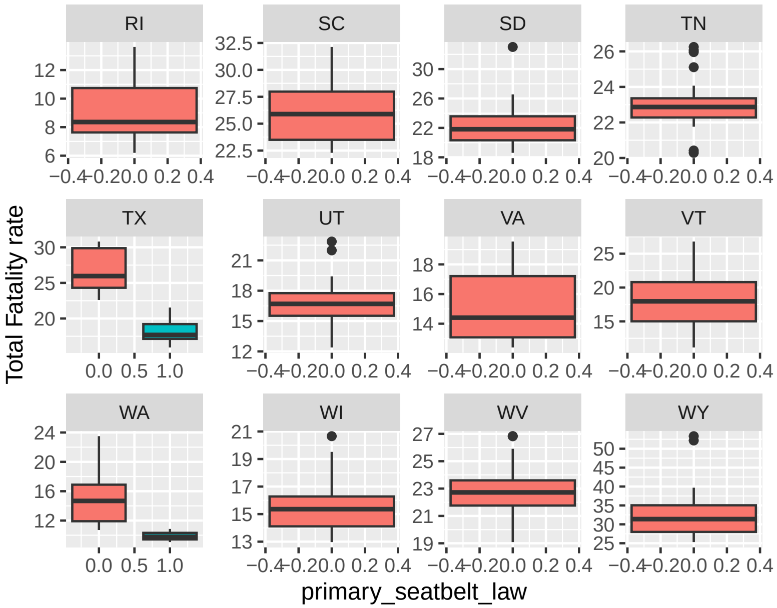


```
p18<- ptraffic_fatalities %>%
  filter(as.integer(state) <= 12 ) %>%
  data.frame() %>%
  ggplot(aes(x = primary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(primary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free") +
  labs(x = "primary_seatbelt_law",  y = "Total Fatality rate")

p19<- ptraffic_fatalities %>%
  filter(as.integer(state) > 12 & as.integer(state) <= 24 ) %>%
  data.frame() %>%
  ggplot(aes(x = primary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(primary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "primary_seatbelt_law",  y = "Total Fatality rate")

p20<- ptraffic_fatalities %>%
  filter(as.integer(state) > 24 &as.integer(state) <= 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = primary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(primary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "primary_seatbelt_law",  y = "Total Fatality rate")

p21<- ptraffic_fatalities %>%
  filter(as.integer(state) > 36 ) %>%
  data.frame() %>%
  ggplot(aes(x = primary_seatbelt_law, y = total_fatalities_rate,
             fill = factor(primary_seatbelt_law))) +
  geom_boxplot() +
  theme(legend.position = "none") +
  facet_wrap(~ state, nrow = 3,scales = "free")+
  labs(x = "primary_seatbelt_law",  y = "Total Fatality rate")

grid.arrange(p18, p19,p20, p21, nrow = 2, ncol = 2)
```


AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
IA
ID
Total Fatality rate
primary_seatbelt_law

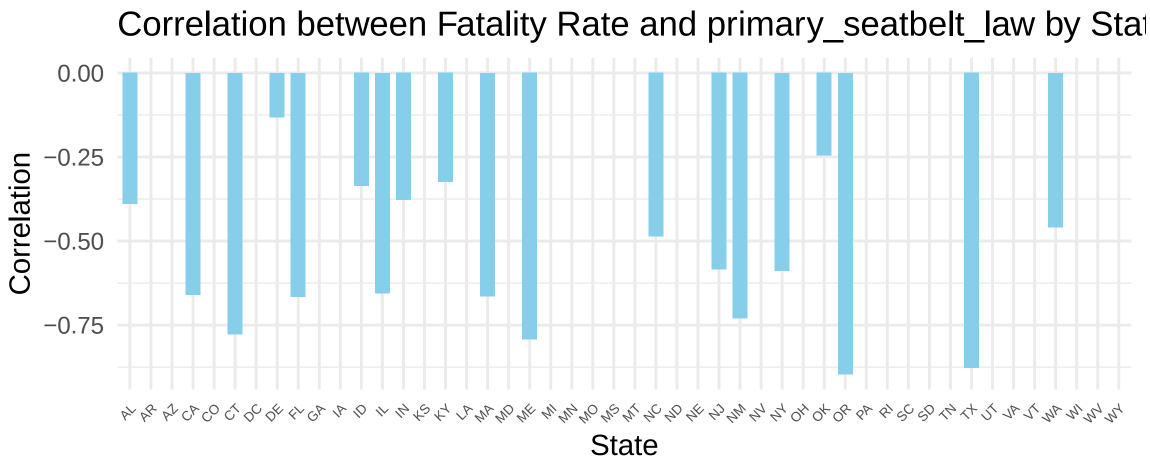

IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
Total Fatality rate
primary_seatbelt_law


MT
NC
ND
NE
NJ
NM
NV
NY
OH
OK
OR
PA
Total Fatality rate
primary_seatbelt_law


RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Total Fatality rate
primary_seatbelt_law

