lab5_samir

Samir Datta

December 10, 2017

```r
library(ggplot2)
library(reshape2)
load("driving.Rdata")
drivedata <- data
```

bac, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc

Structure of the data

```r
head(drivedata)
```

```
##   year state sl55 sl65 sl70 sl75 slnone seatbelt minage zerotol gdl bac10
## 1 1980     1    1    0    0    0      0        0     18       0   0     1
## 2 1981     1    1    0    0    0      0        0     18       0   0     1
## 3 1982     1    1    0    0    0      0        0     18       0   0     1
## 4 1983     1    1    0    0    0      0        0     18       0   0     1
## 5 1984     1    1    0    0    0      0        0     18       0   0     1
## 6 1985     1    1    0    0    0      0        0     20       0   0     1
##   bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm wkndfatpvm
## 1     0     0    940     422     236      3.20      1.437      0.803
## 2     0     0    933     434     248      3.35      1.558      0.890
## 3     0     0    839     376     224      2.81      1.259      0.750
## 4     0     0    930     397     223      3.00      1.281      0.719
## 5     0     0    932     421     237      2.83      1.278      0.720
## 6     0     0    882     358     224      2.51      1.019      0.637
##   statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem perc14_24
## 1  3893888     24.14      10.84       6.06   29.37500  8.8      18.9
## 2  3918520     24.07      11.08       6.33   27.85200 10.7      18.7
## 3  3925218     21.37       9.58       5.71   29.85765 14.4      18.4
## 4  3934109     23.64      10.09       5.67   31.00000 13.7      18.0
## 5  3951834     23.58      10.65       6.00   32.93286 11.1      17.6
## 6  3972527     22.20       9.01       5.64   35.13944  8.9      17.3
##   sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88 d89 d90 d91
## 1        0      0       0   1   0   0   0   0   0   0   0   0   0   0   0
## 2        0      0       0   0   1   0   0   0   0   0   0   0   0   0   0
## 3        0      0       0   0   0   1   0   0   0   0   0   0   0   0   0
## 4        0      0       0   0   0   0   1   0   0   0   0   0   0   0   0
## 5        0      0       0   0   0   0   0   1   0   0   0   0   0   0   0
## 6        0      0       0   0   0   0   0   0   1   0   0   0   0   0   0
##   d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04 vehicmilespc
## 1   0   0   0   0   0   0   0   0   0   0   0   0   0     7543.874
## 2   0   0   0   0   0   0   0   0   0   0   0   0   0     7107.785
## 3   0   0   0   0   0   0   0   0   0   0   0   0   0     7606.622
## 4   0   0   0   0   0   0   0   0   0   0   0   0   0     7879.802
## 5   0   0   0   0   0   0   0   0   0   0   0   0   0     8333.562
## 6   0   0   0   0   0   0   0   0   0   0   0   0   0     8845.614
```

```
unique(drivedata$year)
```

```
##  [1] 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993
## [15] 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
```

```
unique(drivedata$state)
```

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

The data is in long form. There are 25 rows per state, each representing a different year from 1980 to 2005. The data description data frame tells us that the "state" variable is a number that simply represents the 50 states in alphabetical order.

Some variables like the BAC and seatbelt law variable have more than two levels represented by multiple dummy coded variables. For example, BAC has three levels - no BAC law, BAC limit = .08, and BAC limit = .10. These three levels are coded in the two variables bac08 and bac10. Each of the years is also dummy coded in its own column.

