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

Shishir Agarwal, Vasanth Ramani, Vineetha Nalini

U.S. traffic fatalities: 1980-2004

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

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

```
library(foreign)
library(gplots)
library(ggplot2)
library(dplyr)
library(corrplot)
library(lattice)
library(plm)
library(viridis)
library(tsibble)
library(forecast)
library(gridExtra)
```

Exercises:

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

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

#check rows and columns
dim(driving.df)
```

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

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

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

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

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

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

Panel Index
year: 1980 through 2004
state: numeric id of 48 continental states, ordered alphabetically, ranging from 1 to 51.

Dependent Variable
totfatrte: total fatalities per 100,000 population by year by state. Values range from 6.2 to 53.32

Speed Limit Variables
sl55: 1 if speed limit == 55 for the whole year. If the law was in effect only during part of the year, it is set to fractions of 12. This applies for all indicator variables.
sl65: 1 if speed limit == 65
sl70: 1 if speed limit == 70
sl75: 1 if speed limit == 75
slnone: 1 if no speed limit
sl70plus: sl70 + sl75 + slnone

Drinking Laws
minage: minimum drinking age, ranges from 18 years to 21 years.
zerotol: 1 if zero tolerance law was in effect, and 0 if not. If the law was in effect only during part of the year, it is set to fractions of 12.
bac10: 1 if blood alcohol limit .10 in effect, and 0 if not. Fractions used to denote partial years, as above.
bac08: 1 if blood alcohol limit .08 in effect, and 0 if not. Fractions used to denote partial years, as above.
per se: 1 if administrative license revocation (per se law) in effect, and 0 if not. Fractions used to denote partial years, as above.

Seatbelt Laws
sbprim: 1 if primary seatbelt law was in effect, 0 otherswise. There are no fractions in this variable.
sbsecon: 1 if secondary seatbelt law was in effect, 0 otherswise. There are no fractions in this variable.
seatbelt: 0 if none, =1 if primary, =2 if secondary. There are no fractions in this variable.

Age iimit Laws
gdl: 1 if graduated drivers license law was in effect, and 0 if not. Fractions used to denote partial years, similar to speed limit.

Demographic variables
statepop: state population by year by state. Values range from 453,401 to 35,894,000
vehicmiles: vehicle miles traveled, billions. Values range from 3.7027 to 329.6
unem: unemployment rate, percent. Values range from 3.2 to 18
perc14_24: percent population aged 14 through 24. Values range from 11.7 to 20.3
vehiclemilespc: vehicle miles driven per capita. Values range from 4,372 to 18,390

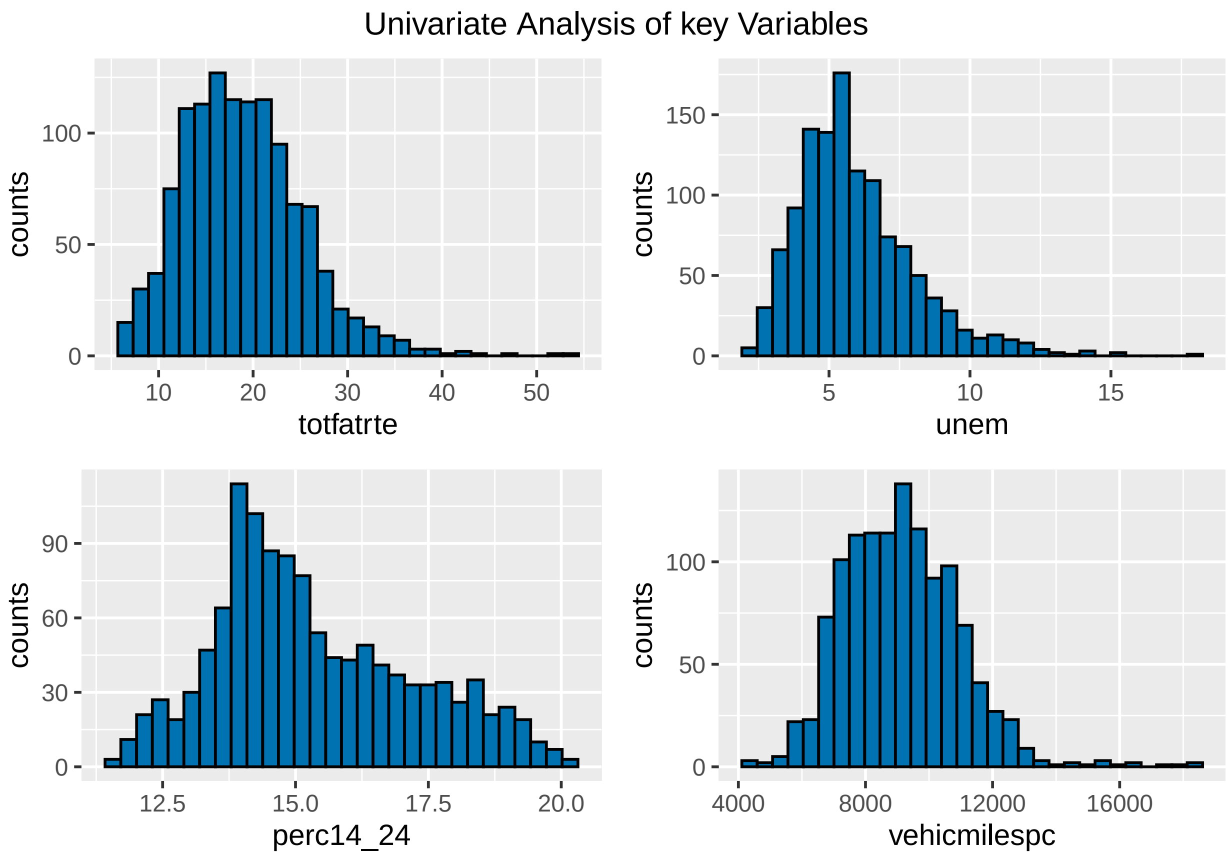
Year Dummy
Dummy variables $d80$ - $d04$ indicating years

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

```
#get state name
us.states = read.csv("usstates.csv", header = TRUE, sep = ",", dec = ".")
data.with.name <- merge(data, us.states, by=c("state","state"))
```

To start EDA, we perform univariate analyses of important variables fatality rate, unemployment, % of younger population, and vehiclemilespc to examine the distribution.