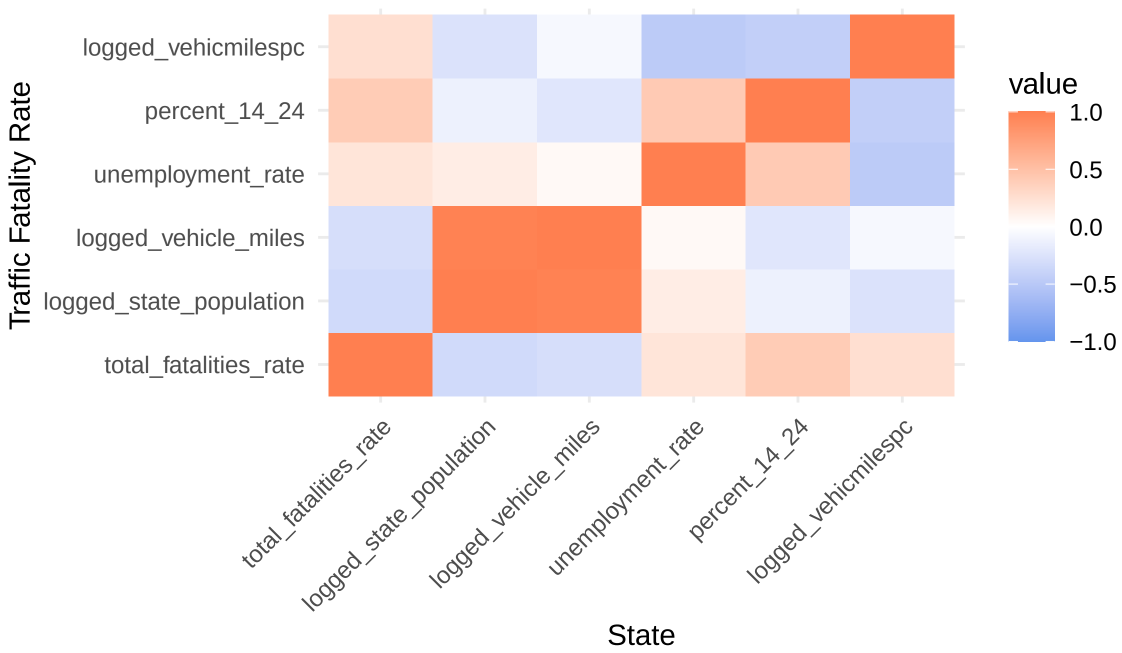
It appears from the above graphs that the range of total fatality rates is generally lower when there is a primary seatbelt law in place for some states. Wherever there is a change in primary seatbelt laws, there will be lower total fatality rates.

```
correlation_by_state <- ptraffic_fatalities %>%
    group_by(state) %>%
    dplyr::summarize(correlation = cor(total_fatalities_rate,
                                       primary_seatbelt_law,
                                       use = "complete.obs"))

ggplot(data = correlation_by_state, aes(x = state, y = correlation)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(x = "State", y = "Correlation",
       title
       = "Correlation between Fatality Rate and primary_seatbelt_law by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5))
```



```
selected_data <- ptraffic_fatalities %>%
  select(c(total_fatalities_rate, state_population, vehicle_miles,
           unemployment_rate, percent_14_24, vehicmilespc))

selected_data$logged_vehicmilespc <- log(selected_data$vehicmilespc)
```

```r
selected_data$logged_state_population <- log(selected_data$state_population)
selected_data$logged_vehicle_miles <- log(selected_data$vehicle_miles)

selected_data <- selected_data %>% select(total_fatalities_rate,
                                          logged_state_population, logged_vehicle_miles,
           unemployment_rate, percent_14_24, logged_vehicmilespc)

correlation_matrix <- selected_data %>%
  cor()

ggplot(melt(correlation_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile() +
  theme_minimal() +
  scale_fill_gradient2(low = "cornflowerblue", high = "coral", mid = "white",
                       midpoint = 0, limit = c(-1, 1)) +
  xlab("State") +
  ylab("Traffic Fatality Rate") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```

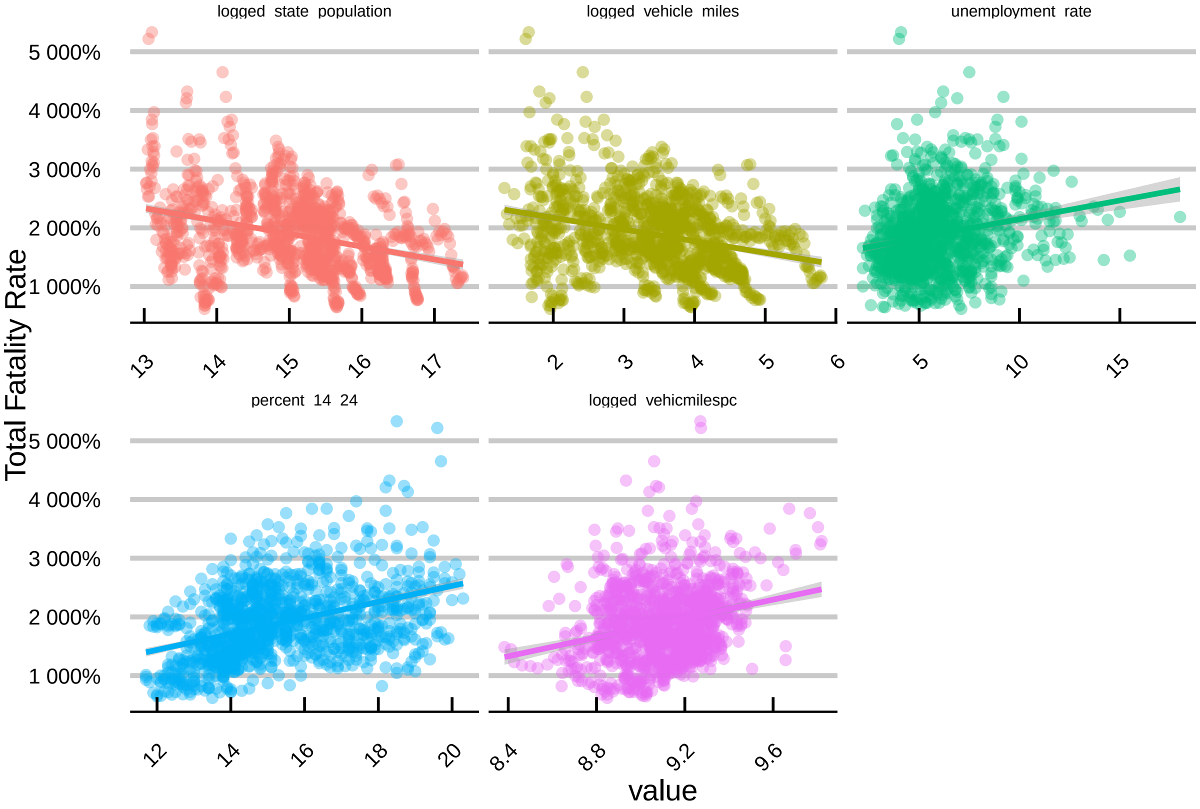


Certain variables like `logged_state_population` and `logged_vehicle_miles` seem highly correlated while others like `unemployment_rate` and `vehicle_miles` are not correlated. Total traffic fatality rate seems to be positively correlated with `unemployment_rate`, `percent_14_24`, `logged_vehiclemilespc` and somewhat negatively correlated with `logged_state_population` and `logged_vehicle_miles`.

To summarise, let's look at the relationship of all our variables of interest and check which variables affect traffic fatality rate, combined across all states.

```r
# Melt the data into a long format
melted_data <- selected_data %>%
  melt(id.vars=c("total_fatalities_rate"))