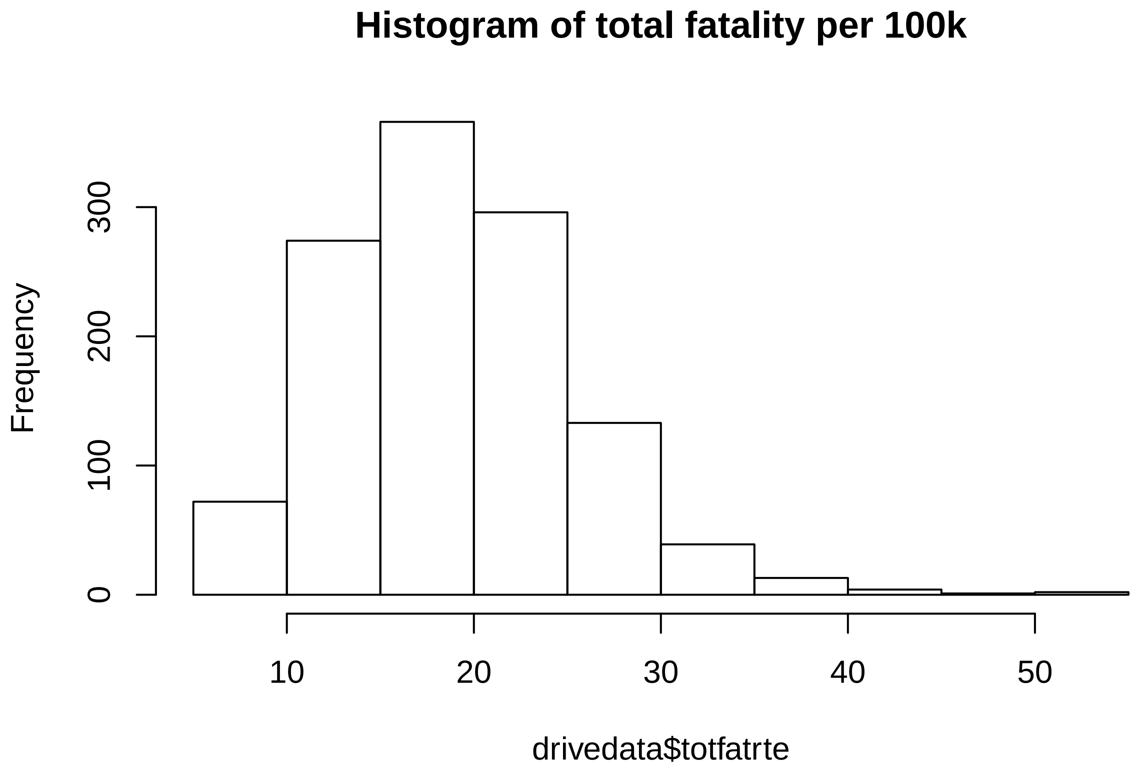
An odd fluke in the data noticed above that there is no state number 2. The number is suppose to represent the states in alphabetical order, but 2 is skipped and the maximum number is 51 instead of 50. This will not affect our modeling in any way since the number is effectively a categorical variable and the values don't matter.

EDA

Total fatality rate

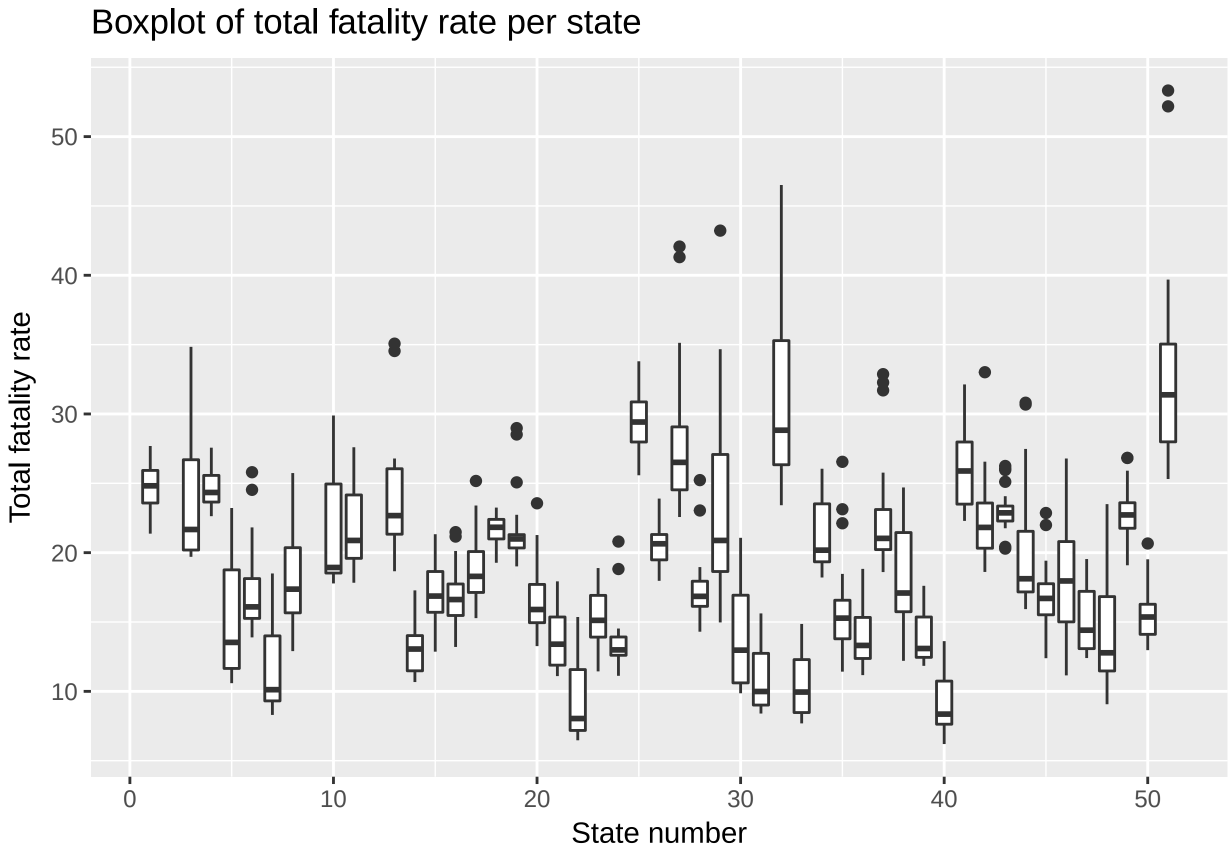
The primary variable of interest is "totfatrte", the total fatalities per 100,000 population.

```
hist(drivedata$totfatrte, main="Histogram of total fatality per 100k")
```



