W271 Lab3

Isaac Law, Mayukh Dutta, Mike King

Question 1

```
# load data
load(file = "driving.RData")
sum(is.na(data))
```

```
## [1] 0
```

```
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 29
## 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 25
## 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
```

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

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

There are no missing data and this is a **balanced panel**. For each state there are 25 observations, correponding to 25 years. For each year, there are 48 observations, corresponding to 48 states.

Our primary outcome variable is *totfatrte*. Since we are interested in how changes in law associate with changes in fatalities, the potential explanatory variables are those pertaining to legal matters. The variables pertaining to legal matters are *sl55*, *sl65*, *sl70*. *sl75*, *slnone*, *sbprim*, *sbsecon*, *minage*, *zerotol*, *gdl*, *bac10*, *bac08*, and *perse*. All of these are binary indicator variables (with a few exceptions if a law changed in the middle of the year for that observation's state.) The *sl70plus* variable is a calculation of *sl70* or *sl75* or *slnone*. The *seatbelt* variable is an integer-coded factor indicating no seat belt law, secondary seatbelt law (*sbsecon*), and primary seatbelt law (*sbprim*). Vairables *unem*, *perc14_24*, and *vehicmilespc* are not laws, but they could be relevant either standalone or as interaction terms. The remaining variables are variations on fatality rates measured in different ways, or are values necessary to calculate the various fatality rates. Therefore they are not candidates to explain *totfatrte*.

We can begin with a Durbin-Watson test to see if there is correlation in the residuals of a basic linear model:

```
dwtest(totfatrte ~ seatbelt+minage+zerotol+sl70plus+gdl+bac08+bac10+
          perse+vehicmilespc+unem+perc14_24, data=data)
```

```
##
##  Durbin-Watson test
##
## data:  totfatrte ~ seatbelt + minage + zerotol + sl70plus + gdl + bac08 +     bac10 + perse
## DW = 0.42474, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

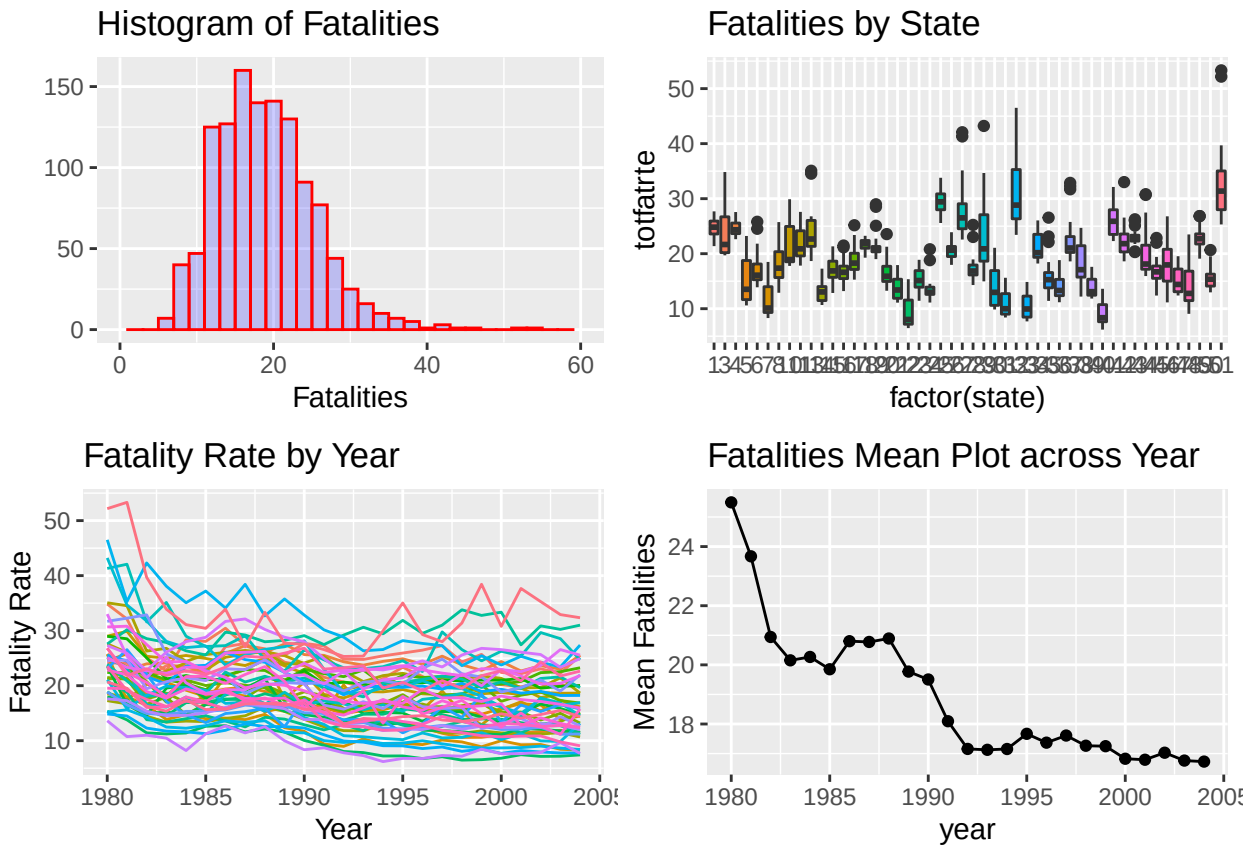
The null hypothesis of non-correlation is rejected, confirming the violation of independence assumption. Therefore we will have to use panel methods for analysis.

We must be careful when thinking about our primary outcome variable as a function of these explanatory variables. Many other changes occurred over this time period that could be explanatory for fatality rates. Automotive technology changed (airbags, anti-lock brakes, influx of Japanese-built cars), national infrastructure changed (more multi-lane highways, nighttime lighting), and many other technology and cultural changes occurred as well (cell phones, urbanization, life expectancy.) There are potentially many unobserved explanatory variables. The premise of our endeavor implies causality, asking if changes in laws can drive changes in fatalities. It must be noted that this is observational, not experimental, data. Although we may think of hypothetical changes to model parameters "causing" changes in estimated outcomes (question 6), we certainly make no claims of real-world causality should such changes be induced by intervention.

A look at various distributions of our outcome variable, *totfatrte*:

```
# dependent variable, totfatrte
p1<-qplot(data$totfatrte,geom="histogram",binwidth =2,main = "Histogram of Fatalities",
          xlab = "Fatalities",fill=I("blue"),col=I("red"),alpha=I(.2),xlim=c(0,60))
p2 <- ggplot(data, aes(factor(state), totfatrte))+ggtitle("Fatalities by State")+
    theme(plot.title = element_text(lineheight=1))+geom_boxplot(aes(fill = factor(state)), sho
p3 <-ggplot(data, aes(year, totfatrte)) + geom_line(aes(col = as.factor(state))) +
    ggtitle("Fatality Rate by Year") + xlab("Year") + ylab("Fatality Rate") + theme(legend.pos
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(totfatrte))%>%ggplot(aes(x=year, y=mea
grid.arrange(p1,p2,p3,p4,nrow=2)
```

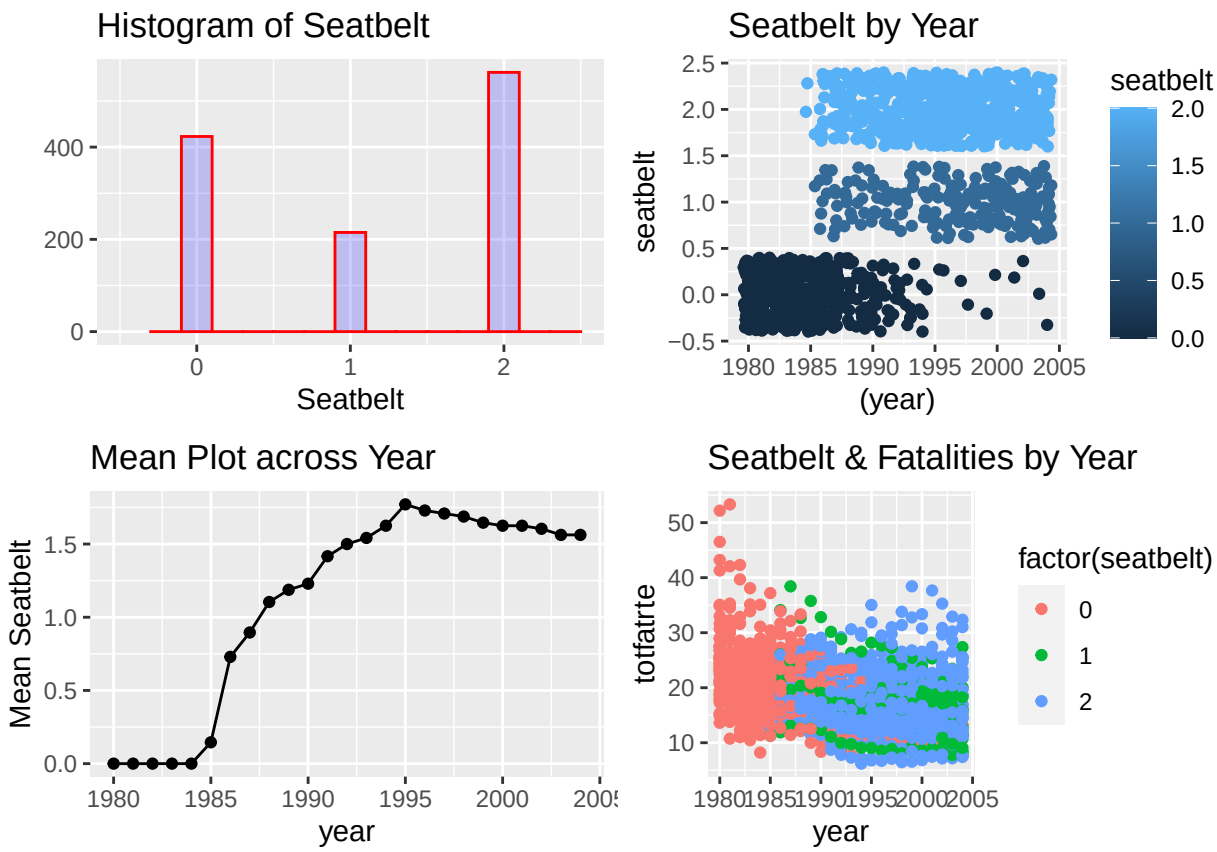
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Most of values of fatalities are between 10 to 30 per 100,000 population. We observed that state 32 and state 51 have higher fatalities rate. Over the years, we observed that there are 2 periods which fatality rate changed more than any other time period. The first is 1980 through 1983 and the second is 1988 through 1992. The other timespans (1983-1988 and 1992-2004) show nearly stationary fatality rates. The overall trend is downward, particularly from 1980 through 1992.

A look at seatbelt law changes over the timespan:

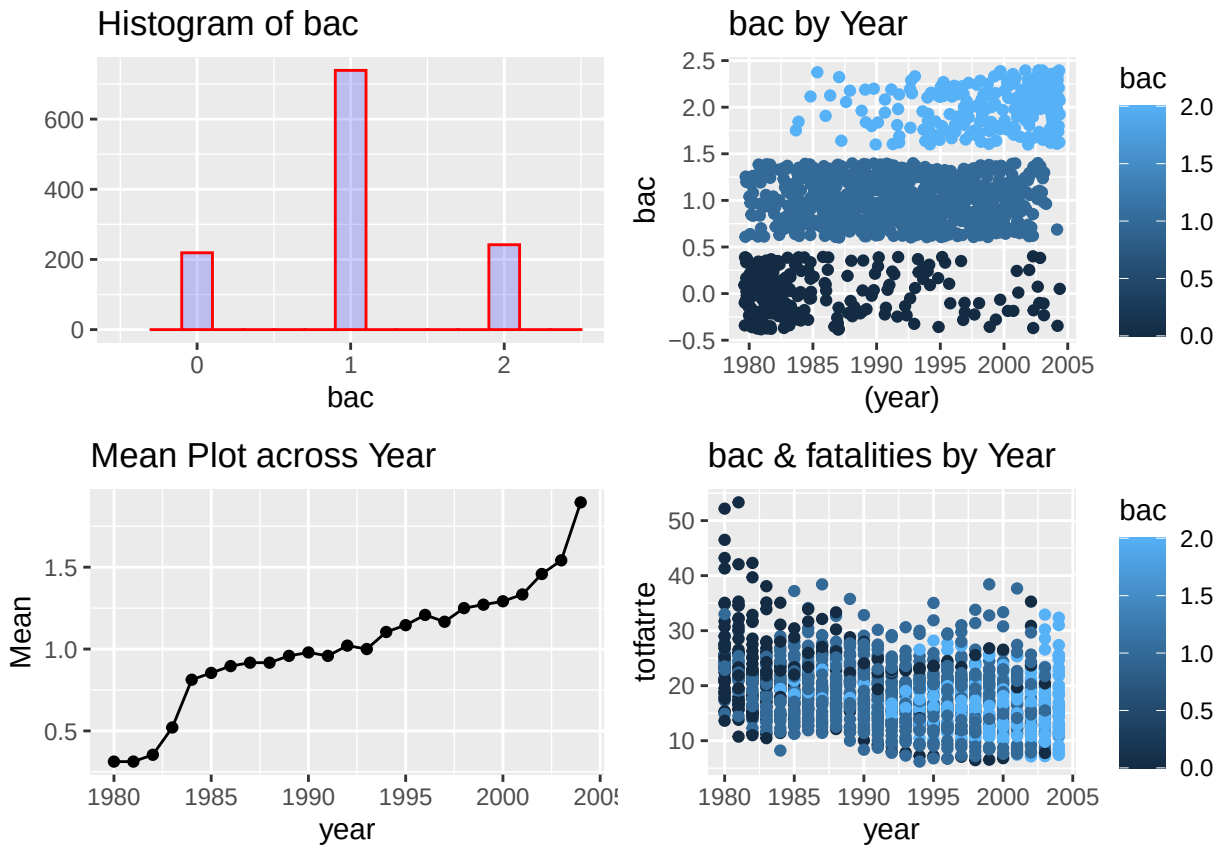
```
p1<-qplot(data$seatbelt,geom="histogram",binwidth =0.2,main = "Histogram of Seatbelt", xlab = "
p2<-ggplot(data, aes((year), seatbelt))+geom_jitter(aes(color=seatbelt)) +ggtitle("Seatbelt by
p3<-data %>% group_by(year)%>%summarise(mean_group=mean(seatbelt))%>%ggplot(aes(x=year, y=mean_
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=factor(seatbelt)))+ggtitle("Seatbe
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Over the years, states started with mostly no seatbelt laws until 1985. Since 1985 states changed to have primary seatbelt law and secondary seatbelt law. From 1995, some more states change from secondary seatbelt law to primary seatbelt law. Primary seatbelt law is stricter than secondary seatbelt law, we observe stricter seatbelt law implemented by states over years. The implementation of more strict seatbelt laws appears to correlate negatively with fatality rates, suggesting that seatbelt laws are worthwhile to consider as explanatory variables.

A look at the limits of blood-alcohol content for a DWI violation:

```
data$bac <- ifelse(round(data$bac08)==1, 2, ifelse(round(data$bac10)==1,1,0))
p1<-qplot(data$bac,geom="histogram",binwidth =0.2,main = "Histogram of bac", xlab = "bac",fill=
p2<-ggplot(data, aes((year), bac))+geom_jitter(aes(color=bac)) +ggtitle("bac by Year") + theme
p3<-data %>% group_by(year)%>%summarise(mean_group=mean(bac))%>%ggplot(aes(x=year, y=mean_grou
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=bac))+ ggtitle("bac & fatalities
grid.arrange(p1,p2,p3,p4,nrow=2)
```

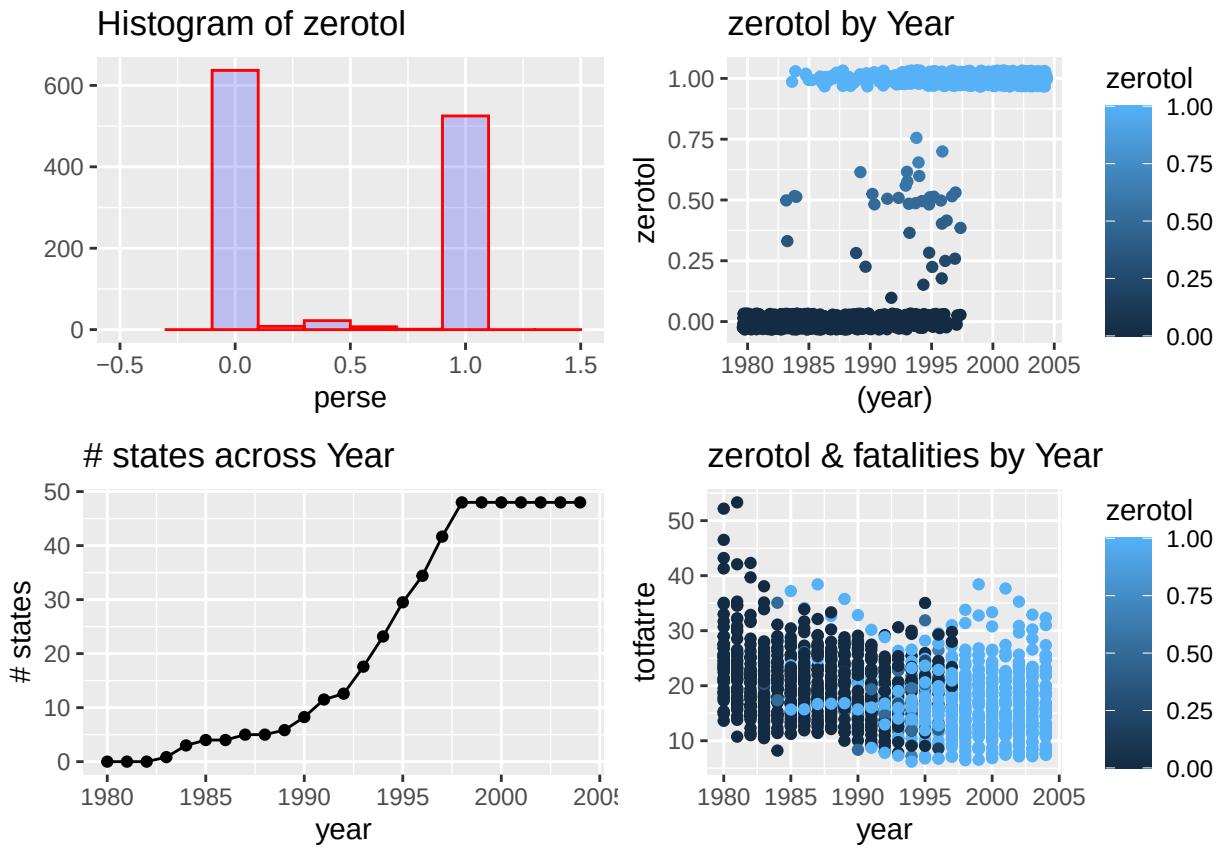


In these plot, value 2 corresponds to blood alcohol level 0.08% and value 1 correspond to blood alcohol level 0.10%. Majority of data points in dataset have value of 1. Over year we observe that majority of states had no law on blood alcohol level, and then states started to implement law on blood alcohol level with 0.10%, and then more states implemented law on blood alcohol level 0.08%. And the end of the time period, majority of states had law on blood alcohol level with 0.08%. There is pattern of stricter law over years, and this is correlated with lower fatality rates over years.