# Plot the data
ggplot(melted_data, aes(x = value, y = total_fatalities_rate,
                        color = variable)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  facet_wrap(~variable, scales = "free_x") +
  theme_economist_white(gray_bg = F) +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45,
                                                             hjust = 1,
                                                             vjust = 1,
                                                             size = 8),
        axis.text.y = element_text(size = 8)) +
  theme(strip.text = element_text(size = 6)) +
  scale_y_continuous(label = percent) +
  ylab("Total Fatality Rate")
```

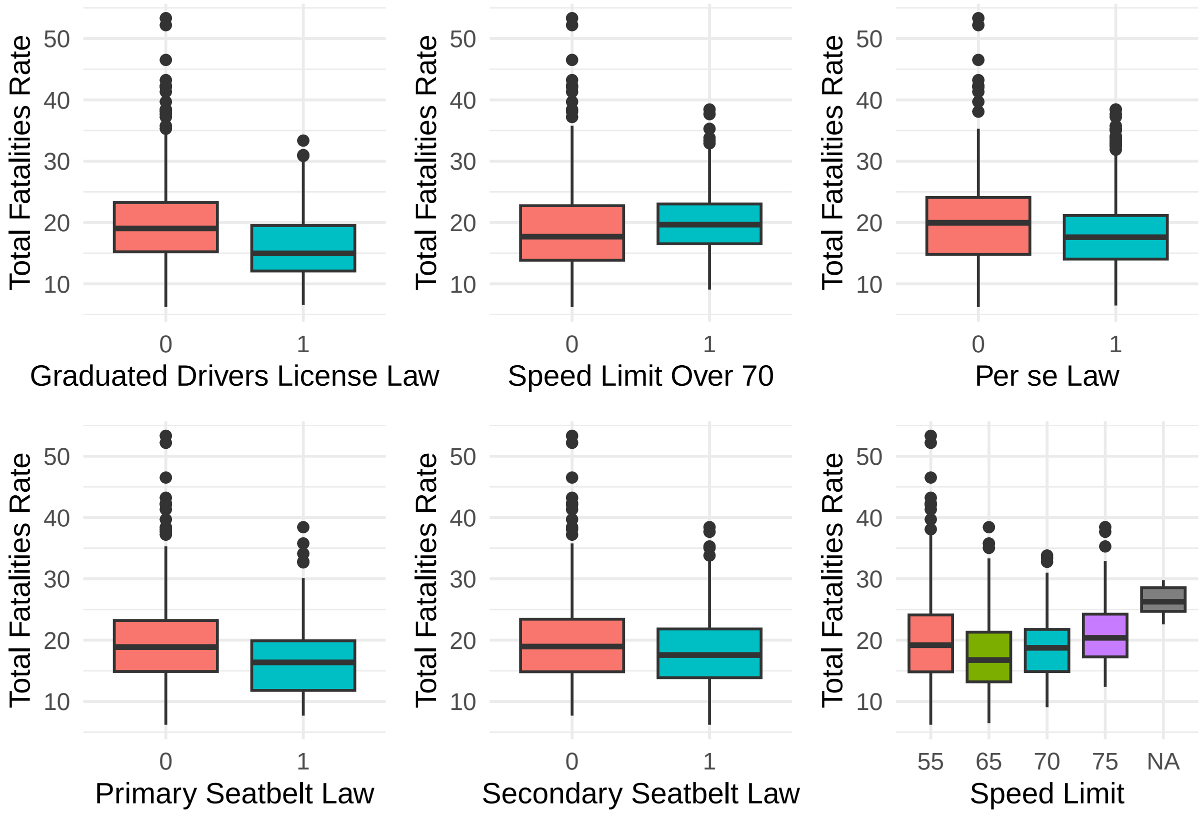
```
## `geom_smooth()` using formula = 'y ~ x'
```



It looks like `logged_state_population` and `logged_vehicle_miles` affect traffic fatality rate negatively while `unemployment_rate`, `percent_14_24` and `logged_vehiclemilespc` affect traffic fatality rate positively.

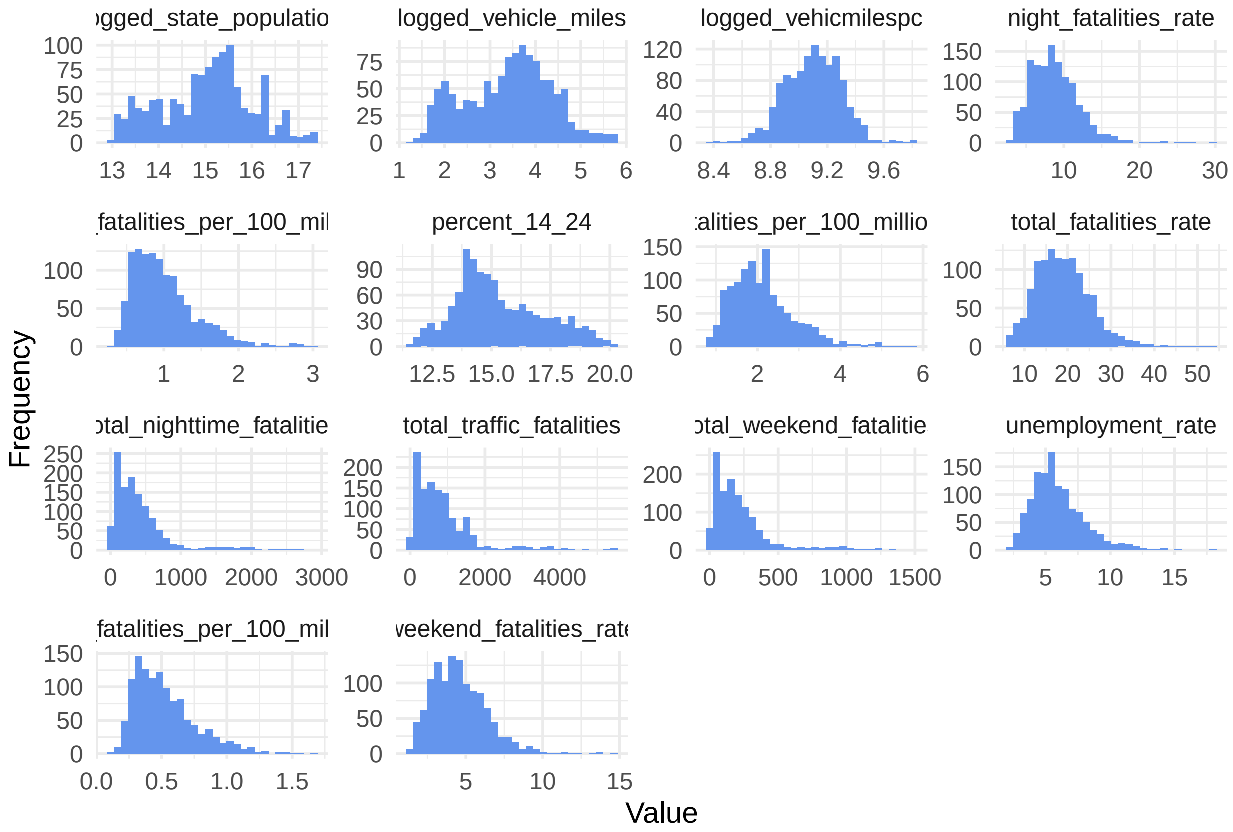
```
p1 <- ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = factor(graduated_drivers_license_law), y =
               total_fatalities_rate,
             fill = factor(graduated_drivers_license_law))) +
  geom_boxplot() +
  labs(x = "Graduated Drivers License Law", y = "Total Fatalities Rate") +
  theme_minimal() +
  theme(legend.position = "none")

p2 <- ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = factor(speed_limit_over_70), y = total_fatalities_rate,
             fill = factor(speed_limit_over_70))) +
  geom_boxplot() +
  labs(x = "Speed Limit Over 70", y = "Total Fatalities Rate") +
  theme_minimal() +
  theme(legend.position = "none")

p3 <- ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = factor(per_se_law), y = total_fatalities_rate,
             fill = factor(per_se_law))) +
  geom_boxplot() +
  labs(x = "Per se Law", y = "Total Fatalities Rate") +
  theme_minimal() +
  theme(legend.position = "none")

p4 <- ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = factor(primary_seatbelt_law), y = total_fatalities_rate,
             fill = factor(primary_seatbelt_law))) +
  geom_boxplot() +
  labs(x = "Primary Seatbelt Law", y = "Total Fatalities Rate") +
  theme_minimal() +
  theme(legend.position = "none")
```

```r
p5 <- ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = factor(secondary_seatbelt_law), y = total_fatalities_rate,
             fill = factor(secondary_seatbelt_law))) +
  geom_boxplot() +
  labs(x = "Secondary Seatbelt Law", y = "Total Fatalities Rate") +
  theme_minimal() +
  theme(legend.position = "none")

p6 <- ptraffic_fatalities %>%
  data.frame() %>%
  ggplot(aes(x = factor(speed_limit), y = total_fatalities_rate,
             fill = factor(speed_limit))) +
  geom_boxplot() +
  labs(x = "Speed Limit", y = "Total Fatalities Rate") +
  theme_minimal() +
  theme(legend.position = "none")

(p1 | p2 | p3) /
  (p4 | p5 | p6)
```



Overall, across all the states, we see a decrease in total fatality rate when `graduated_drivers_license_law`, `primary_seatbelt_law`, `secondary_seatbelt_law` and `per_se_law` were modified. What's interesting for the speed variables from the above plots is that there seems to be an increase in fatality rate at higher speed limits overall.

```
numeric_vars <- c("total_traffic_fatalities", "total_nighttime_fatalities",
                  "total_weekend_fatalities",
                  "total_fatalities_per_100_million_miles",
                  "nighttime_fatalities_per_100_million_miles",
                  "weekend_fatalities_per_100_million_miles",
                  "state_population", "total_fatalities_rate",
                  "night_fatalities_rate", "weekend_fatalities_rate",
                  "vehicle_miles", "unemployment_rate", "percent_14_24",
                  "vehicmilespc")

# Select numeric variables and convert to long format
long_data <- ptraffic_fatalities %>%
  select(numeric_vars)
long_data$logged_vehicmilespc <- log(long_data$vehicmilespc)
long_data$logged_state_population <- log(long_data$state_population)
long_data$logged_vehicle_miles <- log(long_data$vehicle_miles)
long_data <- long_data %>%
  select(total_traffic_fatalities, total_nighttime_fatalities,
         total_weekend_fatalities, total_fatalities_per_100_million_miles,
         nighttime_fatalities_per_100_million_miles,
         weekend_fatalities_per_100_million_miles, logged_state_population,
         total_fatalities_rate, night_fatalities_rate, weekend_fatalities_rate,
         logged_vehicle_miles, unemployment_rate, percent_14_24,
         logged_vehicmilespc)

long_data <- long_data %>%
  gather(key = "variable", value = "value")