```
totfatrte.hist <- ggplot(driving.df, aes(x = totfatrte)) +   geom_histogram(bins = 30, fill="#0
unem.hist <- ggplot(driving.df, aes(x = unem)) +  geom_histogram(bins = 30, fill="#0072B2", col
perc14_24.hist <- ggplot(driving.df, aes(x = perc14_24)) +  geom_histogram(bins = 30, fill="#00
vehicmilespc.hist <- ggplot(driving.df, aes(x = vehicmilespc)) +  geom_histogram(bins = 30, fil
grid.arrange( totfatrte.hist, unem.hist, perc14_24.hist, vehicmilespc.hist, ncol = 2, nrow = 2
```

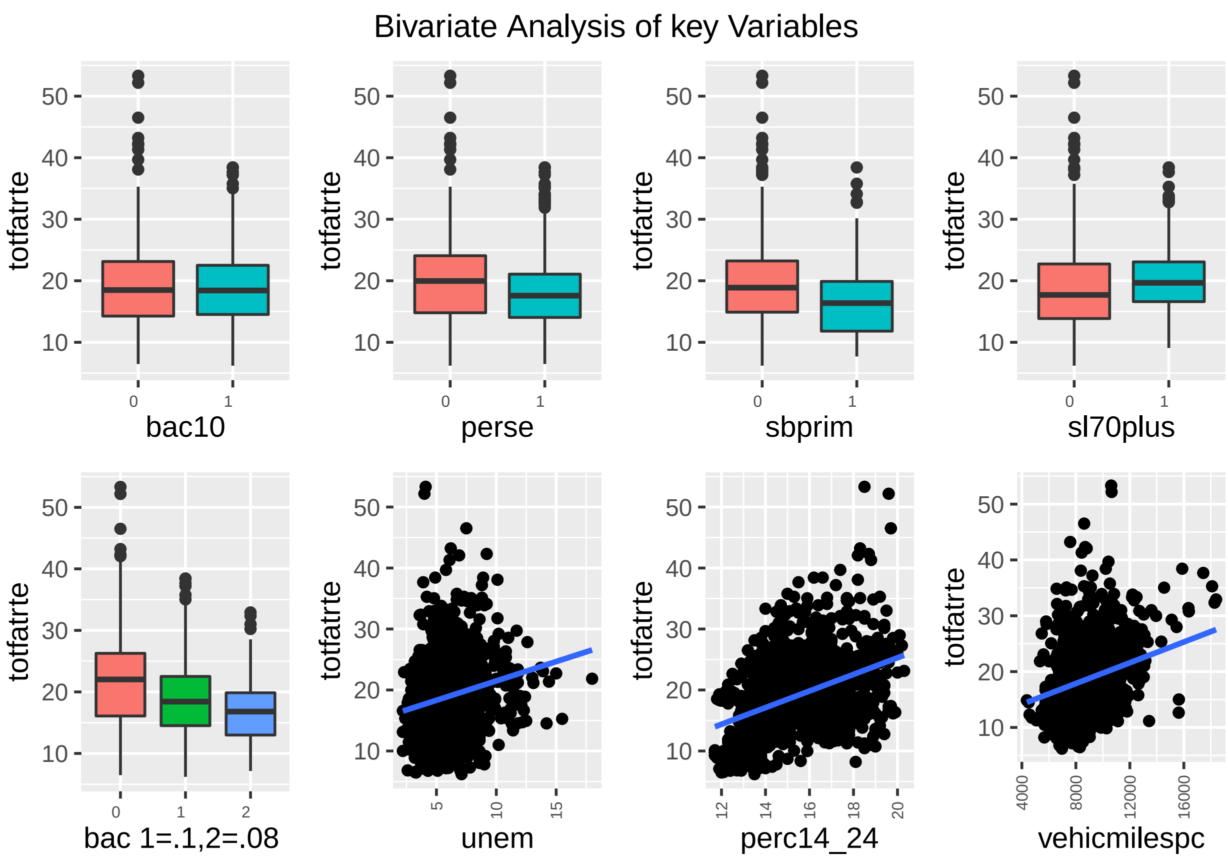


The distribution looks approximately normal with some tail for $totfatrte$, $unem$, and $vehiclemilespc$. It looks normal with higher slope at the top and lower slope at the bottom for $perc14 - 24$.

Next, we examine the bivariate relationship between some of the important explanatory variables and fatality rate.

```
totfatrte.unem.scatter <- ggplot(driving.df, aes(unem, totfatrte)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) + theme(axis.text.x = element_text(an
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
totfatrte.perc14_24.scatter <- ggplot(driving.df, aes(perc14_24, totfatrte)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) + theme(axis.text.x = element_text(an
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
totfatrte.vehicmilespc.scatter <- ggplot(driving.df, aes(vehicmilespc, totfatrte)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) + theme(axis.text.x = element_text(an
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")

driving.df.mutate <- driving.df %>% mutate(bac08 = ifelse(bac08 > 0.5,1,0)) %>%
  mutate(bac10 = ifelse(bac10 > 0.5,1,0)) %>% mutate(perse = ifelse(perse > 0.5,1,0)) %>%
  mutate(sl70plus = ifelse(sl70plus > 0.5,1,0)) %>% mutate(gdl = ifelse(gdl > 0.5,1,0)) %>% mut

totfatrte.bac10.box <- ggplot(driving.df.mutate, aes(x = factor(bac10), y = totfatrte)) +
```