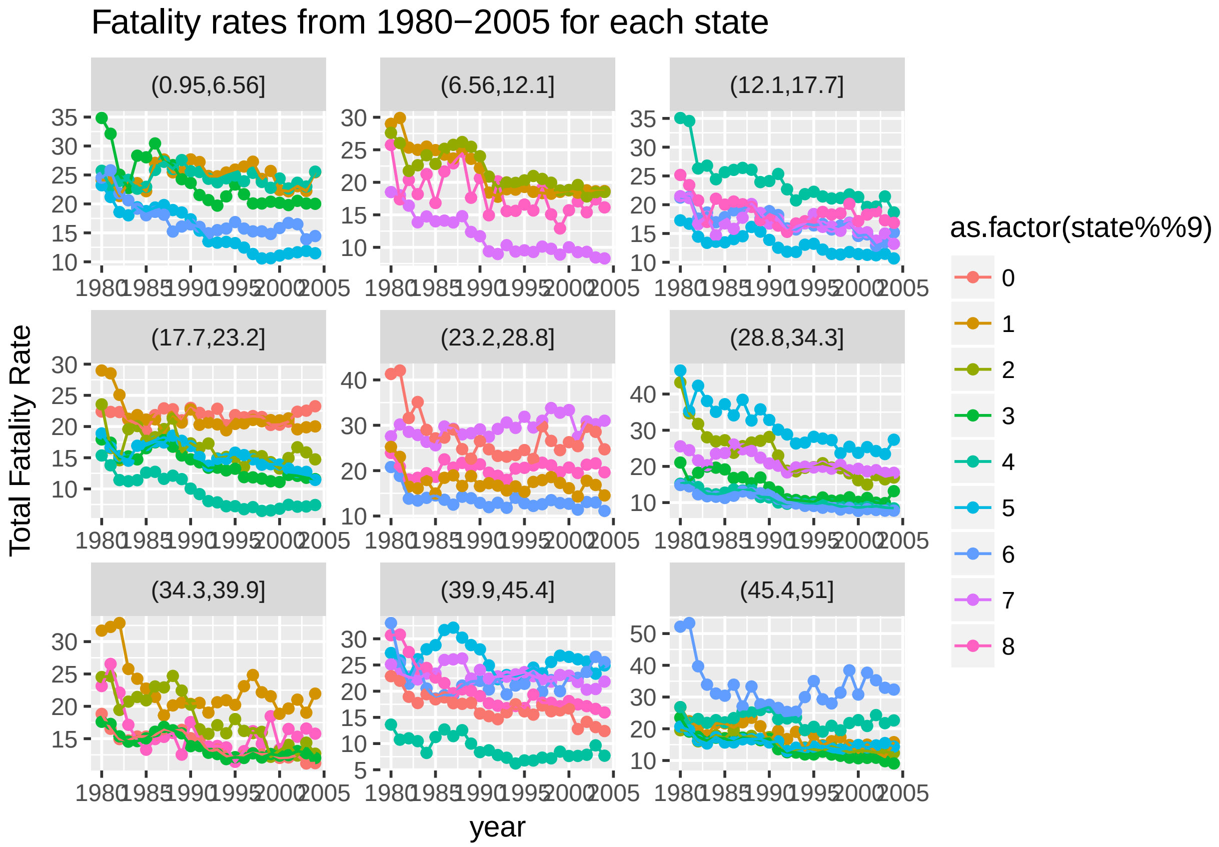
The data has a slight positive skew. Note that this histogram is of all values of total fatality rate, across all states and years.

```
ggplot(drivedata, aes(x=state, y=totfatrte, group=state)) + geom_boxplot()+
ylab("Total fatality rate")+xlab("State number")+ggtitle("Boxplot of total fatality rate per state")
```