# Plot histograms
ggplot(long_data, aes(x = value)) +
  geom_histogram(fill = "cornflowerblue") +
  facet_wrap(~ variable, scales = "free") +
  theme_minimal() +
  labs(x = "Value", y = "Frequency")
```



Lastly, here is the overall distribution of the data before we move to the modeling.

(15 points) Preliminary Model

Estimate a linear regression model of *totfatrte* on a set of dummy variables for the years 1981 through 2004 and interpret what you observe. In this section, you should address the following tasks:

- Why is fitting a linear model a sensible starting place?
- What does this model explain, and what do you find in this model?
- Did driving become safer over this period? Please provide a detailed explanation.

- What, if any, are the limitation of this model. In answering this, please consider **at least**:
  - Are the parameter estimates reliable, unbiased estimates of the truth? Or, are they biased due to the way that the data is structured?
  - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

```
pooled_ols_year <- plm(total_fatalities_rate ~ year_of_observation,
    index= c("state", "year_of_observation"),
    data = ptraffic_fatalities,
    model="pooling")

summary(pooled_ols_year)
```

```
## Pooling Model
##
## Call:
## plm(formula = total_fatalities_rate ~ year_of_observation, data = ptraffic_fatalities,
##     model = "pooling", index = c("state", "year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -12.93021  -4.34682  -0.73052   3.74875  29.64979
##
## Coefficients:
##                          Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)             25.49458    0.86712 29.4015 < 2.2e-16 ***
## year_of_observation1981 -1.82438    1.22629 -1.4877 0.1370936
## year_of_observation1982 -4.55208    1.22629 -3.7121 0.0002152 ***
## year_of_observation1983 -5.34167    1.22629 -4.3560 1.440e-05 ***
## year_of_observation1984 -5.22708    1.22629 -4.2625 2.183e-05 ***
## year_of_observation1985 -5.64313    1.22629 -4.6018 4.644e-06 ***
## year_of_observation1986 -4.69417    1.22629 -3.8279 0.0001360 ***
## year_of_observation1987 -4.71979    1.22629 -3.8488 0.0001251 ***
## year_of_observation1988 -4.60292    1.22629 -3.7535 0.0001829 ***
## year_of_observation1989 -5.72229    1.22629 -4.6663 3.418e-06 ***
## year_of_observation1990 -5.98938    1.22629 -4.8841 1.182e-06 ***
## year_of_observation1991 -7.39979    1.22629 -6.0343 2.137e-09 ***
## year_of_observation1992 -8.33667    1.22629 -6.7983 1.681e-11 ***
## year_of_observation1993 -8.36688    1.22629 -6.8229 1.425e-11 ***
## year_of_observation1994 -8.33938    1.22629 -6.8005 1.656e-11 ***
## year_of_observation1995 -7.82604    1.22629 -6.3819 2.512e-10 ***
## year_of_observation1996 -8.12521    1.22629 -6.6258 5.246e-11 ***
## year_of_observation1997 -7.88396    1.22629 -6.4291 1.863e-10 ***
## year_of_observation1998 -8.22917    1.22629 -6.7106 3.007e-11 ***
## year_of_observation1999 -8.24417    1.22629 -6.7228 2.774e-11 ***
## year_of_observation2000 -8.66896    1.22629 -7.0692 2.666e-12 ***
## year_of_observation2001 -8.70188    1.22629 -7.0961 2.214e-12 ***
## year_of_observation2002 -8.46500    1.22629 -6.9029 8.316e-12 ***
## year_of_observation2003 -8.73104    1.22629 -7.1199 1.877e-12 ***
## year_of_observation2004 -8.76563    1.22629 -7.1481 1.542e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48612
## Residual Sum of Squares: 42407
## R-Squared:      0.12765
## Adj. R-Squared: 0.10983
## F-statistic: 7.16387 on 24 and 1175 DF, p-value: < 2.22e-16
```

This model ignores the panel data structure and assumes that each of the data points are independent. From this model, almost all the variables are statistically significant except for `year_of_observation1981`. We can also see a decreasing trend of the coefficient estimates for the dummy variable, which means that the total fatalities rate is decreasing as the years progress.

- Why is fitting a linear model a sensible starting place?

Fitting a linear model provides us with a baseline model which we can build off of later for a better fitting model. This also gives us a brief sense of the data. Fitting a stargazer in the end with the output from all other the models we've tried would give us a sense of how far we've come with the modeling by not ignoring panel structured data.

- What does this model explain, and what do you find in this model?

The model shows us the estimates for total fatality rate for each of the year, with almost all of them being statistically significant. This model fully ignores the other variables holding explanatory power in the data. The model implies that the year of observation is causing a decrease in total fatality rate, except for the year 1981. The output from this model is biased as it assumes the OLS assumptions are satisfied (such as iid, linearity), which is not the case completely, leading to incorrect interpretation.

- Did driving become safer over this period? Please provide a detailed explanation.

Yes, it does look like driving became safer over this period according to this preliminary model. We can see that as the years progress, the coefficient of `year_of_obeservation*` is decreasing. That is all this model can provide us with since there are no other explanatory variables involved.

- What, if any, are the limitation of this model. In answering this, please consider **at least**:
  - Are the parameter estimates reliable, unbiased estimates of the truth? Or, are they biased due to the way that the data is structured?
  - Are the uncertainty estimate reliable, unbiased estimates of sampling based variability? Or, are they biased due to the way that the data is structured?

The pooled linear regression model requires that the records in the data be independent and identically distributed. The independence assumption is violated since the panel structure of the data means that multiple records exist for each state, and a certain state may have a consistent range of total fatality rates. This results in clustering, which can negatively impact the validity of the model. Next, we are only considering year of observation, which means that other variables that could potentially hold significant explanatory power - like the presence of alcohol laws and unemployment rate - are omitted from the model. Hence, we have omitted variable bias. We also do not consider the unobserved time-invariant state-specific effects (intercepts) while using this model, again allowing it to possess omitted variable bias. If these effects are correlated with the dummy variables, we would have positive or negative bias. If they are not correlated, these effects are added to the residuals therefore making our standard errors wrong.

Therefore, the parameter estimates are not reliable. They are biased due to the way the data and model is structured. The same applies to the uncertainty estimates.

(15 points) Expanded Model

Expand the **Preliminary Model** by adding variables related to the following concepts:

- Blood alcohol levels
- Per se laws
- Primary seat belt laws (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.)
- Secondary seat belt laws
- Speed limits faster than 70
- Graduated drivers licenses
- Percent of the population between 14 and 24 years old
- Unemployment rate
- Vehicle miles driven per capita.

If it is appropriate, include transformations of these variables. Please carefully 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 blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.
- Do *per se laws* have a negative effect on the fatality rate?
- Does having a primary seat belt law?