```
  geom_boxplot(aes(fill = factor(bac10))) + xlab("bac10") + theme(axis.text.x = element_text(ar
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
totfatrte.perse.box <- ggplot(driving.df.mutate, aes(x = factor(perse), y = totfatrte)) +
  geom_boxplot(aes(fill = factor(perse))) + xlab("perse") + theme(axis.text.x = element_text(ar
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
totfatrte.sbprim.box <- ggplot(driving.df.mutate, aes(x = factor(sbprim), y = totfatrte)) +
  geom_boxplot(aes(fill = factor(sbprim)))+ xlab("sbprim") + theme(axis.text.x = element_text(a
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
totfatrte.sl70plus.box <- ggplot(driving.df.mutate, aes(x = factor(sl70plus), y = totfatrte)) +
  geom_boxplot(aes(fill = factor(sl70plus)))+ xlab("sl70plus") + theme(axis.text.x = element_te
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
totfatrte.bacall.box <- ggplot(driving.df.mutate, aes(x = factor(bacall), y = totfatrte)) +
  geom_boxplot(aes(fill = factor(bacall)))+ xlab("bac 1=.1,2=.08 ") + theme(axis.text.x = eleme
  theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")
grid.arrange( totfatrte.bac10.box, totfatrte.perse.box, totfatrte.sbprim.box, totfatrte.sl70plu
```



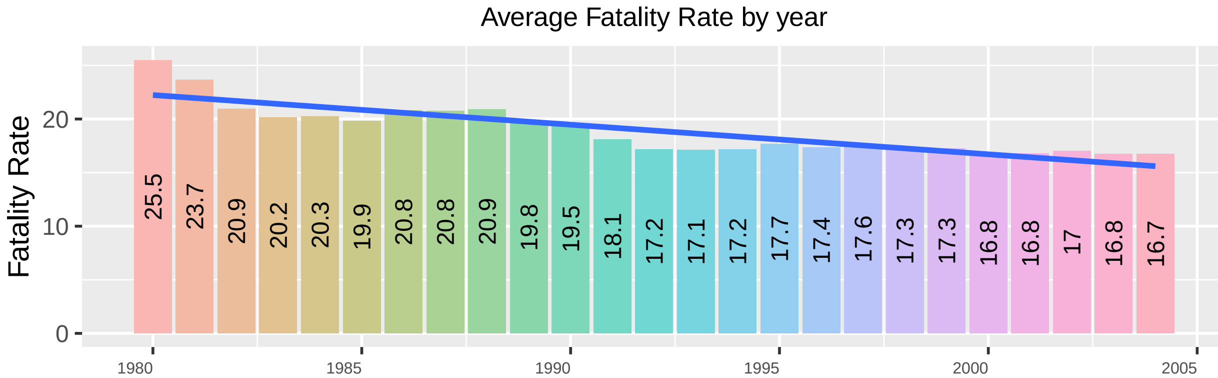
Note that fractions are rounded for representing box plots. We see that blood alcohol limit 10 has a muted effect while per se and primary seatbelt laws have reducing effect on the fatality rate. In addition, fatality rate with bac08 is slightly lower than bac10. Also note the higher fatality rate on the states with speed limit 70 and above or none. We are also exploring the relationship between fatality rate and demographic variables *unem*, $perc14_24$ and *vehicmilespc* to understand the impact.

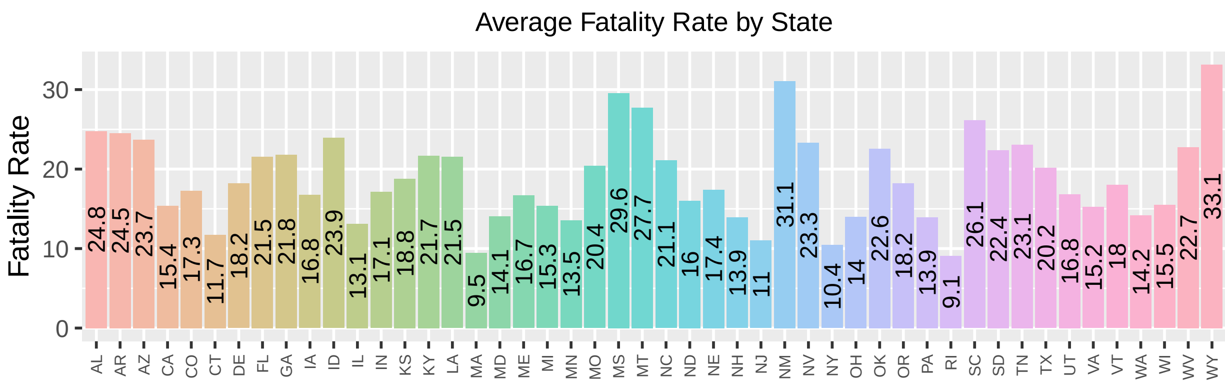
Then, to examine both the average fatality pattern over the years, and individual fixed effect of US

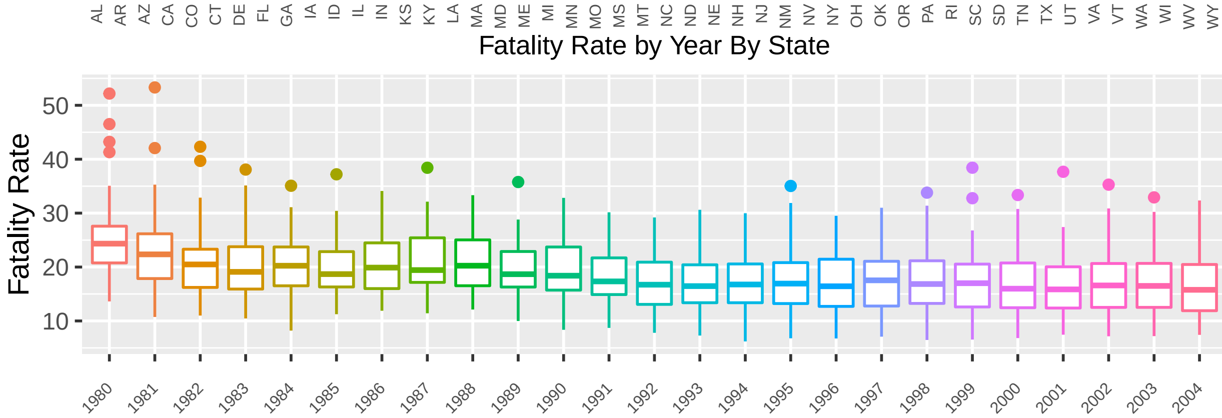
States across time, we'll analyze the aggregate of traffic laws in US across time and across states.

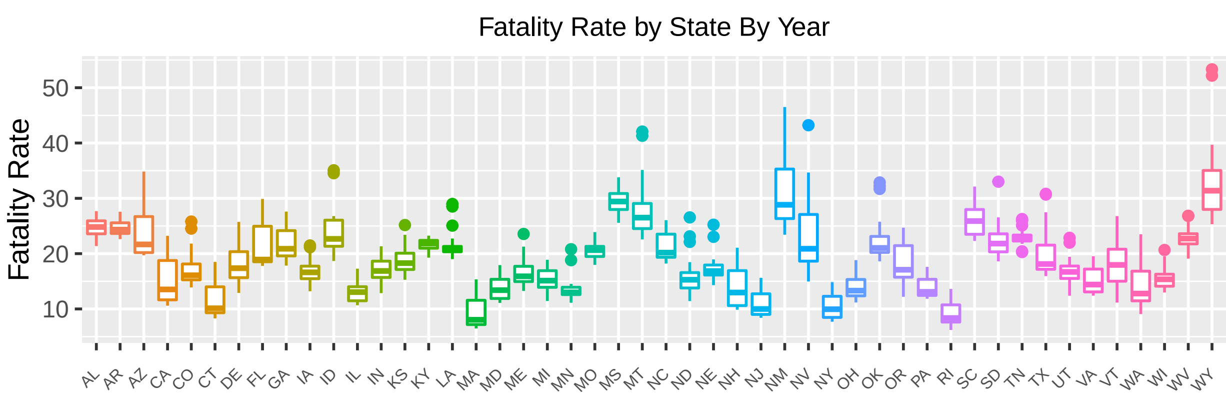
Below we analyze the fatality rate change by year and overall change by state .

```
#fatality change by year
traffic.yearly.aggr <- data %>%   group_by(year) %>%  summarise_at(vars(totfatrte, nghtfatrte,w

#fatality change by state
traffic.state.perc.aggr <- data.with.name %>% group_by(shortname) %>% summarise_at(vars(totfat

#common functions
overall_plot = function(data,condvar,title,angle) {
  g <- ggplot(data, aes(condvar, totfatrte)) +
  geom_bar(aes(fill = factor(condvar)), position = "dodge", stat="identity") + ggtitle(title) +
   geom_text(data = data, aes(x = condvar, y = totfatrte, label = round(totfatrte,1)), size = 3
}
conditional_plot = function(data, plotvar, condvar, title) {
  g <- ggplot(data, aes(as.factor(condvar), plotvar, color = as.factor(condvar)))
  g + geom_boxplot() + ggtitle(title) + theme(axis.text.x = element_text(angle = 45, size = 6,
}

year.plot.1 <- overall_plot(traffic.yearly.aggr,traffic.yearly.aggr$year,"Average Fatality Rate
state.plot.1 <- overall_plot(traffic.state.perc.aggr,traffic.state.perc.aggr$shortname,"Average

cplot.1 <- conditional_plot(data.with.name, data.with.name$totfatrte, data.with.name$year, "Fa
cplot.2 <- conditional_plot(data.with.name, data.with.name$totfatrte, data.with.name$shortname