A look at zero-tolerance policy laws:

```
p1<-qplot(data$zerotol,geom="histogram",binwidth =0.2,main = "Histogram of zerotol", xlab = "pe
p2<-ggplot(data, aes((year), zerotol))+geom_jitter(aes(color=zerotol)) +ggtitle("zerotol by Yea
p3<-data %>% group_by(year)%>%summarise(sum_group=sum(zerotol))%>%ggplot(aes(x=year, y=sum_grou
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=zerotol))+ggtitle("zerotol & fatal
grid.arrange(p1,p2,p3,p4,nrow=2)
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

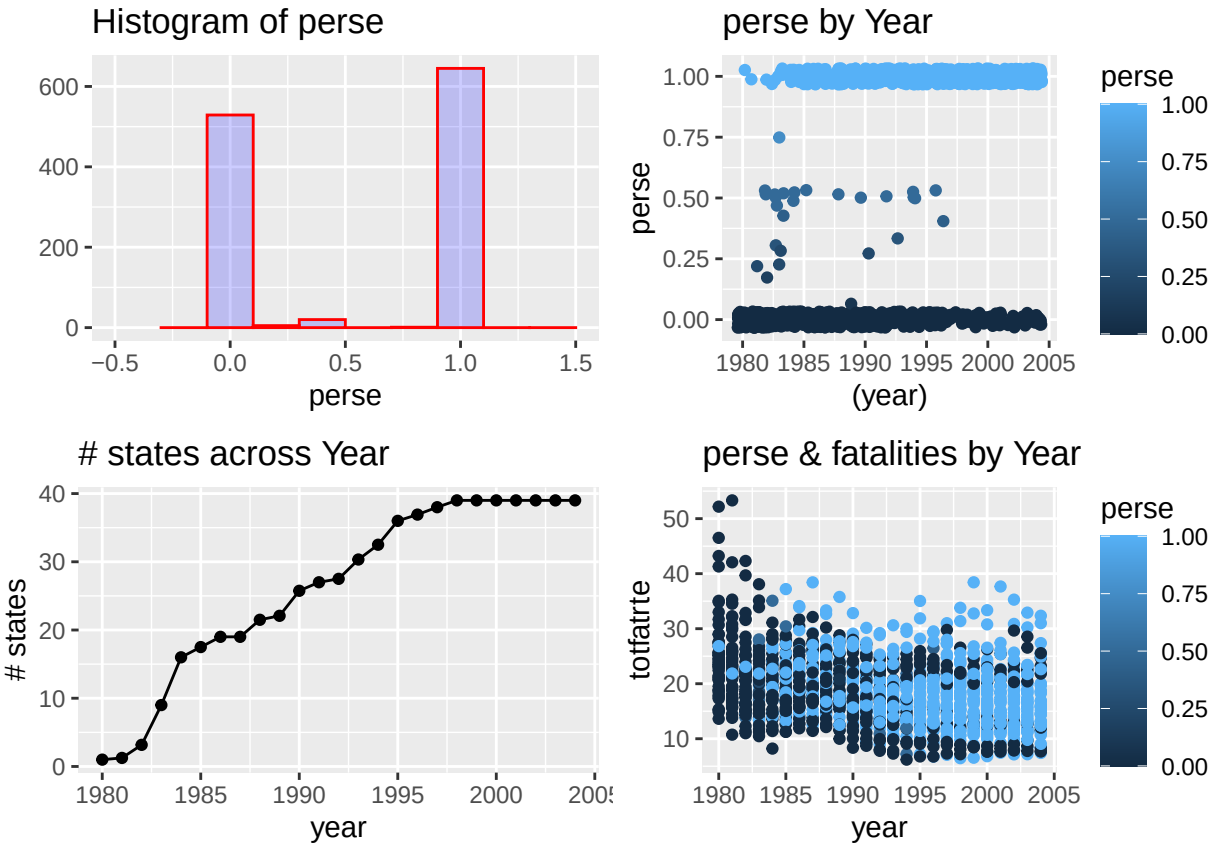


At the start of time period, no states have zero tolerance law. Starting from year 1983, some states started to have zero tolerance law and more states followed in subsequent years. Since year 1996, all states have zero tolerance law. The implementation of this stricter law appears to be correlated with the decreasing fatality rates over years.

A look at per-se laws which allow a judge to revoke a driver's license without a "per-se" related driving violation:

```
p1<-qplot(data$perse,geom="histogram",binwidth =0.2,main = "Histogram of perse", xlab = "perse"
p2<-ggplot(data, aes((year), perse))+geom_jitter(aes(color=perse)) +ggtitle("perse by Year") +
p3<-data %>% group_by(year)%>%summarise(sum_group=sum(perse))%>%ggplot(aes(x=year, y=sum_group)
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=perse)) + ggtitle("perse & fatali
grid.arrange(p1,p2,p3,p4,nrow=2)
```

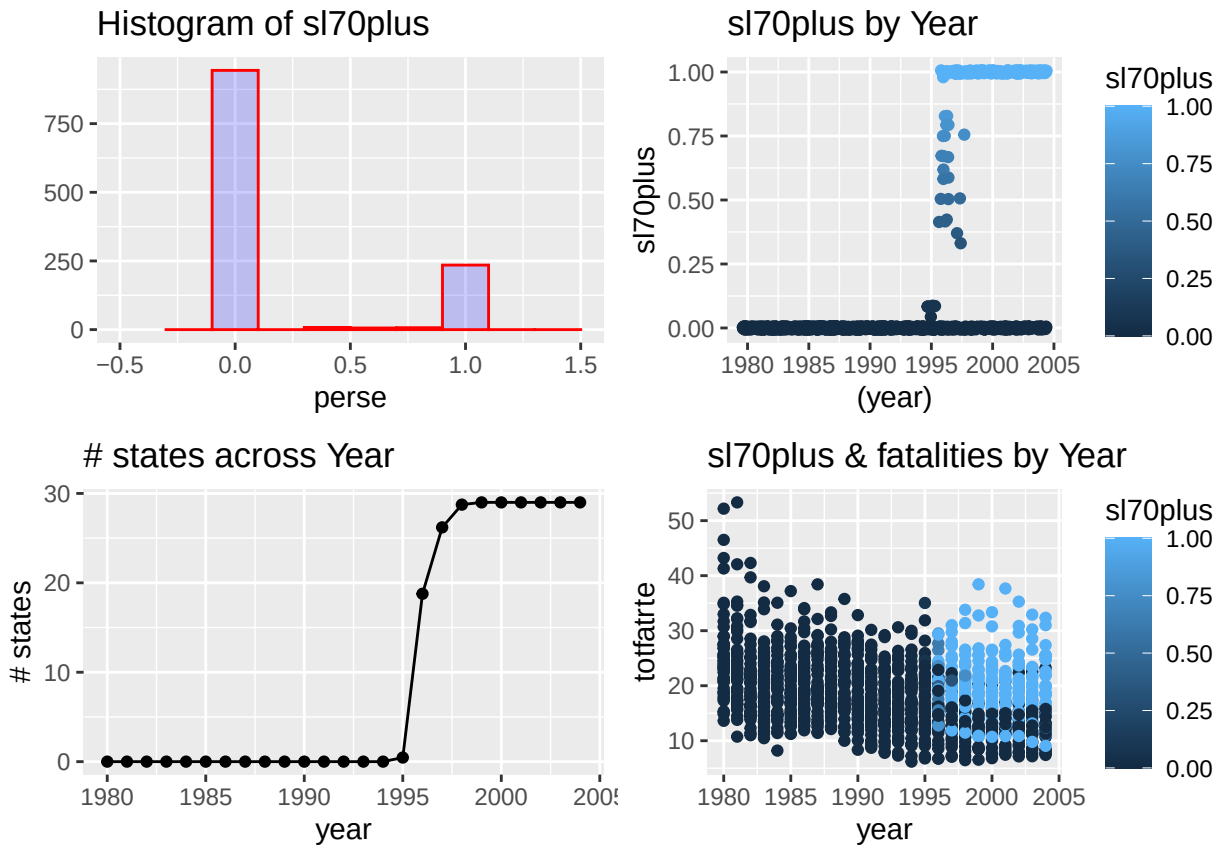
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



Around half of data points have value of 0 and the other half have values of 1. Over time, we observe that more states implemented the per se law. This increasing trend show some correlation with fatality rates. The handful of data points valued neither 0 nor 1 are observations where the law was changed mid-year. In subsequent discussion we will recall that these points are a small minority, and also that our outcome variable *totfatrte* was not adjusted analogously (i.e. fatalities in first 6 months averaged with fatalities in last 6 months).

Now we look at speed limits. Speed limits are different from the other explanatory variables because they have been relaxed over the years, whereas the other laws have become more strict. The *sl70plus* variable is a reasonable proxy for the various speed limit changes over time, so we will focus on that.

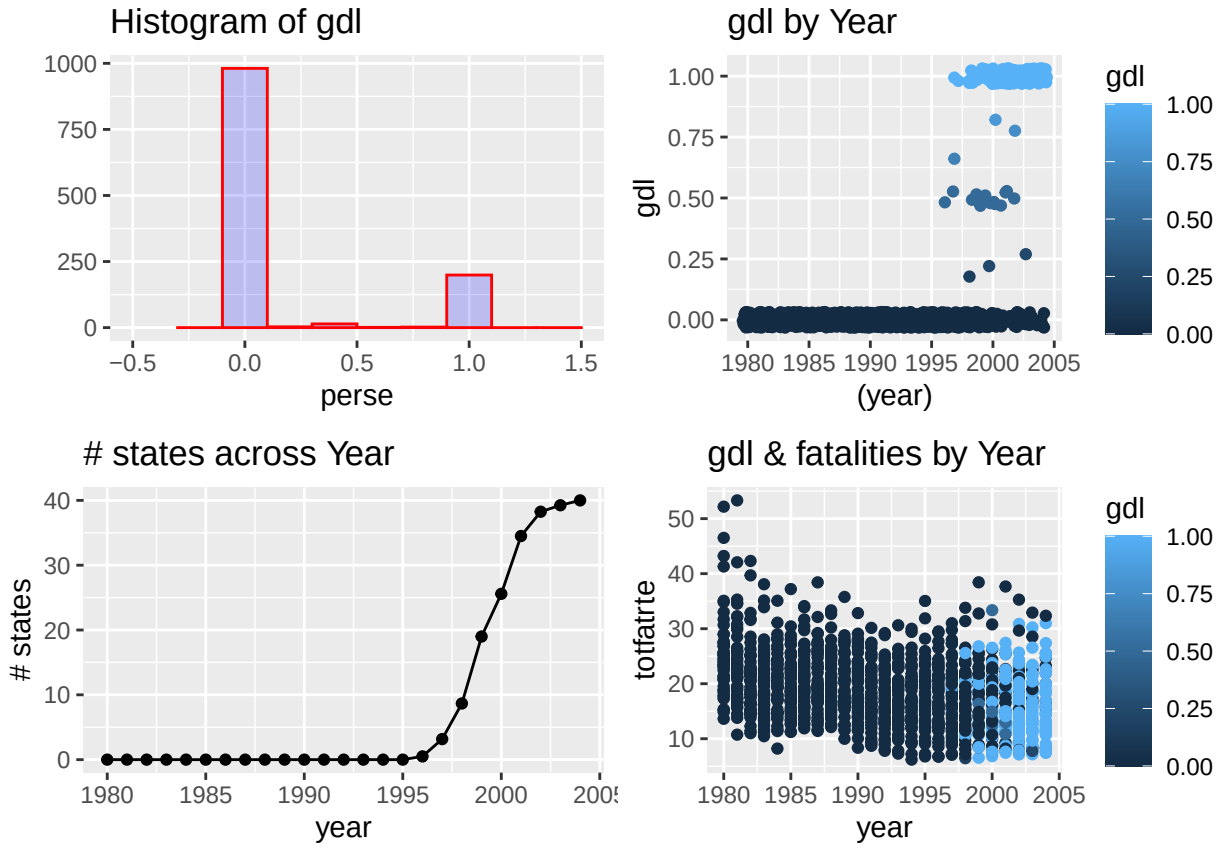
```
p1<-qplot(data$sl70plus,geom="histogram",binwidth =0.2,main = "Histogram of sl70plus", xlab = '
p2<-ggplot(data, aes((year), sl70plus))+geom_jitter(aes(color=sl70plus)) +ggtitle("sl70plus by
p3<-data %>% group_by(year)%>%summarise(sum_group=sum(sl70plus))%>%ggplot(aes(x=year, y=sum_gr
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=sl70plus)) + ggtitle("sl70plus & t
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Most data points are with value 0. All states have stricter speed limit at before year 1995. Since 1995, some states started to relax the limit on speed and after year 1999, around 30 states have speed limit of 70, 75 or no limit.

A look at laws offering graduated driver's licensing:

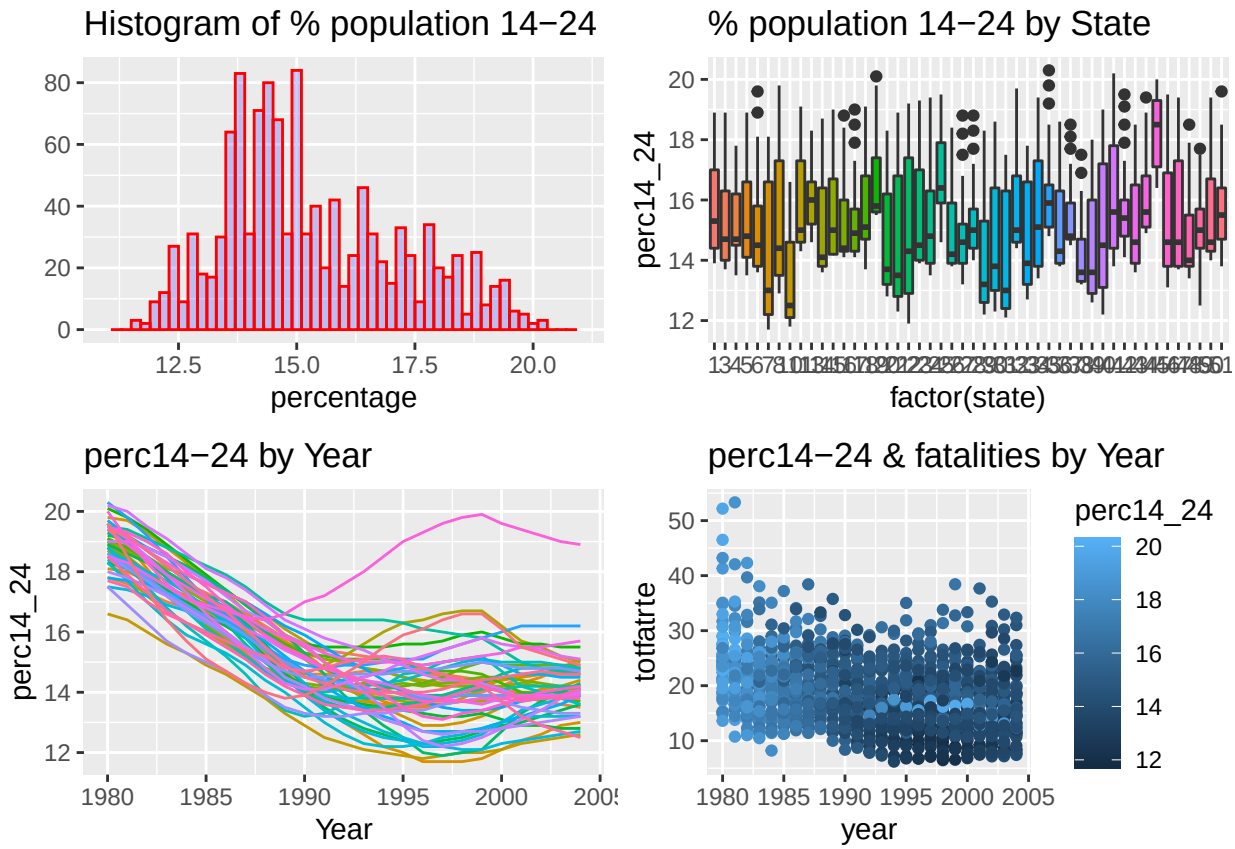
```
p1<-qplot(data$gdl,geom="histogram",binwidth =0.2,main = "Histogram of gdl", xlab = "perse",fi
p2<-ggplot(data, aes((year), gdl))+geom_jitter(aes(color=gdl)) +ggtitle("gdl by Year") + theme
p3<-data %>% group_by(year)%>%summarise(sum_group=sum(gdl))%>%ggplot(aes(x=year, y=sum_group))
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=gdl))+ggtitle("gdl & fatalities by
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Most data points in dataset are with value 0. From plots across years, we observed that states started to have graduated drivers license law since 1996, and more states follow this trend. At then end of time period, most of states implemented this law. During this period of increasing states implementing GDL, fatality rates does not seem to show decreasing trend.

A look at what may be considered the fraction of "youth" drivers (those aged 14-24 years):

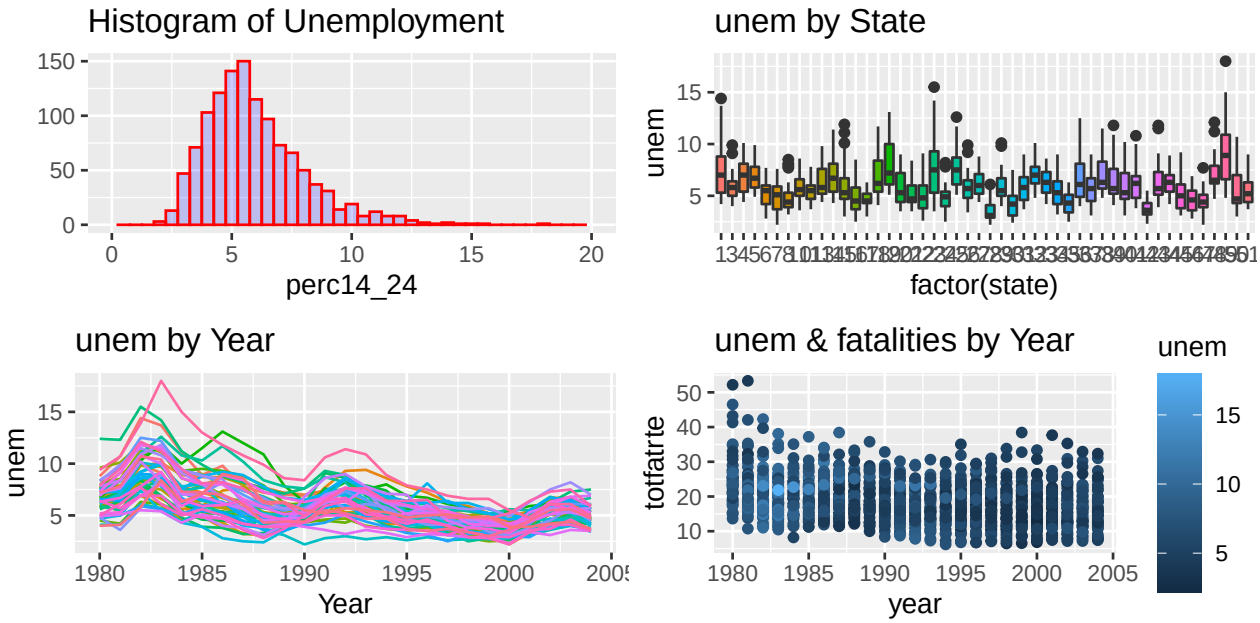
```
p1<-qplot(data$perc14_24,geom="histogram",binwidth = 0.2,main = "Histogram of % population 14-
p2<-ggplot(data, aes(factor(state), perc14_24))+geom_boxplot(aes(fill = factor(state)),show.leg
p3 <-ggplot(data, aes(year, perc14_24)) + geom_line(aes(col = as.factor(state))) + ggtitle("pe
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=perc14_24))+ggtitle("perc14-24 & 
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Majority of values are in range of 13% to 18%. We observed that state 45 has higher percentage of 14-24 years old population, especially after year 1988. The decreasing trend of % population of 14-24 show some degree of correlation to the decreasing trend of fatality rate. There is one state that is a clear outlier in the trend of older population, showing anomalous increase in *perc_14_24* from 1987 through 1997. It appears that there are a few other states showing increases as well, although perhaps to a lesser degree.