```r
expanded_model <- plm(total_fatalities_rate ~ year_of_observation +
                           bac08 +
                           bac10 +
                           per_se_law +
                           primary_seatbelt_law +
                           secondary_seatbelt_law +
                           speed_limit_over_70 +
                           graduated_drivers_license_law +
                           percent_14_24 +
                           unemployment_rate +
                           log(vehicmilespc),
    index=c("state, year_of_observation"),
   data=ptraffic_fatalities,
   model="pooling")

summary(expanded_model)
```

```
## Pooling Model
##
## Call:
## plm(formula = total_fatalities_rate ~ year_of_observation + bac08 +
##     bac10 + per_se_law + primary_seatbelt_law + secondary_seatbelt_law +
##     speed_limit_over_70 + graduated_drivers_license_law + percent_14_24 +
##     unemployment_rate + log(vehicmilespc), data = ptraffic_fatalities,
##     model = "pooling", index = c("state, year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -11.64559  -2.60708  -0.31874   2.34551  20.38756
```

```
##
## Coefficients:
##                                  Estimate Std. Error  t-value  Pr(>|t|)    
## (Intercept)                   -234.597773   7.859222 -29.8500 < 2.2e-16 ***
## year_of_observation1981         -2.220465   0.813407  -2.7298  0.006432 ** 
## year_of_observation1982         -6.876396   0.839433  -8.1917 6.716e-16 ***
## year_of_observation1983         -7.900288   0.857227  -9.2161 < 2.2e-16 ***
## year_of_observation1984         -6.488094   0.862316  -7.5240 1.059e-13 ***
## year_of_observation1985         -7.280247   0.879467  -8.2780 3.398e-16 ***
## year_of_observation1986         -6.852329   0.915374  -7.4858 1.398e-13 ***
## year_of_observation1987         -7.537278   0.952513  -7.9130 5.809e-15 ***
## year_of_observation1988         -7.848464   0.999828  -7.8498 9.392e-15 ***
## year_of_observation1989         -9.403590   1.038684  -9.0534 < 2.2e-16 ***
## year_of_observation1990        -10.339748   1.063643  -9.7211 < 2.2e-16 ***
## year_of_observation1991        -12.484542   1.089477 -11.4592 < 2.2e-16 ***
## year_of_observation1992        -14.366602   1.110149 -12.9411 < 2.2e-16 ***
## year_of_observation1993        -14.159857   1.123912 -12.5987 < 2.2e-16 ***
## year_of_observation1994        -13.713986   1.144336 -11.9842 < 2.2e-16 ***
## year_of_observation1995        -13.290918   1.167297 -11.3861 < 2.2e-16 ***
## year_of_observation1996        -15.373006   1.213175 -12.6717 < 2.2e-16 ***
## year_of_observation1997        -15.366648   1.232418 -12.4687 < 2.2e-16 ***
## year_of_observation1998        -16.140792   1.244721 -12.9674 < 2.2e-16 ***
## year_of_observation1999        -16.023094   1.262923 -12.6873 < 2.2e-16 ***
## year_of_observation2000        -16.291474   1.280963 -12.7181 < 2.2e-16 ***
## year_of_observation2001        -16.802797   1.309439 -12.8321 < 2.2e-16 ***
## year_of_observation2002        -17.285529   1.321777 -13.0775 < 2.2e-16 ***
## year_of_observation2003        -17.653802   1.328635 -13.2872 < 2.2e-16 ***
## year_of_observation2004        -17.324184   1.351368 -12.8197 < 2.2e-16 ***
## bac08                           -2.013943   0.482237  -4.1763 3.184e-05 ***
## bac10                           -0.953493   0.356022  -2.6782  0.007506 ** 
## per_se_law                      -0.760129   0.286958  -2.6489  0.008184 ** 
## primary_seatbelt_law             0.222187   0.483305   0.4597  0.645800    
## secondary_seatbelt_law           0.273116   0.421926   0.6473  0.517560    
## speed_limit_over_70              3.215077   0.425285   7.5598 8.149e-14 ***
## graduated_drivers_license_law   -0.935830   0.495152  -1.8900  0.059008 .  
## percent_14_24                    0.170533   0.120340   1.4171  0.156725    
## unemployment_rate                0.817532   0.076741  10.6532 < 2.2e-16 ***
## log(vehicmilespc)               28.395886   0.879511  32.2860 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    48612
## Residual Sum of Squares: 18417
## R-Squared:      0.62115
## Adj. R-Squared: 0.61009
## F-statistic: 56.1796 on 34 and 1165 DF, p-value: < 2.22e-16
```

- If it is appropriate, include transformations of these variables. Please carefully 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.

During our EDA, we found variables such as `state_population`, `vehicle_miles`, and `vehicmilespc` having significantly higher numerical ranges than those of other numerical variables such as unemployment_rate. To reduce the disparity, we applied a log transformation to make the ranges of the former variables smaller. This helps ensure that the modeling process is not heavily influenced by extreme values of a particular variable.

- How are the blood alcohol variables defined? Interpret the coefficients that you estimate for this concept.

The `bac` variables are two indicator variables for each legal driving BAC limit, 0.08 and 0.1. In our model estimates, we see that `bac08` is highly statistically significant with a negative coefficient of -2.013, which means that with all else being equal, states having a BAC limit of 0.08 decrease their total fatality rate by about 2.01 units keeping all else equal. Similarly, `bac10` is also statistically significant at the 0.01 level again with a negative coefficient of -0.953. This again means that states having a BAC limit of 0.10 decrease their total fatality rate by about 0.95 units keeping all else equal. Overall, it looks like states having a BAC limit of 0.08 decrease their total fatality rate better compared to states with a higher limit of 0.10.

- Do *per se laws* have a negative effect on the fatality rate?

Yes, per se law is statistically significant at the 0.01 level, having a negative effect on total fatality rate. The coefficient is -0.76. This means that holding all the variables constant, the `total_fatalities_rate` is expected to decrease by 0.76 units in states with per se laws compared to states without per se laws.

- Does having a primary seat belt law?

Weirdly, having a primary seat belt law appears to have an increasing effect on the total fatality rate, but it is not statistically significant. With this model, primary seat belt law estimates seem to hold no explanatory power.

(15 points) State-Level Fixed Effects

Re-estimate the **Expanded Model** using fixed effects at the state level.

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?
- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all?
- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

Which set of estimates do you think is more reliable? Why do you think this?

- What assumptions are needed in each of these models?

- Are these assumptions reasonable in the current context?

```r
within_model<- plm(total_fatalities_rate ~ year_of_observation +
                           bac08 +
                           bac10 +
                           per_se_law +
                           primary_seatbelt_law +
                           secondary_seatbelt_law +
                           speed_limit_over_70 +
                           graduated_drivers_license_law +
                           percent_14_24 +
                           unemployment_rate +
                           log(vehicmilespc),
                   data = ptraffic_fatalities,
                   index = c("state", "year_of_observation"),
                   effect = "individual",
                   model = "within")