grid.arrange(year.plot.1, state.plot.1, nrow = 2,  ncol = 1);grid.arrange(cplot.1, cplot.2, nr
```







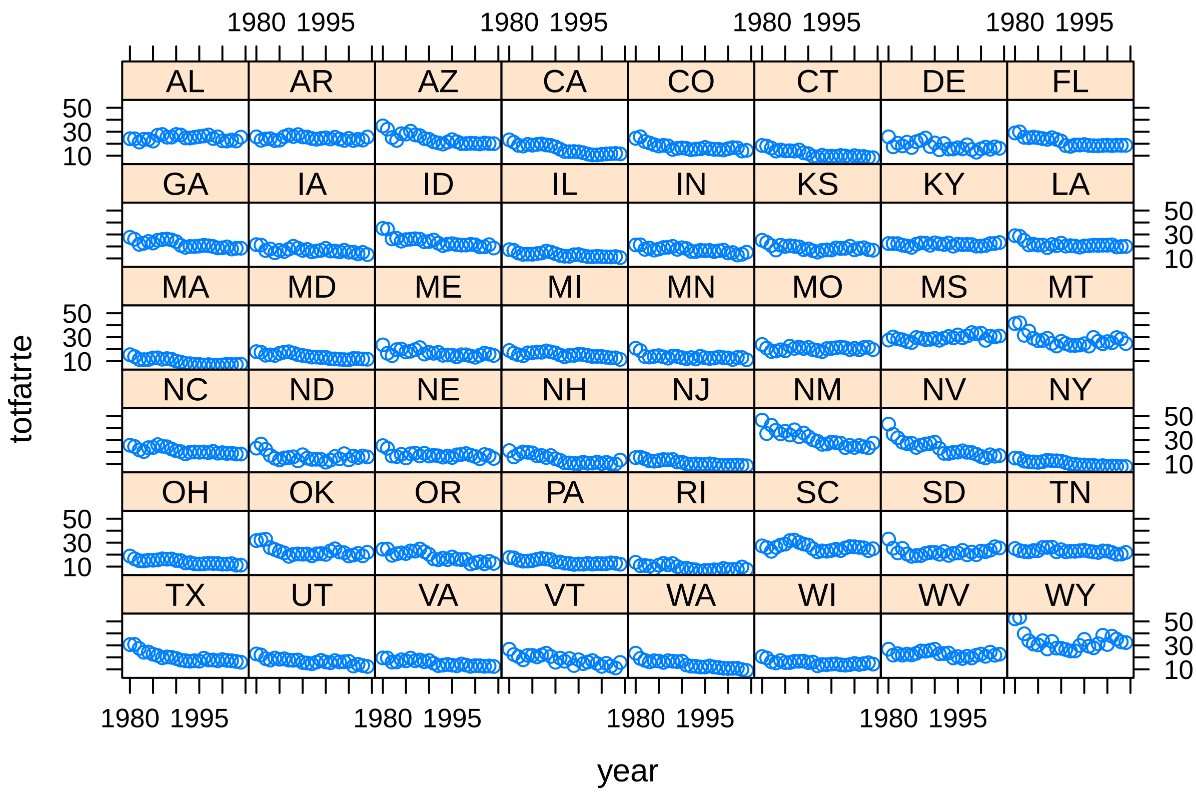


We can see that the fatality rate is largely decreasing from 1980 to 2004. The overall fatality rates range from ~9 to ~34 between states. *Wyoming*, *New Mexico*, *Mississippi*, *Montana*, and *South*

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

The boxplots graph shows there is heterogeneity across states, but very little heterogenetiy across years. Next we analyze how the fatality rates varied over the years, in individual states.

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

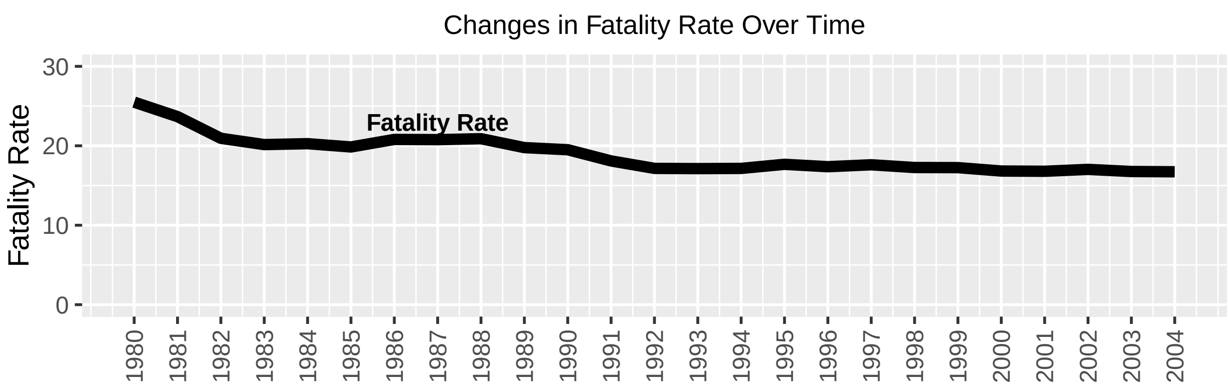


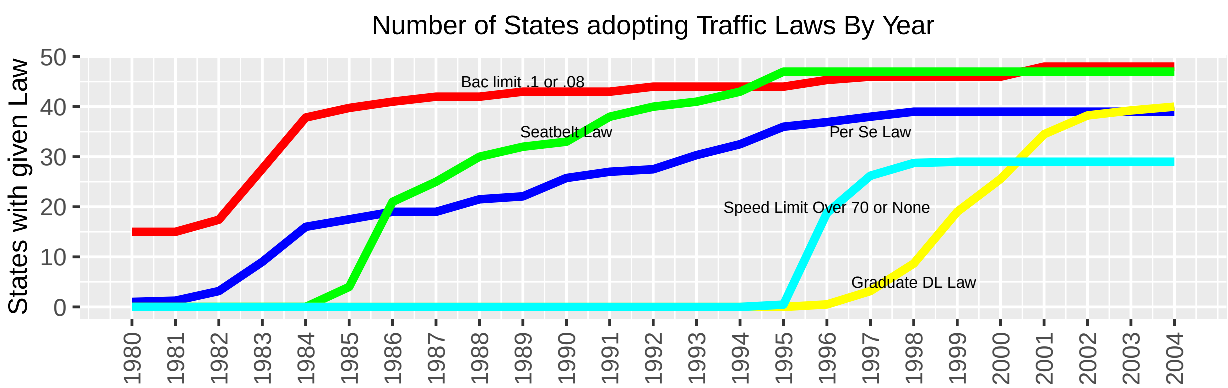
The above xyplot confirms that most of the States shows an overall decrease in the traffic fatality rate, except states like *Mississippi* . We can see that New Mexico(NM) and Wyoming(WY) has high variance in the data, with NM consistently reducing the traffic fatality rate over years. However, WY reduced the fatality rate from 80's to mid 90's and had a gradual increase after. One interesting point is that the traffic fatality rate is not dependent on the state area or population - the top 2 states with size and population, *Texas* and *California*, are not among the top states in traffic fatality rate.

Below, we explore how the traffic laws over the years across states, and whether they show a correlation with fatality rate. We first plot the fatality rate over years, and then plot the count of states that adopt the traffic laws, grouped by year and specific law. We hypothesize that the fatality rate is influenced the most by drinking and overspeeding and proceed to examine the applicable laws.

```
# summarize the number of states adopting specific traffic laws
data <- data %>%  mutate(sball = case_when(sbprim + sbsecon >= 0.5  ~ 1,TRUE ~ 0 ))
```

```
bac.df <- data %>% group_by(year) %>%
summarise(bac10 = sum(bac10), bac08 = sum(bac08), bac.all = sum(bac08 + bac10), perse = sum(pe

bac.plot.f <- ggplot(bac.df, aes(x = year)) +
  geom_line(aes(y=totfatrte, color='totfatrte'), size = 2, group = 1) + ylim(0,30)+
  scale_x_continuous(breaks = seq(min(bac.df$year), max(bac.df$year), 1)) + theme(axis.text.x =
  annotate("text", x = 1987, y = 23, label = "Fatality Rate", size = 3, fontface = "bold") +
   theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position = "none")

bac.plot <- ggplot(bac.df, aes(x = year)) +
geom_line(aes(y = bac.all, color='bac.all'), size = 1.5, group = 1) +   geom_line(aes(y=perse,
  geom_line(aes(y=sball, color='sball'), size = 1.5, group = 1) + geom_line(aes(y=gdl, color='
      values = c( bac.all="red", perse="blue", sball="green", gdl = "yellow", sl70plus = "cyan
x = "",size = 2) +
  annotate("text", x = 1998, y = 5, label = "Graduate DL Law", size = 2) +  annotate("text", x
  annotate("text", x = 1990, y = 35, label = "Seatbelt Law", size = 2) +  annotate("text", x =
  annotate("text", x = 1989, y = 45, label = "Bac limit .1 or .08", size = 2) + theme(plot.titl