A look at unemployment:

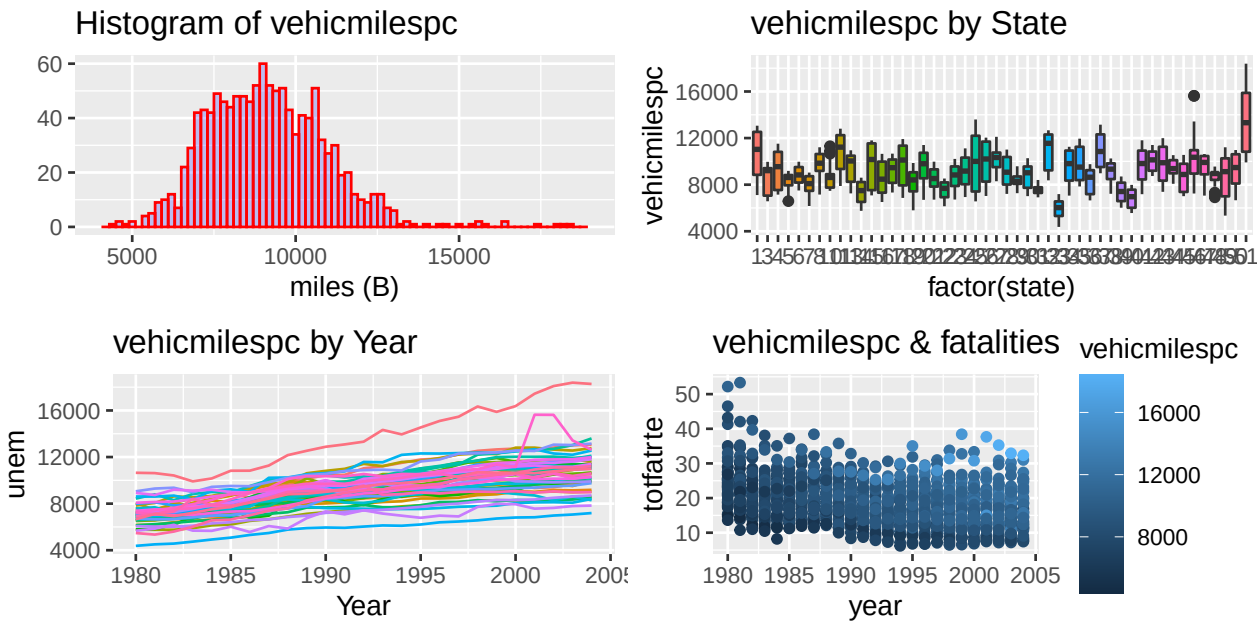
```
p1<-qplot(data$unem,geom="histogram",binwidth = 0.5,main="Histogram of Unemployment",xlab = "pe
p2<-ggplot(data, aes(factor(state), unem))+geom_boxplot(aes(fill = factor(state)),show.legend=F
p3 <-ggplot(data, aes(year, unem)) + geom_line(aes(col = as.factor(state))) + ggtitle("unem by
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=unem))+ggtitle("unem & fatalities
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Most of values are between 3% to 10%, with some outliers above 14%. We observe that state 41 has higher unemployment rate across years. The decreasing trend of unemployment rate show correlation with the decreasing trend of fatality rates.

A look at vehicle miles per capita:

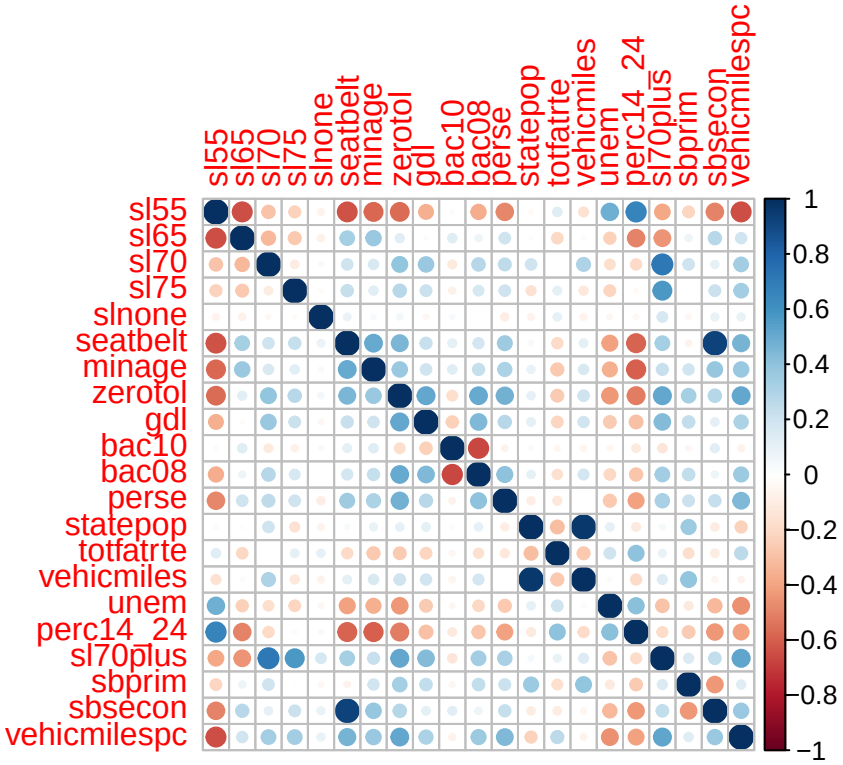
```
p1<-qplot(data$vehicmilespc,geom="histogram",binwidth = 200,main = "Histogram of vehicmilespc"
p2<-ggplot(data, aes(factor(state), vehicmilespc))+geom_boxplot(aes(fill = factor(state)),show
p3 <-ggplot(data, aes(year, vehicmilespc)) + geom_line(aes(col = as.factor(state))) + ggtitle(
p4<-ggplot(data,aes(year, totfatrte)) + geom_point(aes(color=vehicmilespc))+ggtitle("vehicmiles
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Most values are between 5000 to 14000, with some outliers above 15000. State 45 has higher vehicle miles traveled across years consistently. There is a general upward trend across years and this shows

some degree of negative correlation with fatality rates.