summary(within_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = total_fatalities_rate ~ year_of_observation + bac08 +
##     bac10 + per_se_law + primary_seatbelt_law + secondary_seatbelt_law +
##     speed_limit_over_70 + graduated_drivers_license_law + percent_14_24 +
##     unemployment_rate + log(vehicmilespc), data = ptraffic_fatalities,
##     effect = "individual", model = "within", index = c("state",
##         "year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -7.762636 -0.982204  0.012005  0.937225 13.943234
##
## Coefficients:
##                                Estimate Std. Error  t-value  Pr(>|t|)
## year_of_observation1981        -1.551123   0.408288  -3.7991  0.000153 ***
## year_of_observation1982        -3.237791   0.438757  -7.3795 3.093e-13 ***
## year_of_observation1983        -3.927139   0.456162  -8.6091 < 2.2e-16 ***
## year_of_observation1984        -4.865732   0.462293 -10.5252 < 2.2e-16 ***
## year_of_observation1985        -5.475114   0.483421 -11.3258 < 2.2e-16 ***
## year_of_observation1986        -4.558120   0.517880  -8.8015 < 2.2e-16 ***
## year_of_observation1987        -5.387613   0.560507  -9.6120 < 2.2e-16 ***
## year_of_observation1988        -5.979103   0.610968  -9.7863 < 2.2e-16 ***
## year_of_observation1989        -7.466386   0.652021 -11.4511 < 2.2e-16 ***
## year_of_observation1990        -7.655581   0.680364 -11.2522 < 2.2e-16 ***
## year_of_observation1991        -8.380143   0.700115 -11.9697 < 2.2e-16 ***
## year_of_observation1992        -9.344970   0.724352 -12.9011 < 2.2e-16 ***
## year_of_observation1993        -9.682848   0.737160 -13.1353 < 2.2e-16 ***
## year_of_observation1994       -10.085779   0.755301 -13.3533 < 2.2e-16 ***
## year_of_observation1995        -9.922007   0.775481 -12.7946 < 2.2e-16 ***
## year_of_observation1996       -10.406306   0.824208 -12.6258 < 2.2e-16 ***
## year_of_observation1997       -10.541290   0.843479 -12.4974 < 2.2e-16 ***
## year_of_observation1998       -11.265398   0.857713 -13.1342 < 2.2e-16 ***
## year_of_observation1999       -11.396079   0.867769 -13.1326 < 2.2e-16 ***
## year_of_observation2000       -11.920979   0.878262 -13.5734 < 2.2e-16 ***
## year_of_observation2001       -11.585665   0.896342 -12.9255 < 2.2e-16 ***
## year_of_observation2002       -10.934216   0.908008 -12.0420 < 2.2e-16 ***
## year_of_observation2003       -11.026807   0.914294 -12.0605 < 2.2e-16 ***
## year_of_observation2004       -11.514777   0.930725 -12.3718 < 2.2e-16 ***
## bac08                          -0.721372   0.323170  -2.2322  0.025801 *
## bac10                          -0.583702   0.225312  -2.5906  0.009704 **
## per_se_law                    -1.095348   0.221950  -4.9351 9.227e-07 ***
## primary_seatbelt_law          -1.153566   0.338701  -3.4059  0.000683 ***
## secondary_seatbelt_law        -0.295780   0.248735  -1.1891  0.234637    
## speed_limit_over_70            0.135663   0.256593   0.5287  0.597112    
## graduated_drivers_license_law -0.451428   0.275826  -1.6366  0.101987    
## percent_14_24                  0.191798   0.093628   2.0485  0.040744 *  
## unemployment_rate             -0.529827   0.060371  -8.7763 < 2.2e-16 ***
## log(vehicmilespc)             12.130865   1.162685  10.4335 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4426.5
## R-Squared:      0.6352
## Adj. R-Squared: 0.60877
## F-statistic: 57.2557 on 34 and 1118 DF, p-value: < 2.22e-16
```

- What do you estimate for coefficients on the blood alcohol variables? How do the coefficients on the blood alcohol variables change, if at all?

The estimated coefficients in the within model for `bac08` is -0.721 and for `bac10` is -0.584. Comparing these values to that of the previous pooled model where `bac08` is -2.013 and `bac10` is -0.953, it looks like the coefficients of both the `bac*` variables have *increased* in the within model but are still negative. Additionally, the level of statistical significance has changed for the `bac08` variable, but nonetheless, it is statistically significant.

- What do you estimate for coefficients on per se laws? How do the coefficients on per se laws change, if at all? The estimated coefficient in the within model for `per_se_law` is -1.095. Comparing this value to that of the previous model -0.760, we see a further *decrease* in this value which means that this variable is now contributing more towards decrease in total fatality rate.
- What do you estimate for coefficients on primary seat-belt laws? How do the coefficients on primary seatbelt laws change, if at all?

The estimated coefficient in the within model for `primary_seatbelt_law` is -1.154. Comparing this value to that of the previous pooled model 0.222, we see a *decreased* value in the estimate, with a changed sign to negative. This means that the `primary_seatbelt_law` in the within model now contributes in decreasing the total fatality rate. Additionally, in this model, `primary_seatbelt_law` is highly statistically significant whereas in the pooled model, it was not significant.

Which set of estimates do you think is more reliable? Why do you think this?

The within model is reliable is it removes the omitted variable bias in the presence of any unobserved effects, which we would expect to occur when evaluating driving laws in different states. For instance, each state has its unique geography such as elevation or the flatness of terrain that may influence driving conditions. The within model controls for all the time invariant differences in the data by estimating the changes within a specifc state. The pooling model provides us with biased estimates by treating the data as cross-sectional data, ignoring all the characteristics specific to each state.

- What assumptions are needed in each of these models?
- Are these assumptions reasonable in the current context?

For the pooled OLS model, the following are the assumptions: - Linearity - IID - No perfect collinearity - Homoscedastic and uncorrelated errors.

The IID assumption is broken since we have multiple observations for each state, which can introduce clustering and thus non-independent observations.

For the fixed effects (within model), the following are the assumptions: - Linearity - IID - No perfect collinearity - Zero conditional mean (strict exogeneity)

Based on the correlation plots from our EDA section, we have shown that the explanatory variables significant high positive or negative correlations with the total fatality rate. By definition, correlation measures how linearly related two variables are. Therefore, we can describe a linear relationship between the explanatory variables and total fatality rate. Moreover, we satisfy the IID condition since we have demeaned away the unobserved individual effects, leaving our explanatory variables which do not exhibit perfect collinearity.

These assumptions occur before applying linear regression and after we assume a state-level unobserved effect. A formal test for individual effects is performed below.

```r
pFtest(within_model, expanded_model)
```

```
##
##  F test for individual effects
##
## data:  total_fatalities_rate ~ year_of_observation + bac08 + bac10 +  ...
## F = 75.18, df1 = 47, df2 = 1118, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Based on the results of the pooling test above, we obtain a highly significant p-value and thus reject the null hypothesis of no fixed effects. Hence, we should include fixed effects in our model and choose the within model.

(10 points) Consider a Random Effects Model

Instead of estimating a fixed effects model, should you have estimated a random effects model?

- Please state the assumptions of a random effects model, and evaluate whether these assumptions are met in the data.
- If the assumptions are, in fact, met in the data, then estimate a random effects model and interpret the coefficients of this model. Comment on how, if at all, the estimates from this model have changed compared to the fixed effects model.

- If the assumptions are **not** met, then do not estimate the data. But, also comment on what the consequences would be if you were to *inappropriately* estimate a random effects model. Would your coefficient estimates be biased or not? Would your standard error estimates be biased or not? Or, would there be some other problem that might arise?

Aside from including all the fixed effect assumptions, the core assumption of the random effects model is that the residuals are uncorrelated with the explanatory variables and the unobserved effects are independent of the explanatory variables in all the time periods, as opposed to the fixed effects model that assumes a correlation. Overall, looking at the data and based on our first intuition, we believe that Random Effect assumptions are not satisfied as there could be several unobserved effects which can be correlated to our variables. For example, variables related to change in seat belt laws or traffic fatalities can be related to incidents caused by bad road conditions or weather. Another example would be unemployment rate or population percentage in the ages 14 to 24, these variables are prone to change based on laws/policy and economic conditions.

Fitting a random effects model when the assumptions are not met would lead to inefficient and biased estimations of coefficients and its standard errors, making prone to incorrect hypothesis testing. In general, the model would be misleading as it does not represent the true relationships between the variables and we would not be able to draw reliable conclusions from it.