grid.arrange(bac.plot.f, bac.plot, nrow = 2,  ncol = 1)
```





Over the years, more states are adopting stricter alcohol limits. In 2004, over 45 states have a bac limit of 0.08 or 0.1, compared to ~10 in 1980. Similarly, ~40 states have adopted the Per Se law and Graduate DL law in 2004 compared to 0 states in 1980. We see a similar trend in seatbelt adoption as well. This is consistent with the decrease in fatality rates over time that observed before.

Regarding speed limit, states had lower speed limit in 1980 - however, the speed limits were more relaxed in the later years as can be seen by the increase in the number of states with speed limit 70 or above, as seen in the above graph.

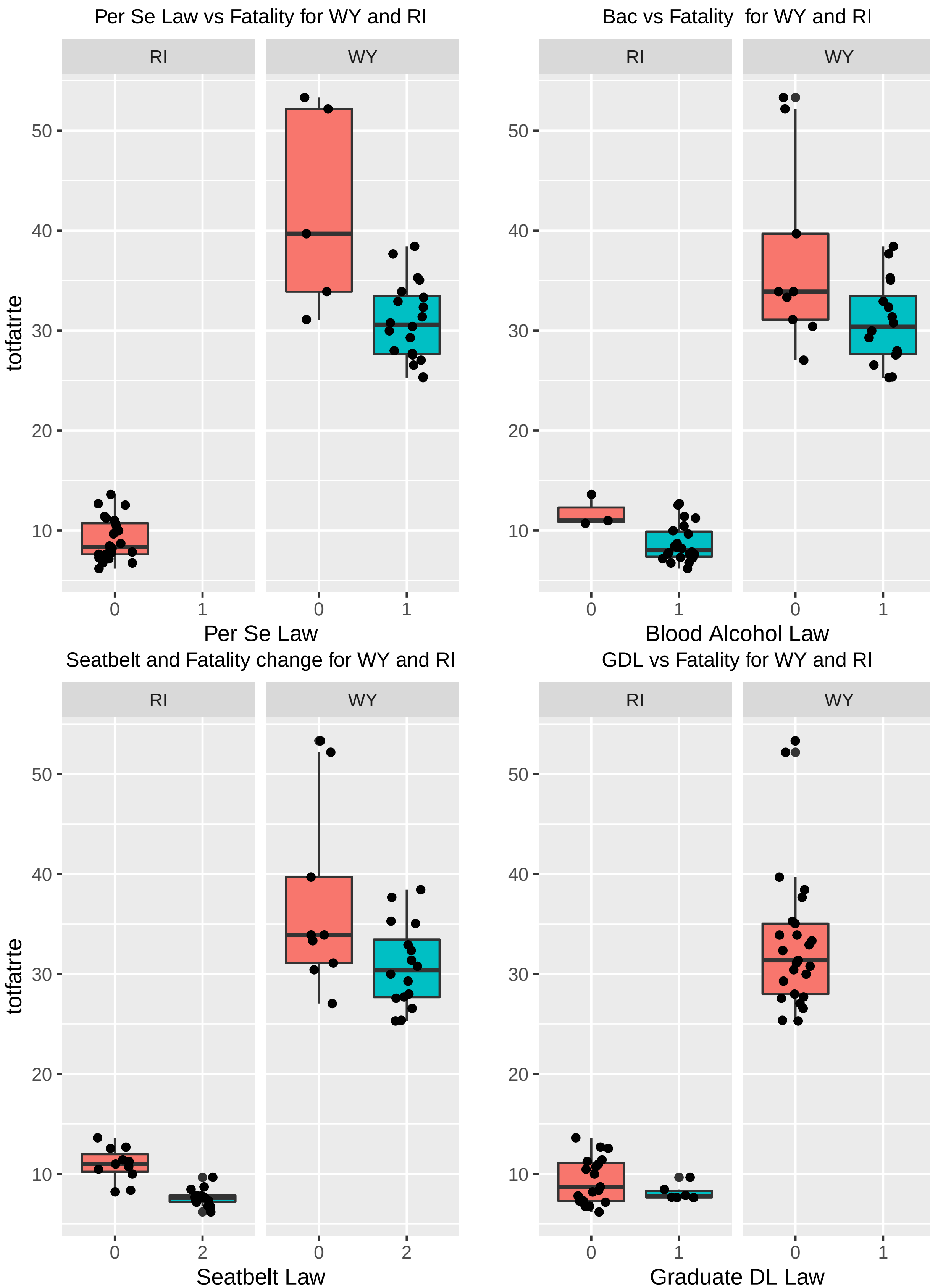
Lets proceed to examine the individual state behavior in the panel. First, we'll examine how the traffic fatality rates changed over years for the first 3 States from the top and bottom of the fatality rate scale.

```
#Top 3 and Bottom 3 states in overall fatality scale
traffic.state.aggr <- data.with.name %>%   group_by(shortname) %>%   summarise_at(vars(totfat,
top.3.fatalities <- traffic.state.perc.aggr %>%  filter(rank(desc(totfatrte))<=3) %>% arrange(d
bottom.3.fatalities <- traffic.state.perc.aggr %>% filter(rank((totfatrte))<=3) %>% arrange((to
data.top.filtered <- data.with.name %>% filter(shortname %in% c("WY"))
data.bottom.filtered <- data.with.name %>% filter(shortname %in% c("RI"))
data.merged <- union(data.top.filtered,data.bottom.filtered)

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

```
##   top3states totfatrte bottom3states totfatrte
## 1         WY   33.1408            RI    9.0900
## 2         NM   31.0608            MA    9.4512
## 3         MS   29.5548            NY   10.4380
```

The top 3 are *Wyoming*, *New Mexico* and *Mississippi*. The bottom 3 are *Rhode Island*, *New York* and *Massachussets*. We see that the fatality rate in Wyoming is nearly 4 times as high as in Rhode Island in 2004. Below, we see the how the fatality rate varies for seat belt, bac each state across years.

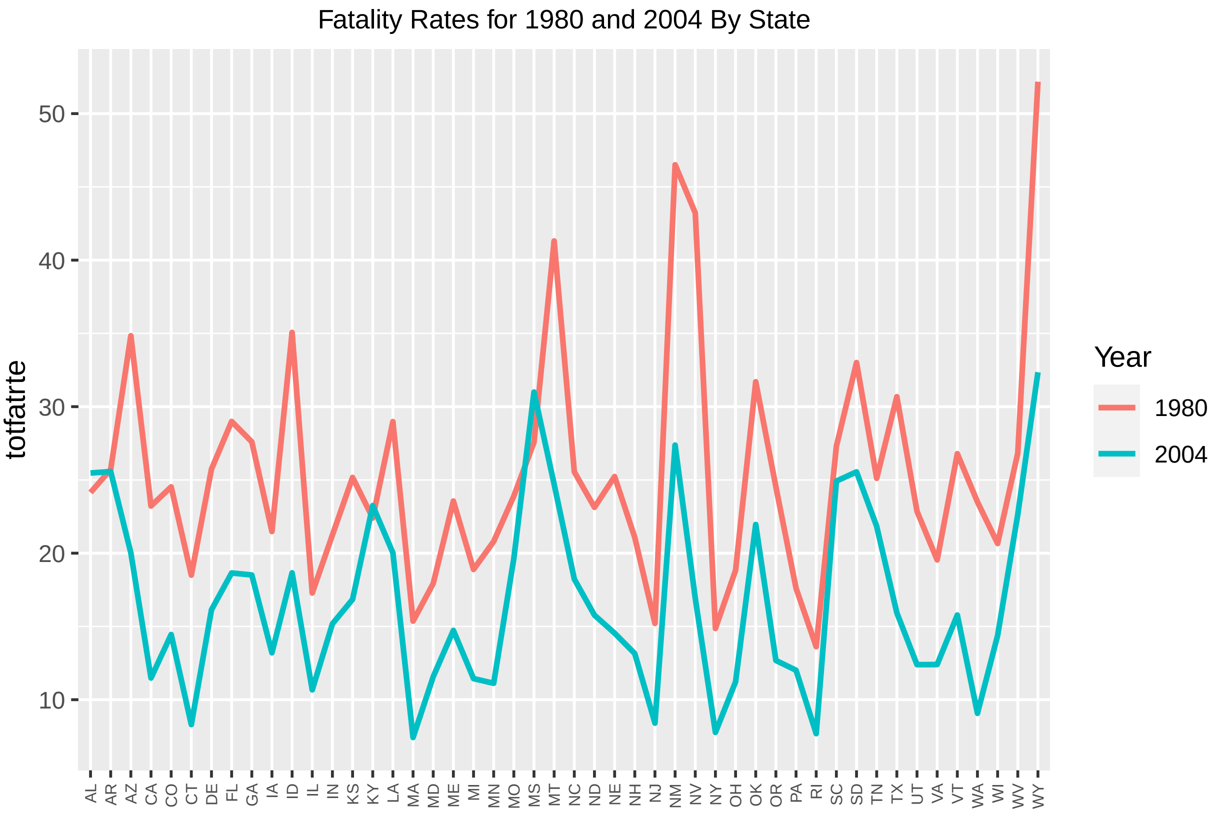
```
df.transformed <- data.merged %>%   mutate( perse = case_when(perse >= 0.5  ~ 1,TRUE ~ 0),baca

g.1 <- ggplot(df.transformed, aes(as.factor(perse), totfatrte)) + geom_boxplot(aes(fill = facto

g.2 <- ggplot(df.transformed, aes(as.factor(bacat), totfatrte)) + geom_boxplot(aes(fill = facto

g.3 <- ggplot(df.transformed, aes(as.factor(seatbelt), totfatrte)) + geom_boxplot(aes(fill = fa

g.4 <- ggplot(df.transformed, aes(as.factor(gdl), totfatrte)) + geom_boxplot(aes(fill = factor

grid.arrange(g.1, g.2, nrow = 1,  ncol = 2);grid.arrange(g.3, g.4, nrow = 1,  ncol = 2)
```