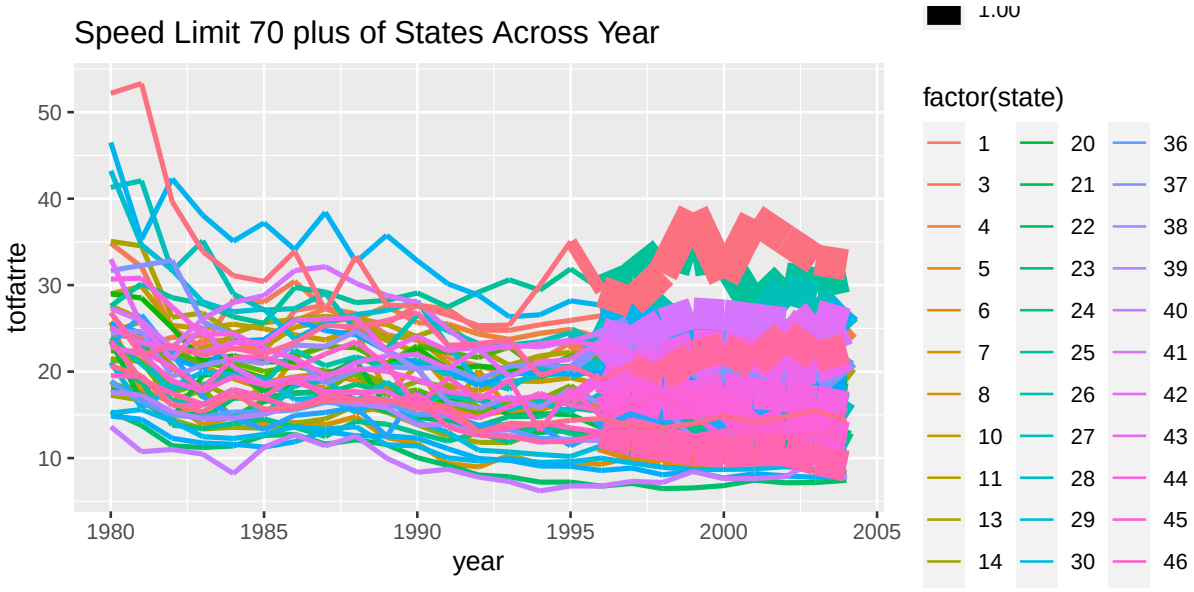
```
# Correlation plot
driving.panel <- pdata.frame(data, c("state","year"), drop.index = TRUE);corrplot(cor(driving.p
```

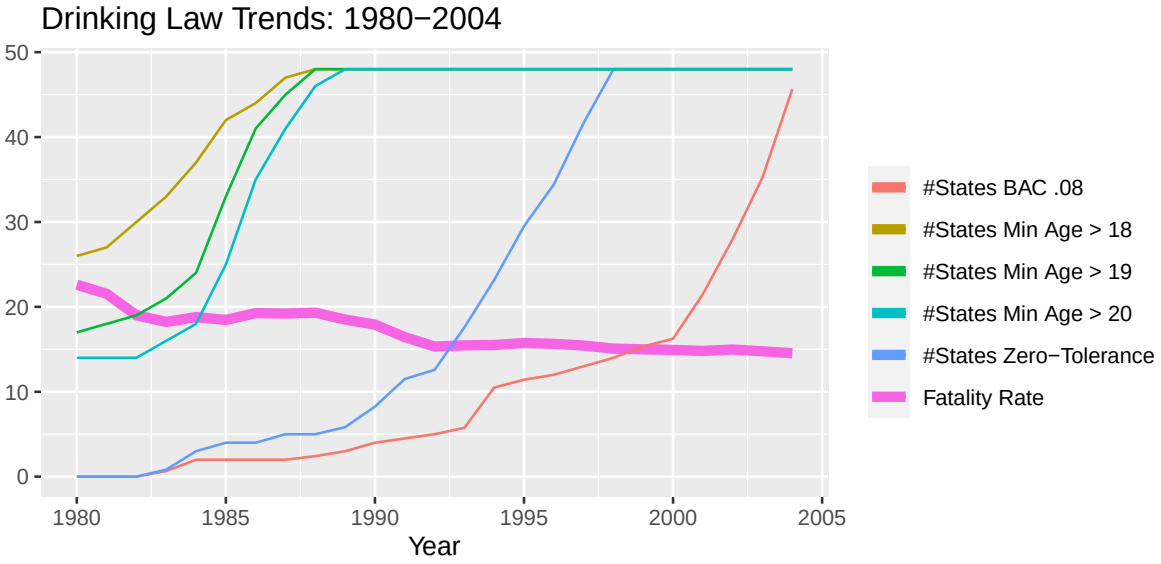


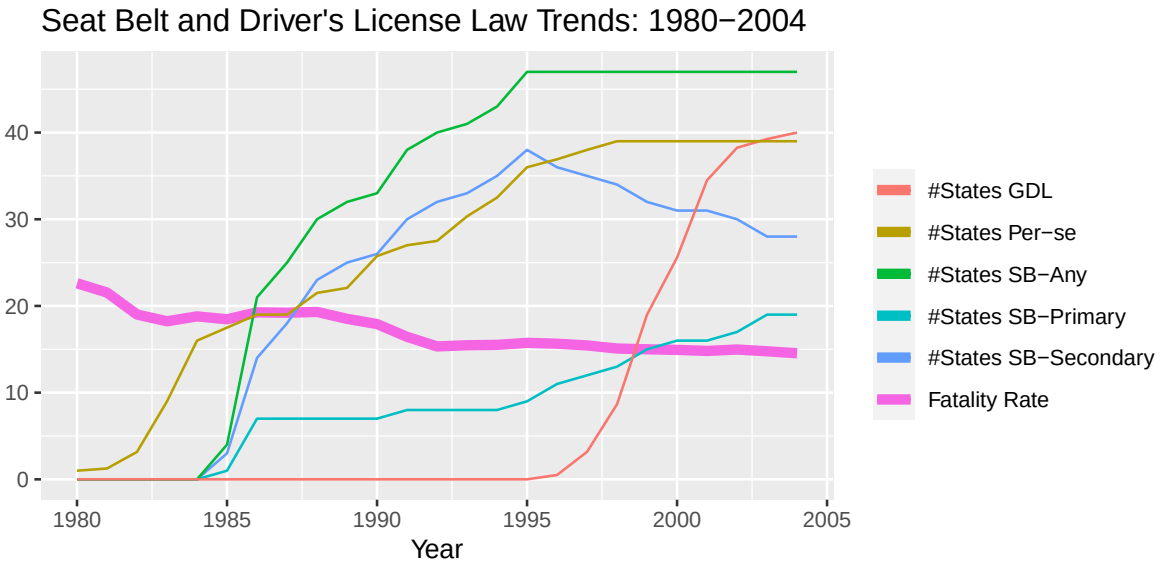
We see that the "perc_14_24" which is the percent population between 14 and 24 has a very strong correlation with the "fatality rate". The 'minimum age' has a negative correlation with the fatality rate, so is the 'zero tolerance' even though it is not very strong.

Next we move to a detail bivariate EDA to explore relationship between explanatory variables and outcome variable. We aggregate the data, combine the separate values by state into appropriate sums/averages for the nation. The intention here is to see how states have changed their laws over time. We look at how speed limit laws, drinking laws, seatbelt and driver's license law have changed:

```
aggregatedData = data %>% mutate(sbany = as.integer(seatbelt > 0),slGt55 = as.integer((sl65 + s
p1<-ggplot(data, aes(x=year, y=totfatrte, size=sl70plus, color=factor(state))) + geom_line()+gg
p2<-aggregatedData %>% ggplot(aes(year))+geom_line(aes(y = totfatrte, col = "Fatality Rate"), s
p3<-aggregatedData %>% ggplot(aes(year))+geom_line(aes(y = totfatrte, col = "Fatality Rate"), s
grid.arrange(p1,p2,p3,nrow=3)
```


Speed Limit 70 plus of States Across Year
totfatrte
year
factor(state)
1.00
1
3
4
5
6
7
8
10
11
13
14
20
21
22
23
24
25
26
27
28
29
30
36
37
38
39
40
41
42
43
44
45
46


Drinking Law Trends: 1980−2004
Year
#States BAC .08
#States Min Age > 18
#States Min Age > 19
#States Min Age > 20
#States Zero−Tolerance
Fatality Rate


Seat Belt and Driver's License Law Trends: 1980−2004
Year
#States GDL
#States Per−se
#States SB−Any
#States SB−Primary
#States SB−Secondary
Fatality Rate

The first plot above shows that while several of the states moved to a 70mph speed limit, the effect of the change on the fatality rate was different for different states. Observe, how, for some states the fatality rates increased after the change to *sl70plus* while for a very few they continued on the downward trend. This makes us believe that there are some unobserved city effects that effect the fatality rate for a given change in speed limit. Therefore, the explanatory variable *sl70plus* may not be an important explanatory variable for the fatality rate.

From second plot, the drop in fatality rates from 1980-1983 correlates with the increased minimum drinking age in several states. The drop in fatality rates from 1988-1992 coincide with the increase in zero-tolerance states and the increase in BAC08 states. As with the speed limit laws, there are no years with reversions to prior values. The minimum drinking age either increases or remains the same, and the same for zero-tolerance and BAC08.

From third plot, no obvious correlation between seatbelt laws is seen with the 1980-1983 drop in fatality rate, but the 1988-1992 drop occurs over a time when seatbelt laws and per-se laws were increasing quickly. The graduated license law does not seem correlated with any meaningful drop in fatality rate, especially compared to seatbelt-primary over the time when the GDL laws were adopted.