To verify this, we can perform a Hausman test to evaluate whether a random effects model is appropriate compared to a fixed effects model. The null hypothesis of the Hausman test is that the random effects model is the most efficient estimator, meaning it is more useful than a fixed-effects model. We can use this test to determine if a model is appropriate, then can observe the random effects model if so.

```
# Estimate the RE model and conduct the Hausman Test

re.model <- plm(total_fatalities_rate ~ year_of_observation +
                           bac08 +
                           bac10 +
                           per_se_law +
                           primary_seatbelt_law +
                           secondary_seatbelt_law +
                           speed_limit_over_70 +
                           graduated_drivers_license_law +
                           percent_14_24 +
                           unemployment_rate +
                           log(vehicmilespc),
                   data = ptraffic_fatalities,
       index=c("state", "year_of_observation"),
       model = "random")

# Hausman test
phtest(within_model, re.model)
```

```
##
```

```
##  Hausman Test
##
## data:  total_fatalities_rate ~ year_of_observation + bac08 + bac10 +  ...
## chisq = 26.632, df = 34, p-value = 0.812
## alternative hypothesis: one model is inconsistent
```

Interestingly, we fail to reject the null hypothesis that the random effects model is appropriate compared to the fixed effect (within) model. Let us look at the estimates of the random effects model.

```
summary(re.model)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora’s transformation)
##
## Call:
## plm(formula = total_fatalities_rate ~ year_of_observation + bac08 +
##     bac10 + per_se_law + primary_seatbelt_law + secondary_seatbelt_law +
##     speed_limit_over_70 + graduated_drivers_license_law + percent_14_24 +
##     unemployment_rate + log(vehicmilespc), data = ptraffic_fatalities,
##     model = "random", index = c("state", "year_of_observation"))
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Effects:
##                 var std.dev share
## idiosyncratic 3.959   1.990 0.309
## individual    8.857   2.976 0.691
## theta: 0.8675
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -7.69132 -1.10652 -0.15471  0.88435 15.23429
##
## Coefficients:
##                                   Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)                    -102.081080   10.120708 -10.0864 < 2.2e-16 ***
## year_of_observation1981          -1.589496    0.419530  -3.7888 0.0001514 ***
## year_of_observation1982          -3.463847    0.449855  -7.6999 1.362e-14 ***
## year_of_observation1983         -4.192994    0.467206  -8.9746 < 2.2e-16 ***
## year_of_observation1984         -5.025477    0.473355 -10.6167 < 2.2e-16 ***
## year_of_observation1985         -5.666652    0.494315 -11.4636 < 2.2e-16 ***
## year_of_observation1986         -4.797990    0.528892  -9.0718 < 2.2e-16 ***
## year_of_observation1987         -5.670933    0.571113  -9.9296 < 2.2e-16 ***
## year_of_observation1988         -6.292627    0.621354 -10.1273 < 2.2e-16 ***
## year_of_observation1989         -7.812804    0.662311 -11.7963 < 2.2e-16 ***
## year_of_observation1990         -8.068538    0.690258 -11.6892 < 2.2e-16 ***
## year_of_observation1991         -8.869994    0.710157 -12.4902 < 2.2e-16 ***
## year_of_observation1992         -9.916073    0.733779 -13.5137 < 2.2e-16 ***
## year_of_observation1993        -10.234221    0.746684 -13.7062 < 2.2e-16 ***
## year_of_observation1994        -10.607270    0.764954 -13.8666 < 2.2e-16 ***
## year_of_observation1995        -10.448834    0.785125 -13.3085 < 2.2e-16 ***
## year_of_observation1996        -11.014470    0.833889 -13.2086 < 2.2e-16 ***
## year_of_observation1997        -11.160633    0.853011 -13.0838 < 2.2e-16 ***
## year_of_observation1998        -11.914566    0.866757 -13.7461 < 2.2e-16 ***
## year_of_observation1999        -12.054578    0.876785 -13.7486 < 2.2e-16 ***
## year_of_observation2000        -12.576431    0.887466 -14.1712 < 2.2e-16 ***
## year_of_observation2001        -12.311609    0.905340 -13.5989 < 2.2e-16 ***
## year_of_observation2002        -11.742490    0.916436 -12.8132 < 2.2e-16 ***
## year_of_observation2003        -11.854972    0.922677 -12.8485 < 2.2e-16 ***
## year_of_observation2004        -12.316494    0.939334 -13.1119 < 2.2e-16 ***
## bac08                           -0.802392    0.329719  -2.4336 0.0149511 *
## bac10                           -0.628184    0.230189  -2.7290 0.0063528 **
## per_se_law                      -1.054157    0.225039  -4.6843 2.809e-06 ***
## primary_seatbelt_law            -1.087148    0.344669  -3.1542 0.0016095 **
## secondary_seatbelt_law          -0.275819    0.254618  -1.0833 0.2786897
## speed_limit_over_70              0.244258    0.262460   0.9306 0.3520360
## graduated_drivers_license_law   -0.443798    0.282795  -1.5693 0.1165726
## percent_14_24                    0.201807    0.095003   2.1242 0.0336525 *
## unemployment_rate               -0.452146    0.061100  -7.4001 1.361e-13 ***
## log(vehicmilespc)               14.340255    1.127970  12.7133 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Total Sum of Squares:    12775
## Residual Sum of Squares: 4871.5
## R-Squared:      0.61867
## Adj. R-Squared: 0.60754
## Chisq: 1890.09 on 34 DF, p-value: < 2.22e-16
```

For `bac08` the coefficient of the within model is -0.72 whereas for the random effects model it is -0.8. For `bac10` the coefficient of the within model is -0.58 whereas for the random effects model, it is -0.62. For `per_se_law`, the coefficient of the within model is -1.096 whereas for the random effects model, it is -1.054. For `primary_seatbelt_law`, the coefficient of the within model is -1.15 whereas for the random effects model, it is -1.087.

Overall, the coefficients do not vary significantly and have similar effects in both the within model and the random effects model. It also looks like there is no change in statistical significance between the two models. This slight variation could be attributed to the difference in assumptions between the two models.

(10 points) Model Forecasts

The COVID-19 pandemic dramatically changed patterns of driving. Find data (and include this data in your analysis, here) that includes some measure of vehicle miles driven in the US. Your data should at least cover the period from January 2018 to as current as possible. With this data, produce the following statements:

- Comparing monthly miles driven in 2018 to the same months during the pandemic:
  - What month demonstrated the largest decrease in driving? How much, in percentage terms, lower was this driving?
  - What month demonstrated the largest increase in driving? How much, in percentage terms, higher was this driving?

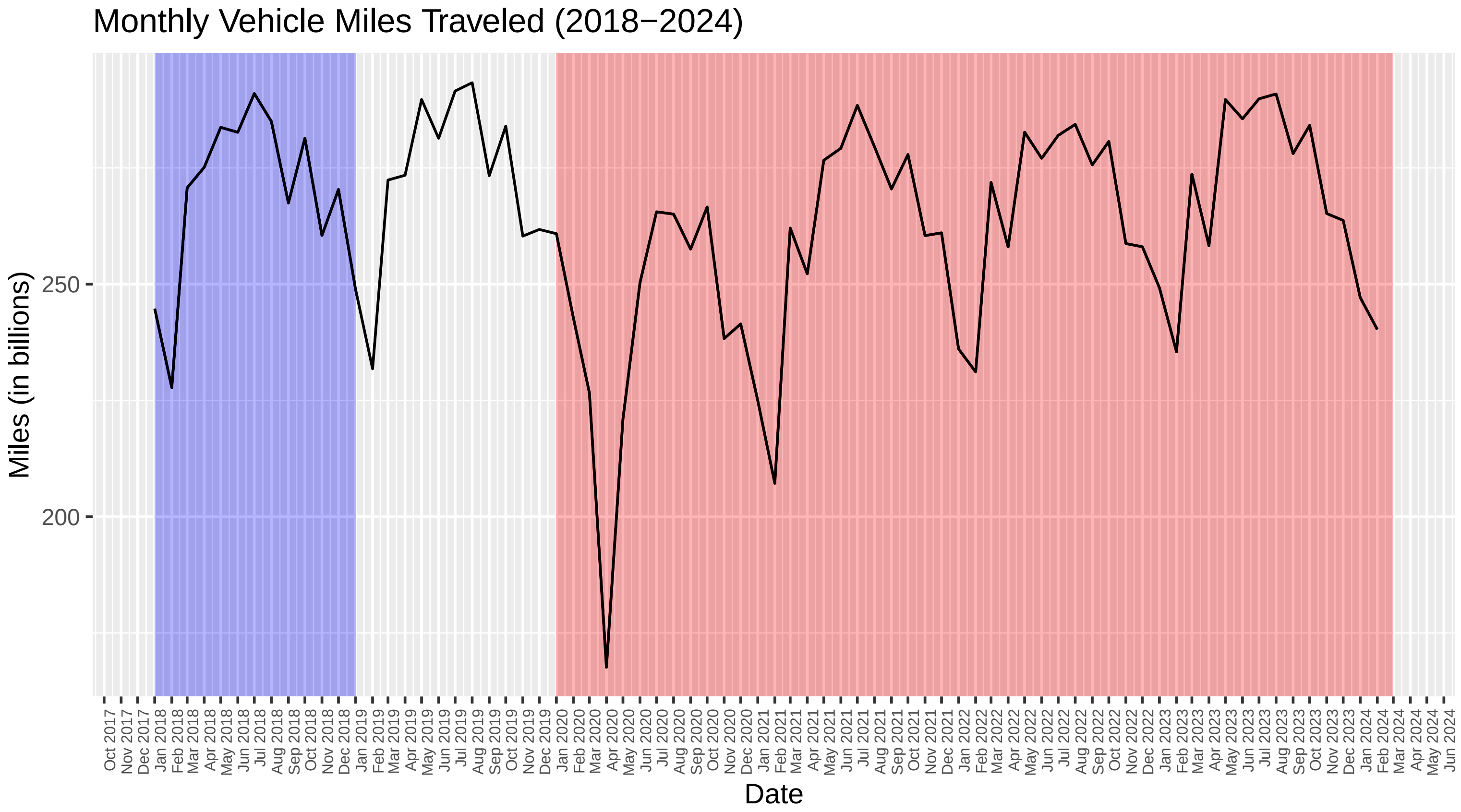
Now, use these changes in driving to make forecasts from your models.