Per Se Law vs Fatality for WY and RI
Bac vs Fatality for WY and RI
Seatbelt and Fatality change for WY and RI
GDL vs Fatality for WY and RI
RI
WY
totfatrte
10
20
30
40
50
0
1
2
Per Se Law
Blood Alcohol Law
Seatbelt Law
Graduate DL Law

In the above graphs, seatbelt=1 indicates the the states followed the primary or secondary seatbelt laws for more than half of the year. Similar logic is applied to blood alcohol laws, Per Se and Gdl laws . *We can see that WY has lower fatality rate in the years they adopted the Per Se, Bac, and Seatbelt laws.* They have not adopted the Gdl law as of 2004. Rhode Island shows a similar pattern, eventhough the fatality rate was already low. WY also has higher variance over the years than RI regardless of the law adoption.

Now we will look at how fatality rate differs between 1980 and 2004 for each state.

```
df.80.04 <- data.with.name %>% filter(year %in% c('1980','2004')) %>% dplyr::select(year,shortn

ggplot(df.80.04, aes(shortname, totfatrte, group = year, colour = as.factor(year))) +
  geom_line(aes(y=totfatrte), size = 1) + ggtitle("Fatality Rates for 1980 and 2004 By State")
```



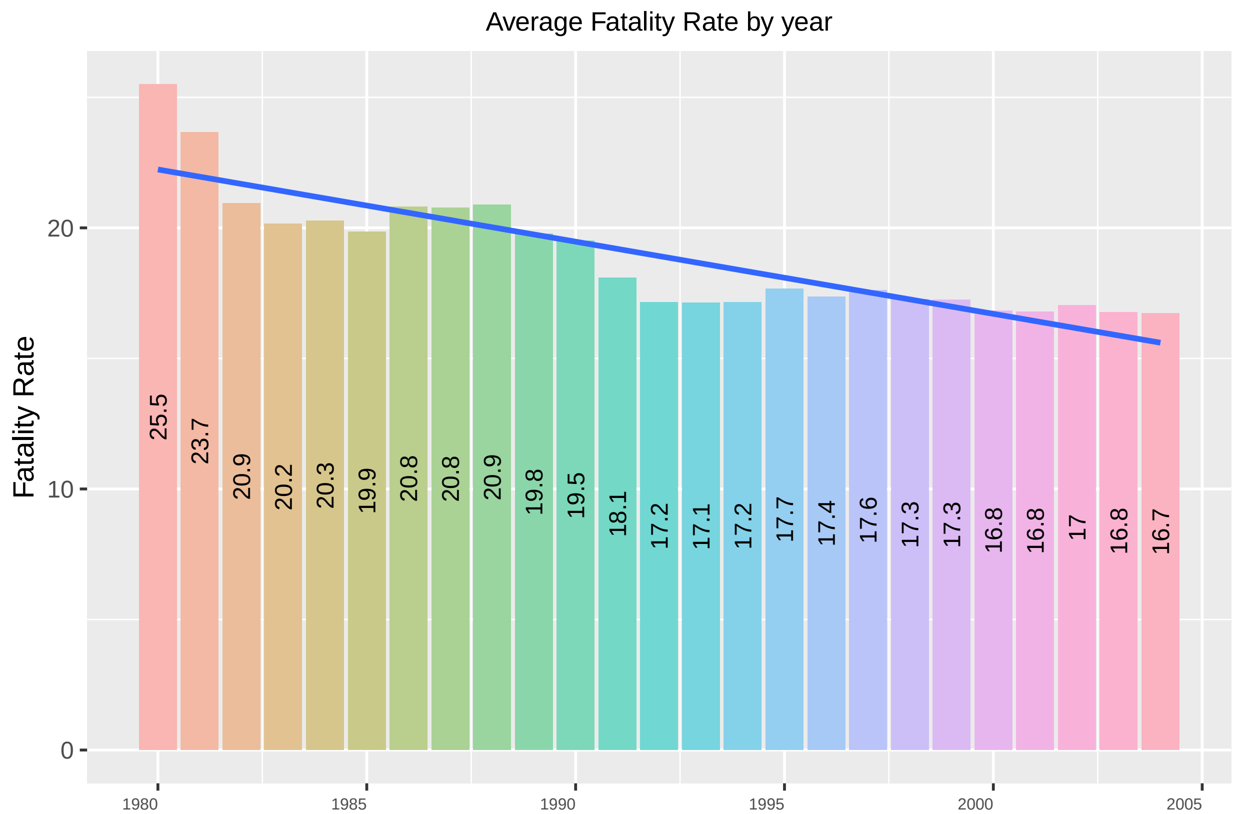
For most of the states, fatality rate in 2004 is less than 1980 barring exceptions like *Arkansas(AR)* and *Mississippi(MS)*. States show more variance in 1980 than in 2004.

Thus both the overall and state level EDA indicates that there is a reduction in fatality rate that is consistent with traffic laws adoption. However the actual impact of the traffic laws adoption, and the demographic variables on fatality rate can be assessed only through the detailed regression analysis below.

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

The dependent variable *totfatrte* is defined as the total fatalities per 100,000 population, grouped by the index variables year and state. The average of this *totfatrte* variable per year is computed for EDA in variable $traffic.yearly.aggr$ and plotted using $year.plot.1$.

```
year.plot.1
```