Question 2

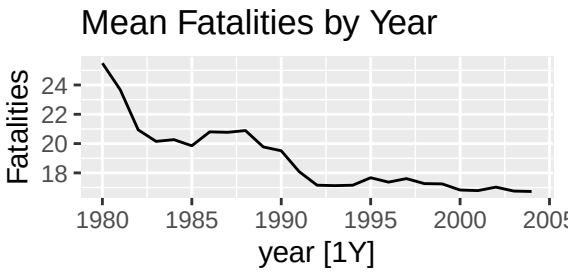
Variable *totfatrte* is defined as total number of fatalities in 100,000 population. This calculation is per-state and states have (very) different populations. Since our data set has one row per state per year, we are giving states with lower/higher populations equal weight in the model(s). This is reasonable because the various explanatory variables (laws) change at the state level from year to year.

As requested we can calculate the average of this variable for each year in the data set:

```
byYear.mean <- aggregate(data, by=list(data$year), FUN=mean)
round(byYear.mean$totfatrte, 2)
```

```
##  [1] 25.49 23.67 20.94 20.15 20.27 19.85 20.80 20.77 20.89 19.77 19.51 18.09
## [13] 17.16 17.13 17.16 17.67 17.37 17.61 17.27 17.25 16.83 16.79 17.03 16.76
## [25] 16.73
```

```
mean.totfatrte.df = round(data.frame(year=1980:2004, mean.totfatrte=byYear.mean$totfatrte), 2)
as_tsibble(mean.totfatrte.df,index=year)%>%autoplot(mean.totfatrte)+ggtitle("Mean Fatalities by
```



Mean of total fatalities show decreasing trend over years. After year 1992, when mean fatalities drop below 18, this number show a stable trend. A model with each year as a dummy variable(factor):

$$totfatrte_{it} = \beta_0 + \delta_{81}d81 + \delta_{82}d82 + ... + \delta_{04}d04 + u_{it}$$

```r
# Linear Regression
fit.lm <- lm(totfatrte ~ factor(year), data=data);summary(fit.lm)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year), data = data)
##
## 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 ***
## factor(year)1981   -1.8244     1.2263  -1.488 0.137094
## factor(year)1982   -4.5521     1.2263  -3.712 0.000215 ***
## factor(year)1983   -5.3417     1.2263  -4.356 1.44e-05 ***
## factor(year)1984   -5.2271     1.2263  -4.263 2.18e-05 ***
## factor(year)1985   -5.6431     1.2263  -4.602 4.64e-06 ***
## factor(year)1986   -4.6942     1.2263  -3.828 0.000136 ***
## factor(year)1987   -4.7198     1.2263  -3.849 0.000125 ***
## factor(year)1988   -4.6029     1.2263  -3.754 0.000183 ***
## factor(year)1989   -5.7223     1.2263  -4.666 3.42e-06 ***
## factor(year)1990   -5.9894     1.2263  -4.884 1.18e-06 ***
## factor(year)1991   -7.3998     1.2263  -6.034 2.14e-09 ***
## factor(year)1992   -8.3367     1.2263  -6.798 1.68e-11 ***
## factor(year)1993   -8.3669     1.2263  -6.823 1.43e-11 ***
## factor(year)1994   -8.3394     1.2263  -6.800 1.66e-11 ***
## factor(year)1995   -7.8260     1.2263  -6.382 2.51e-10 ***
## factor(year)1996   -8.1252     1.2263  -6.626 5.25e-11 ***
## factor(year)1997   -7.8840     1.2263  -6.429 1.86e-10 ***
## factor(year)1998   -8.2292     1.2263  -6.711 3.01e-11 ***
## factor(year)1999   -8.2442     1.2263  -6.723 2.77e-11 ***
## factor(year)2000   -8.6690     1.2263  -7.069 2.67e-12 ***
## factor(year)2001   -8.7019     1.2263  -7.096 2.21e-12 ***
## factor(year)2002   -8.4650     1.2263  -6.903 8.32e-12 ***
## factor(year)2003   -8.7310     1.2263  -7.120 1.88e-12 ***
## factor(year)2004   -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
```

F-statistic is 7.164 with p-value significantly below threshold level. Using year as explanatory is significant at 95% level. This shows that total fatalities is decreasing over time and that the year is sufficient to explain our outcome at a statistically significant level. Of course, the coefficients on

each year (relative to intercept/base year 1980) are identical to the averages of the rows for each year since no other variables are included in the model. Therefore the only additional insight from this model relative to the average by year is the standard errors:

```
average.per.year = byYear.mean$totfatrte
coefficients.vector = as.vector(fit.lm$coefficients)
#first coefficient is 1980, the rest are all relative to 1980
coefficients.relative.to.base = c(coefficients.vector[1], coefficients.vector[1] + coefficients
#we see that they are all the same as the average per year
round(coefficients.relative.to.base - average.per.year, 2)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Driving became safer over time if you use fatality rate as a measure/proxy for safety, however it is unclear if injuries, property damages, or other measures of safety trended similarly to the fatality rates in this data set.

Question 3

Variables bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl are supposed to be binary variables. But due to the fact that some states implemented the law in middle of year, some of the these variables have values between 0 and 1. For correct modeling of binary variables, we need all values to be 0 or 1, for approximation, we will round the values to be 0 or 1. Our model:

$$totfatrte_{it} = \beta_0 + \delta_{81}d81 + ... + \beta_1 bac08_{it} + \beta_2 bac10_{it} + \beta_3 perse_{it} + \beta_4 sbprim_{it} + ... + u_{it}$$

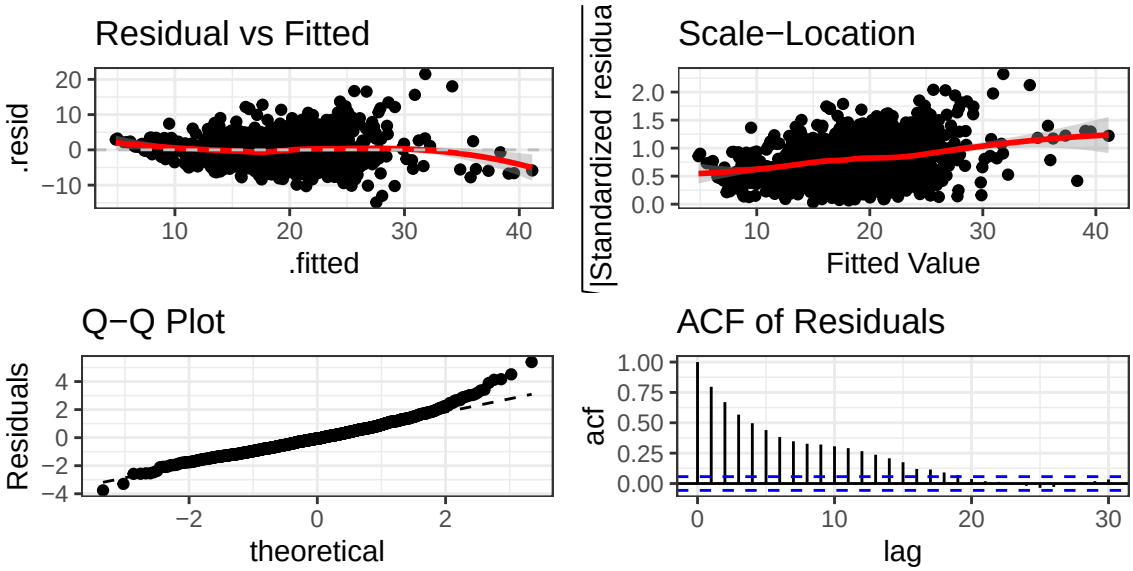
```
# round up binary variables
data.round <- data;data.round$bac08<-factor(round(data$bac08), levels=c(0,1));
data.round$bac10<-factor(round(data$bac10), levels=c(0,1));
data.round$perse<-factor(round(data$perse), levels=c(0,1));
data.round$sbprim<-factor(round(data$sbprim),levels=c(0,1));
data.round$sbsecon<-factor(round(data$sbsecon),levels=c(0,1))
data.round$sl70plus<-factor(round(data$sl70plus), levels=c(0,1));
data.round$gdl<-factor(round(data$gdl), levels=c(0,1))
# fit lm
fit.lm2 <- lm(totfatrte ~ factor(year)+bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24
summary(fit.lm2)
```