- Suppose that the number of miles driven per capita, increased by as much as the COVID boom. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.
- Suppose that the number of miles driven per capita, decreased by as much as the COVID bust. Using the FE estimates, what would the consequences be on the number of traffic fatalities? Please interpret the estimate.

```r
# Data from Federal Reserve Economic Data (https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA)
# Selected from 2018 and up.
# Load data
downloaded_vehicle_miles <- read.csv('./data/TRFVOLUSM227NFWA.csv')

# Convert to date, make to billions from millions of miles, and rename
downloaded_vehicle_miles <- downloaded_vehicle_miles %>%
  dplyr::rename(miles = TRFVOLUSM227NFWA) %>%
  mutate(DATE = as.Date(DATE), miles = miles / 1000)

head(downloaded_vehicle_miles)
```

```
##         DATE   miles
## 1 2018-01-01 244.736
## 2 2018-02-01 227.759
## 3 2018-03-01 270.705
## 4 2018-04-01 275.127
## 5 2018-05-01 283.713
## 6 2018-06-01 282.648
```

```
# Plot to visualize
ggplot(downloaded_vehicle_miles, aes(x = DATE, y = miles)) +
  geom_line() +
  geom_rect(aes(xmin = as.Date("2018-01-01"), xmax = as.Date("2019-01-01"), ymin = -Inf, ymax = Inf),
            fill = "blue", alpha = 0.002) +
  geom_rect(aes(xmin = as.Date("2020-01-01"), xmax = as.Date("2024-03-01"), ymin = -Inf, ymax = Inf),
            fill = "red", alpha = 0.002) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y") +
  labs(x = "Date", y = "Miles (in billions)", title = "Monthly Vehicle Miles Traveled (2018-2024)")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```



In the figure, the period in blue is the year 2018 and the period in red is the pandemic and beyond. Looks like the month of April 2020 had lowest miles driven while the month of August had the highest miles driven. Let's compare how the pandemic miles differ with those accumulated in 2018.

```
# Calculate percentage increase/decrease

# Extract year and month from the date
downloaded_vehicle_miles$Year <- year(downloaded_vehicle_miles$DATE)
downloaded_vehicle_miles$Month <- month(downloaded_vehicle_miles$DATE)

# Calculate the monthly miles driven for 2018 and during the pandemic
miles_2018 <- downloaded_vehicle_miles %>%
  filter(Year == 2018) %>%
  select(Year, Month, miles)

miles_pandemic <- downloaded_vehicle_miles %>%
  filter(Year >= 2020) %>%
  select(Year, Month, miles)

# Join the two dataframes and calculate the percentage change
percentage.change <- full_join(miles_2018, miles_pandemic, by = c("Month"), suffix = c("_2018", "_pandemic")) %>%
  mutate(percent_change = (miles_pandemic - miles_2018) / miles_2018 * 100) %>%
  mutate(change = miles_pandemic - miles_2018)

# Find the month with the largest decrease and increase in driving
largest_decrease <- percentage.change[which.min(percentage.change$change), ]
largest_increase <- percentage.change[which.max(percentage.change$change), ]

print(paste("The largest decrease in driving was in month", largest_decrease$Month, "of year", largest_decrease$Year_pandemic,
            "with a decrease of", largest_decrease$change, "units, or", round(largest_decrease$percent_change, 2), "%"))
```

```
## [1] "The largest decrease in driving was in month 4 of year 2020 with a decrease of -107.51 units, or -39.08 %"
```

```
print(paste("The largest increase in driving was in month", largest_increase$Month, "of year", largest_increase$Year_pandemic,
            "with an increase of", largest_increase$change, "units, or", round(largest_increase$percent_change, 2), "%"))
```

```
## [1] "The largest increase in driving was in month 1 of year 2020 with an increase of 16.111 units, or 6.58 %"
```

```
# Forecast
# Get the coefficient from the within model
vehicle_miles_pc_coef <- coef(within_model)["log(vehicmilespc)"]

# bust
print(paste("Bust: ", (largest_decrease$change * vehicle_miles_pc_coef ) / 100))
```

```
## [1] "Bust:  -13.0418933298609"
```

```r
# boom
print(paste("Boom: ", (largest_increase$change * vehicle_miles_pc_coef ) / 100))
```

```
## [1] "Boom:  1.95440371535103"
```

During the COVID bust, we can see a decrease in total fatality rate by approximately 13.04 for every percent decrease in vehicle miles per capita, keeping all else constant. During the COVID boom, we can see an increase in total fatality rate by approximately 1.95 for every percent increase in vehicle miles per capita, keeping all else constant.

(5 points) Evaluate Error

If there were serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors? Is there any serial correlation or heteroskedasticity?

In case of any serial correlation or heteroskedasticity in the errors of the model, the standard errors would be incorrect. The coefficients would be unbiased and consistent but they would no longer have the lowest variance. To check this, we can conduct a few tests.

```r
# Breusch-Godfrey test for serial correlation
pbgtest(within_model, order=2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  total_fatalities_rate ~ year_of_observation + bac08 + bac10 +  ...
## chisq = 278.23, df = 2, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```r
# Breusch Pagan Test for heteroscedasticity
pcdtest(within_model, test="lm")
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  total_fatalities_rate ~ year_of_observation + bac08 + bac10 +     per_se_law + primary_seatbelt_law + secondary_seatbelt_law +
## chisq = 3132.4, df = 1128, p-value < 2.2e-16
## alternative hypothesis: cross-sectional dependence
```

We conduct the Breusch-Godfrey test for several lags (only order 2 shown here), all of which reject the null hypothesis of no serial correlation in the within model. This means that we need to adjust for this autocorrelation.

To test for heteroskedasticity, we conduct the Breusch Pagan test, we reject the null hypothesis of homoskedasticity in the within model. Which mean that we need to adjust our standard errors and cluster them as appropriate to control for this heteroskedasticity.

Since we reject the null hypothesis in both the above tests, it is recommended to fit the Arellano standard errors as it adjusts for serial correlation as well as cluster our standard errors.

References:

Freeman, Donald G. 2007. "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws." *Contemporary Economic Policy* 25 (3): 293–308.