Let's estimate the linear regression for the dummy variables from 1981 to 2004 below.

```
lm.fit1 <- lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +
                          d90 + d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 +
                          d00 + d01 + d02 + d03 + d04, data=driving.df)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04, data = driving.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9302  -4.3468  -0.7305   3.7488  29.6498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.4946     0.8671  29.401  < 2e-16 ***
```

```
## d81          -1.8244     1.2263  -1.488 0.137094
## d82          -4.5521     1.2263  -3.712 0.000215 ***
## d83          -5.3417     1.2263  -4.356 1.44e-05 ***
## d84          -5.2271     1.2263  -4.263 2.18e-05 ***
## d85          -5.6431     1.2263  -4.602 4.64e-06 ***
## d86          -4.6942     1.2263  -3.828 0.000136 ***
## d87          -4.7198     1.2263  -3.849 0.000125 ***
## d88          -4.6029     1.2263  -3.754 0.000183 ***
## d89          -5.7223     1.2263  -4.666 3.42e-06 ***
## d90          -5.9894     1.2263  -4.884 1.18e-06 ***
## d91          -7.3998     1.2263  -6.034 2.14e-09 ***
## d92          -8.3367     1.2263  -6.798 1.68e-11 ***
## d93          -8.3669     1.2263  -6.823 1.43e-11 ***
## d94          -8.3394     1.2263  -6.800 1.66e-11 ***
## d95          -7.8260     1.2263  -6.382 2.51e-10 ***
## d96          -8.1252     1.2263  -6.626 5.25e-11 ***
## d97          -7.8840     1.2263  -6.429 1.86e-10 ***
## d98          -8.2292     1.2263  -6.711 3.01e-11 ***
## d99          -8.2442     1.2263  -6.723 2.77e-11 ***
## d00          -8.6690     1.2263  -7.069 2.67e-12 ***
## d01          -8.7019     1.2263  -7.096 2.21e-12 ***
## d02          -8.4650     1.2263  -6.903 8.32e-12 ***
## d03          -8.7310     1.2263  -7.120 1.88e-12 ***
## d04          -8.7656     1.2263  -7.148 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.008 on 1175 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1098
## F-statistic: 7.164 on 24 and 1175 DF,  p-value: < 2.2e-16
```

This is a pooled OLS model that explains the impact of time on the total fatality rate. The year dummy variables are highly statistically significant except for 1981. We see total fatality rate trending down with time and it proves that the *driving became safer over this period.* We see that compared to 1980, for each year we have lower fatality rate. The coefficients also indicate that compared to fatality rate between 83 and 85, we see a higher fatality rate for years 86 to 88. The adjusted R^2 0.1098 indicates the model is not a good fit. Diagnostic plots (not shown here due to space constraints) indicate that residuals are not normally distributed. Detailed explanation of why the pooled OLS model is not a good fit for this panel dataset is given in question 4.

3. (15%) Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables.* Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the coefficients on *bac8* and *bac10*. Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)

Is Transformation needed?
First, we looked at the predicted variable *totfatrte* for possible transformations. EDA showed that the variable is already weighted by state population and ranged from ~6 to ~50, and was normally distributed. Hence we did not apply any transformations on the variable.

The variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl* define the year in which the specific law became effective. It ranges from 0-1 with decimals as the law could change in the middle of the year. One possible transformation is to round the fractions into 0 or 1 - but we choose to not round them up, as we will lose precision if the law was adopted in the middle of the year.

Variables *perc14_24, unem, vehicmilespc* are already normalized and so no transformation is necessary. Note that *vehicmilespc* is a candidate for log transformation as there is a range from 4k to 18k miles per capita and the relationship between this and fatality rate is not linear as seen from EDA - but because the variable is already a rate, and we lose interpretability with log transformation, we retain the original format.

Definition of bac08 and bac10 variables and their interpretation
bac10 is defined as the blood alcohol limit of .10 and *bac08* is defined as the blood alcohol limit of .08. Both the variables *bac08* (-2.498) and *bac10* (-1.418) have negative coefficients, with *bac08* having larger negative coefficient (thus larger impact) on reducing the fatality rate. They are statistically significant and it implies that they have a strong negative correlation to the total fatality rate. This pattern was observed in EDA where the boxplot showed lower fatality rate for *bac08* than *bac10*. The takeaway is, if we come up with a stricter law and decrease the blood alcohol limit to .08 then the fatalities rate decreases more.

Effect of perse law on fatality rate
Yes. *perse* variable (p-value 0.037791) has a statistically significant at the 95% confidence interval level, showing negative correlation with the total fatality rate. The coefficient value is -0.6201 which implies a small change in the rate.

Effect of primary seat belt law on fatality rate
Primary seatbelt has a negative coeffient (-7.533e-02) indicating that adoption of primary seatbelt reduces fatality rate. However, with a high p-value (0.878032), we are not able to find any statistically significant effect of the primary seat belt law on the fatality rate based on the regression coefficients shown below.

```r
library(lmtest)
lm.fit2 <- lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 +
                d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 +
                d03 + d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl +
                perc14_24 + unem + vehicmilespc, data=driving.df)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = driving.df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9160  -2.7384  -0.2778   2.2859  21.4203
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.716e+00  2.476e+00  -1.097 0.272847
## d81          -2.175e+00  8.276e-01  -2.629 0.008686 **
## d82          -6.596e+00  8.534e-01  -7.729 2.33e-14 ***
## d83          -7.397e+00  8.690e-01  -8.512  < 2e-16 ***
## d84          -5.850e+00  8.763e-01  -6.676 3.79e-11 ***
## d85          -6.483e+00  8.948e-01  -7.245 7.82e-13 ***
## d86          -5.853e+00  9.307e-01  -6.289 4.52e-10 ***
## d87          -6.367e+00  9.670e-01  -6.585 6.87e-11 ***
## d88          -6.592e+00  1.014e+00  -6.502 1.17e-10 ***
## d89          -8.071e+00  1.053e+00  -7.667 3.68e-14 ***
## d90          -8.959e+00  1.077e+00  -8.319 2.46e-16 ***
## d91          -1.107e+01  1.101e+00 -10.052  < 2e-16 ***
## d92          -1.288e+01  1.123e+00 -11.473  < 2e-16 ***
## d93          -1.273e+01  1.136e+00 -11.204  < 2e-16 ***
## d94          -1.236e+01  1.157e+00 -10.685  < 2e-16 ***
## d95          -1.195e+01  1.184e+00 -10.098  < 2e-16 ***
## d96          -1.388e+01  1.223e+00 -11.343  < 2e-16 ***
## d97          -1.426e+01  1.250e+00 -11.408  < 2e-16 ***
## d98          -1.504e+01  1.265e+00 -11.886  < 2e-16 ***
## d99          -1.509e+01  1.284e+00 -11.750  < 2e-16 ***
## d00          -1.544e+01  1.305e+00 -11.831  < 2e-16 ***
## d01          -1.618e+01  1.334e+00 -12.131  < 2e-16 ***
## d02          -1.672e+01  1.348e+00 -12.406  < 2e-16 ***
## d03          -1.702e+01  1.359e+00 -12.521  < 2e-16 ***
## d04          -1.671e+01  1.387e+00 -12.049  < 2e-16 ***
## bac08        -2.498e+00  5.375e-01  -4.648 3.73e-06 ***
## bac10        -1.418e+00  3.963e-01  -3.577 0.000362 ***
## perse        -6.201e-01  2.982e-01  -2.079 0.037791 *
## sbprim       -7.533e-02  4.908e-01  -0.153 0.878032
## sbsecon       6.728e-02  4.293e-01   0.157 0.875492
## sl70plus      3.348e+00  4.452e-01   7.521 1.09e-13 ***
## gdl          -4.269e-01  5.269e-01  -0.810 0.417978
## perc14_24     1.416e-01  1.227e-01   1.154 0.248675
## unem          7.571e-01  7.791e-02   9.718  < 2e-16 ***
## vehicmilespc  2.925e-03  9.497e-05  30.804  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.046 on 1165 degrees of freedom
## Multiple R-squared:  0.6078, Adjusted R-squared:  0.5963
## F-statistic:  53.1 on 34 and 1165 DF,  p-value: < 2.2e-16
```

```r
bgtest(lm.fit2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  lm.fit2
## LM test = 771.2, df = 1, p-value < 2.2e-16
```

In summary, including the pertinent explanatory variables improved the model fit, as seen in the adjusted R^2 (0.5963) and is better than the model in Q2.

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

Coeefficients comparison for *bac08, bac10, perse, and sbprim* between pooled OLS and Fixed effects model
Effect of coefficients on fatality rate have gone down for *bac08 (-2.498 to -1.437), bac10 (-1.418 to -1.062)* while they increased for *sbprim (-0.07533 to -1.227) and perse(-0.62 to -1.152)* for the fixed effects model. Furthermore *sbprim* has become statistically significant in the Fixed effects model.