```
##
## Call:
## lm(formula = totfatrte ~ factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehicmilespc,
##     data = data.round)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8962  -2.7265  -0.3033   2.3323  21.5064
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.826e+00  2.478e+00  -1.141 0.254236
## factor(year)1981 -2.184e+00  8.290e-01  -2.634 0.008539 **
## factor(year)1982 -6.657e+00  8.547e-01  -7.789 1.49e-14 ***
## factor(year)1983 -7.589e+00  8.671e-01  -8.752  < 2e-16 ***
## factor(year)1984 -5.974e+00  8.730e-01  -6.843 1.25e-11 ***
## factor(year)1985 -6.603e+00  8.915e-01  -7.407 2.47e-13 ***
## factor(year)1986 -5.947e+00  9.290e-01  -6.401 2.23e-10 ***
## factor(year)1987 -6.459e+00  9.656e-01  -6.689 3.48e-11 ***
## factor(year)1988 -6.691e+00  1.013e+00  -6.607 5.97e-11 ***
## factor(year)1989 -8.159e+00  1.052e+00  -7.757 1.89e-14 ***
## factor(year)1990 -9.060e+00  1.076e+00  -8.421  < 2e-16 ***
## factor(year)1991 -1.121e+01  1.099e+00 -10.194  < 2e-16 ***
## factor(year)1992 -1.300e+01  1.121e+00 -11.591  < 2e-16 ***
## factor(year)1993 -1.288e+01  1.134e+00 -11.358  < 2e-16 ***
## factor(year)1994 -1.253e+01  1.154e+00 -10.855  < 2e-16 ***
## factor(year)1995 -1.203e+01  1.183e+00 -10.176  < 2e-16 ***
## factor(year)1996 -1.403e+01  1.224e+00 -11.459  < 2e-16 ***
## factor(year)1997 -1.430e+01  1.242e+00 -11.517  < 2e-16 ***
## factor(year)1998 -1.512e+01  1.262e+00 -11.978  < 2e-16 ***
## factor(year)1999 -1.518e+01  1.276e+00 -11.900  < 2e-16 ***
## factor(year)2000 -1.554e+01  1.296e+00 -11.996  < 2e-16 ***
## factor(year)2001 -1.645e+01  1.316e+00 -12.500  < 2e-16 ***
## factor(year)2002 -1.703e+01  1.331e+00 -12.798  < 2e-16 ***
## factor(year)2003 -1.742e+01  1.336e+00 -13.033  < 2e-16 ***
## factor(year)2004 -1.698e+01  1.369e+00 -12.399  < 2e-16 ***
## bac081           -2.194e+00  4.891e-01  -4.487 7.94e-06 ***
## bac101           -1.238e+00  3.616e-01  -3.423 0.000641 ***
## perse1           -6.499e-01  2.943e-01  -2.208 0.027433 *
## sbprim1          -9.420e-02  4.910e-01  -0.192 0.847868
## sbsecon1          6.430e-02  4.299e-01   0.150 0.881124
## sl70plus1         3.239e+00  4.352e-01   7.443 1.91e-13 ***
## gdl1             -3.476e-01  5.101e-01  -0.682 0.495682
## perc14_24         1.401e-01  1.229e-01   1.140 0.254611
## unem              7.675e-01  7.796e-02   9.844  < 2e-16 ***
## vehicmilespc      2.927e-03  9.485e-05  30.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.052 on 1165 degrees of freedom
## Multiple R-squared:  0.6064, Adjusted R-squared:  0.595
## F-statistic:  52.8 on 34 and 1165 DF,  p-value: < 2.2e-16
```

```
# diagnostic plots
p1<-ggplot(fit.lm2, aes(.fitted, .resid))+geom_point()+geom_smooth(method="loess", col="red", 
p2<-ggplot(fit.lm2, aes(.fitted, sqrt(abs(.stdresid))))+geom_point(na.rm=TRUE)+geom_smooth(for
p3<-ggplot(fit.lm2, aes(sample=.stdresid)) + stat_qq() + stat_qq_line(col="black", linetype="da
```

```
list.acf<-acf(fit.lm2$residuals, plot=FALSE);bacf.df<-with(list.acf, data.frame(lag,acf));N<-as
p4<-ggplot(bacf.df, aes(x=lag, y = acf))+geom_hline(aes(yintercept = 0))+geom_segment(mapping =
grid.arrange(p1,p2,p3,p4,nrow=2)
```



Variables *bac08* and *bac10* are binary indicator variables, indicating if a state had law of blood alcohol content of level 0.08% and 0.10% repectively. From mean plot of variables *bac08* and *bac10* in EDA, we see that majority of state start with no law on blood alcohol content, and then implementing a 0.10% limit, and then a more strict limit of 0.08%. Coefficient of *bac10* can be interpreted as, states with blood alcohol content limit 0.10% law have 1.238 less fatalities per 100,000 population. Coefficient of *bac08* can be interpreted as, states with blood alcohol content limit 0.08% law have 2.194 less fatalities per 100,000 population.

Variable *perse* (per se law) has p-value of 0.027433 in pooled OLS result. This variable is statistically significant at 95% level. It shows that there is empirical evidence that per se law is negatively correlated with fatalities.

Variable *sbprim* (primary seat belt law) has p-value of 0.847868 in pooled OLS result. This variable is not statistically significant at 95% level. It shows that there is no empirical evidence that primary seat belt law is correlated with fatality rate.

Even though the perc_14_24 had a strong correlation with fatality rate as seen in the corrplot, it did not turn out signifianct in the pooled regression. However, speed limit of 70-plus did turn out significiant. It fact it indicates that controlling for all other factors, states with speed limit 70 or above (incl. no limit) have 3.24 more fatalities per 100,000 population.

One thing to note is that, from regression diagnostic, we observed heteroskedasticity on residuals from scale-location plot and serial correlations on residuals from ACF graph. Serial correlations on residuals suggest there are state level unobserved fixed effects. Serial correlations and heteroskedasticity on residuals suggest the test statistics in pooled OLS result are not valid.

Question 4

Adding a fixed effect for state:

```r
data.panel = pdata.frame(data.round, index=c("state", "year"))
fit.plm.fe <-plm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicm
                 data=data.panel, model='within')
summary(fit.plm.fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon +
##     sl70plus + gdl + perc14_24 + unem + vehicmilespc, data = data.panel,
##     model = "within")
##
## Balanced Panel: n = 48, T = 25, N = 1200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -7.196355 -1.199164 -0.068262  1.137700 14.554645
##
## Coefficients:
##                 Estimate  Std. Error  t-value  Pr(>|t|)
## bac081       -1.54934878  0.33484339  -4.6271 4.132e-06 ***
## bac101       -1.15290142  0.23139549  -4.9824 7.250e-07 ***
## perse1       -1.40105536  0.23799390  -5.8869 5.166e-09 ***
## sbprim1      -1.86938834  0.34668462  -5.3922 8.454e-08 ***
## sbsecon1     -0.88032830  0.24914282  -3.5334 0.0004266 ***
## sl70plus1    -1.13047368  0.23850465  -4.7398 2.408e-06 ***
## gdl1         -0.58719959  0.22493208  -2.6106 0.0091577 **
## perc14_24     0.97632522  0.07069974  13.8095 < 2.2e-16 ***
## unem         -0.59813653  0.05100886 -11.7261 < 2.2e-16 ***
## vehicmilespc  0.00024665  0.00010162   2.4271 0.0153745 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12134
## Residual Sum of Squares: 5571.9
## R-Squared:      0.54081
## Adj. R-Squared: 0.51789
## F-statistic: 134.498 on 10 and 1142 DF, p-value: < 2.22e-16
```

In fixed effect model, the coefficient of *bac10* is similar to pooled OLS and and the coefficient of *bac08* is smaller in absolute value. *perse* is highly statistically significant in fixed effects model but it was marginally statistically significant in pooled OLS. *sbprim* is highly statistically significant in fixed effects model but it was not statistically significant in pooled OLS.

Result from fixed effect model is more reliable. In pooled OLS, we have to assume no state level unobserved fixed effects, otherwise test statistics are not valid. While in fixed effects model, we are allowed to have unobserved fixed effects present in population model and this fixed effect is allowed to be correlated with explanatory variables. In ACF graph of pooled OLS residuals, we observe

serial correlations and this suggests the present of unobserved effect. Therefore assumptions of OLS are not met and pooled OLS result is not reliable. Fixed effect model is the preferred choice.

Question 5

```
fit.plm.re <- plm(totfatrte~bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24+unem+vehicr
                  data=data.panel, model='random')
phtest(fit.plm.fe, fit.plm.re)
```

```
##
##  Hausman Test
##
## data:  totfatrte ~ bac08 + bac10 + perse + sbprim + sbsecon + sl70plus +  ...
## chisq = 72875, df = 10, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

P-value is smaller than 0.05 in Hausman test, we can reject null hypothesis that random effect model is preferred. Fixed Effect model should be chosen for our analysis.

This comes as no surprise. We have data from all 48 states, presumptively including any state for which our model may be pertinent. As such, there is no need for a random effect per state. Our data set has balanced data for each.

Regarding model specifications, the random effects model does not allow for arbitrary correlations between unobserved fixed effects and explanatory variables. There could be state level unobserved effects, e.g. culture, geographical features, presence of industries that rely on natural resources. Blood alcohol limit and unemployment are very likely to be correlated with these examples of state level unobserved effects.

Question 6

Increase miles driven per capita by 1000, the expect total fatalities per 100,000 population increase by 0.00024665 * 1000 = 0.24665, holding all other variables constant.

As noted in the EDA, it would be reckless to make such an assertion in the hypothetical situation that such an intervention could be carried out. Also worth noting is that EDA shows *vehicmilespc* trends upwards over the timespan, and *totfatrte* trends downward, suggesting a negative correlation between the two. However after accounting for the effects of the other explanatory variables, the relationship appears to be positive. This isn't hard to believe because more miles offers more chances for a traffic fatality.

Question 7

In this scenario, the estimators are not as efficient. Statistical inferences are not valid. If only heteroskedasticity present but not serial correlation, we can use heteroskedastic robust standard errors for statistical inference.

If unobserved effect is uncorrelated with all explanatory variables, the estimators would be consistent, otherwise estimators would not be consistent.