Model reliability and assumptions
Fixed effects model is more reliable than the pooled OLS model as it considers the panel data into account and models the total fatality rate of each state over time.

For pooled OLS to be valid we need to ensure we account for serial correlation between the years and also account for heterogeneity bias due to omitted time-invariant variables (state). The pooled OLS requires the composite error to be uncorrelated with the explanatory variables. In this model we account for serial correlation across different time periods using the year dummies. However we are not accounting for heterogeneity bias due to unobserved fixed effect.

For the fixed affect model, allows for correlation between unobserved fixed effect and explanatory variables. By applying fixed effects (at the state level), we are accounting for the unobserved time-invariant variables (road conditions, percentage of rural vs urban roads, compliance to laws etc specific to state that doesn't change much over time). The serial correlation across time is accounted for by the year dummies.

Fixed effect model is more reliable because it is accounting for the unobserved time-invariant variable which pooled OLS is not.

Our fixed effect model assumption is that there could be correlation between the unobserved fixed effect (e.g. road conditions, percentage of rural vs urban roads, compliance to laws etc specific to state that doesn't change much over time) and explanatory variables (e.g. bac08, vehicmilespc). This assumption is reasonable because the EDA showed there is correlation between fatality rate and state, and traffic law adoption and state.

```r
# Creating a panel with 'State' and 'Year' variables.
pnldata <- pdata.frame(driving.df, c("state", "year"))

model.fe <- plm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 +
                d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 +
```

```r
                 d03 + d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl +
                 perc14_24 + unem + vehicmilespc, data=pnldata, model = "within")
summary(model.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 +
##     d87 + d88 + d89 + d90 + d91 + d92 + d93 + d94 + d95 + d96 +
##     d97 + d98 + d99 + d00 + d01 + d02 + d03 + d04 + bac08 + bac10 +
##     perse + sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem +
##     vehicmilespc, data = pnldata, model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -8.4273592 -1.0258600 -0.0029547  0.9572345 14.8109310
##
## Coefficients:
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## d81          -1.51107133  0.41321486  -3.6569 0.0002672 ***
## d82          -3.02549578  0.44243119  -6.8383 1.316e-11 ***
## d83          -3.50360069  0.45657705  -7.6736 3.628e-14 ***
## d84          -4.25936110  0.46494255  -9.1610 < 2.2e-16 ***
## d85          -4.72679311  0.48547032  -9.7365 < 2.2e-16 ***
## d86          -3.66118539  0.51769787  -7.0721 2.686e-12 ***
## d87          -4.30578838  0.55532856  -7.7536 2.001e-14 ***
## d88          -4.76712131  0.60155650  -7.9246 5.501e-15 ***
## d89          -6.12997263  0.64019069  -9.5752 < 2.2e-16 ***
## d90          -6.22973766  0.66485076  -9.3701 < 2.2e-16 ***
## d91          -6.91714040  0.68195432 -10.1431 < 2.2e-16 ***
## d92          -7.77417239  0.70288580 -11.0604 < 2.2e-16 ***
## d93          -8.09410864  0.71594741 -11.3055 < 2.2e-16 ***
## d94          -8.50421668  0.73410866 -11.5844 < 2.2e-16 ***
## d95          -8.25540198  0.75623634 -10.9164 < 2.2e-16 ***
## d96          -8.60661913  0.79594975 -10.8130 < 2.2e-16 ***
## d97          -8.70781739  0.81975686 -10.6224 < 2.2e-16 ***
## d98          -9.34924025  0.83373487 -11.2137 < 2.2e-16 ***
## d99          -9.47489124  0.84399083 -11.2263 < 2.2e-16 ***
## d00          -9.99185979  0.85606370 -11.6719 < 2.2e-16 ***
## d01          -9.63121721  0.87255395 -11.0380 < 2.2e-16 ***
## d02          -8.90673015  0.88205263 -10.0977 < 2.2e-16 ***
## d03          -8.93650263  0.88994687 -10.0416 < 2.2e-16 ***
## d04          -9.33936116  0.91107045 -10.2510 < 2.2e-16 ***
## bac08        -1.43722116  0.39421213  -3.6458 0.0002788 ***
## bac10        -1.06266776  0.26883763  -3.9528 8.208e-05 ***
```

```
## perse        -1.15161719  0.23398721  -4.9217 9.867e-07 ***
## sbprim       -1.22739974  0.34271485  -3.5814 0.0003564 ***
## sbsecon      -0.34970784  0.25217091  -1.3868 0.1657826
## sl70plus     -0.06253283  0.26931063  -0.2322 0.8164283
## gdl          -0.41177619  0.29257391  -1.4074 0.1595790
## perc14_24     0.18712169  0.09509969   1.9676 0.0493567 *
## unem         -0.57183997  0.06057851  -9.4397 < 2.2e-16 ***
## vehicmilespc  0.00094005  0.00011104   8.4656 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 4535.3
## R-Squared:      0.62624
## Adj. R-Squared: 0.59916
## F-statistic: 55.0943 on 34 and 1118 DF, p-value: < 2.22e-16
```

The model is

$totfatrte_i = -1.51107133d81 - 3.02549578d82 - 3.50360069d83 - 4.25936110d84 - 4.72679311d85 - 3.66118539d86 - 4.$

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

We will fit the random effects model to the data as shown below.

```
model.re <- plm(totfatrte ~  d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 +
                d91 + d92 + d93 + d94 + d95 + d96 + d97 + d98 + d99 + d00 + d01 + d02 +
                d03 + d04 + bac08 + bac10 + perse + sbprim + sbsecon + sl70plus + gdl +
                perc14_24 + unem + vehicmilespc, data=pnldata, model = "random")

phtest(model.fe, model.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 +  ...
## chisq = 148.69, df = 34, p-value = 2.727e-16
## alternative hypothesis: one model is inconsistent
```

RE model assumes that the variation across entities is random and uncorrelated with the predictors in the model. EDA showed us that the differences across states are consistent over years, and institutional knowledge tells us that there are state specific factors that could be influencing the fatality rate that is omitted in the dataset - for example road conditions specific to state could be correlated to vehicmilespc. FE will clear them out by demeaning over time, but RE will not. Hence the intuition is FE is better. We confirm the intuition with Hausman test.

Hausman test checks whether the unique errors u_i are correlated with the explanatory variables; the null hypothesis is they are not, and a lower p-value indicates that fixed effects model is better than random effects model. In the above test, p-value (2.727e-16) is statistically significant and we can reject the null hypothesis that the unobserved fixed effects are uncorrelated with the explanatory variables. Therefore, We will prefer the Fixed effect model instead of the random effects model in this scenario.

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

The coefficient for the *vehicmilespc* variable is 0.00094005 using the FE estimates and it is highly statistically significant. In other words, There will be an increase of 0.94 fatalities per 100k for an increase of 1000 vehicle miles driven per capita.

7. (5%) If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the estimators and their standard errors?
   #TODO read and update
   If there is Serial correlation then it will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency. With positive serial correlation, the OLS estimates of the standard errors will be smaller than the true standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are. There will be a tendency to reject the null hypothesis when it should not be rejected. Much of the time serial correlation is viewed as the most important problem, because it usually has a larger impact on standard errors and the efficiency of estimators than does heteroskedasticity. The presence of heteroskedasticity, while not causing bias or inconsistency in the coefficients does invalidate the usual standard errors, t statistics, and F statistics. To address this, the usual OLS standard errors, t statistics, and F statistics can be adjusted to allow for the presence of heteroskedasticity of unknown form (Heteroskedasticity-robust statistics)

Conclusion Traffic fatalities from crashes in the United States substantial. Good Statistical models can identify important predictors that can help authorities to direct focus and resources, and policy analysis.