Clearly, each state has different fatality rates, and some states simply have higher or lower rates regardless of the year or other variables.

```
drivedata$statebucket <- cut(drivedata$state, 9)

ggplot(drivedata, aes(x=year, y=totfatrte, group=as.factor(state), color=as.factor(state%%9)))+geom_poi
  geom_line()+facet_wrap(~as.factor(statebucket), scales="free")+
  ylab("Total Fatality Rate")+guides(fill=F)+
  ggtitle("Fatality rates from 1980-2005 for each state")
```



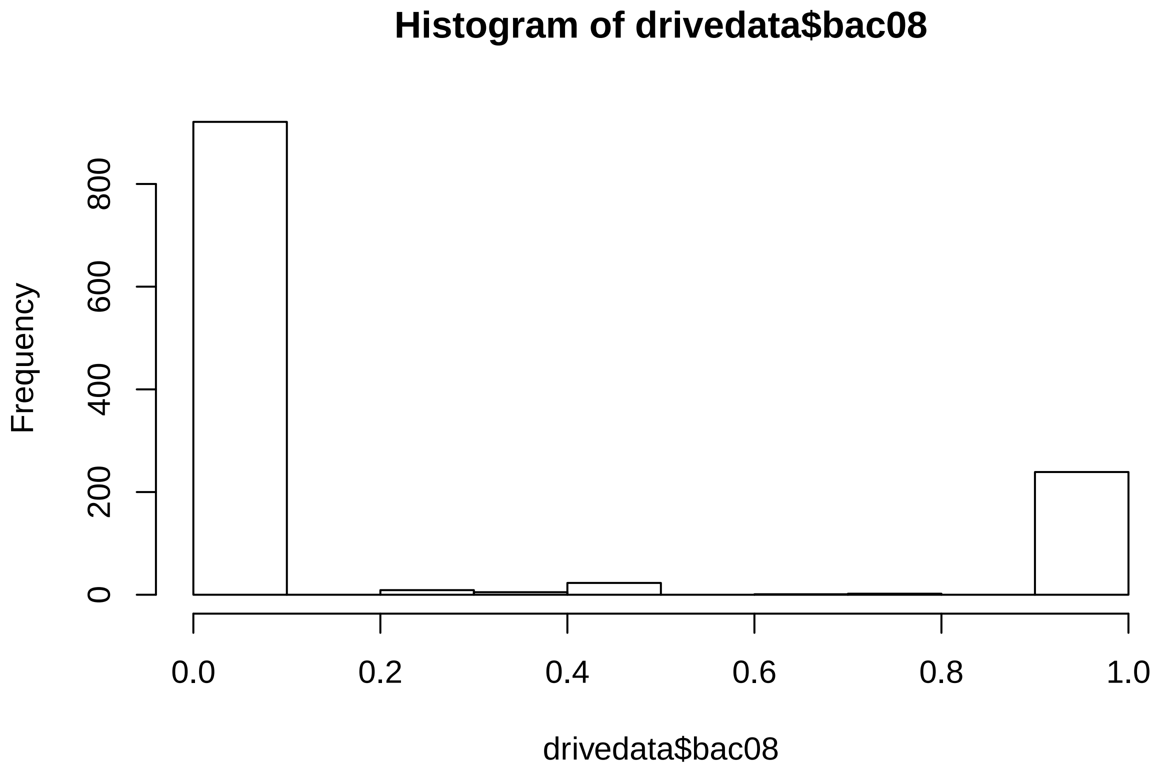
The above plot shows the fatality rates from 1980-205. Each set of dots connected with lines represents the data for a different state. 9 separate graphs are shown simply for the sake of avoiding clutter - the states have been binned by the number in the dataset (their alphabetical order).

This graph is simply to get an idea of the nature of year-to-year changes in fatlity rate. It seems for most states the rates steadily go down, but there are some interesting exceptions. It also seems that for a lot of states, the rates for the first one or two years is much higher, after which the rates drastically drop. The overall decrease in fatality rates over time is likely due to factors such as laws enacted.

Laws

bac, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc

```
hist(drivedata$bac08)
```



Many of the explanatory variables of interest (bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl) appear to be binary - 0 if the law was not in place, 1 if it was. As shown in the example histogram above, there are a very small number of fractional values.

```r
head(drivedata[drivedata$year>1990,c("year", "bac08")], 10)
```

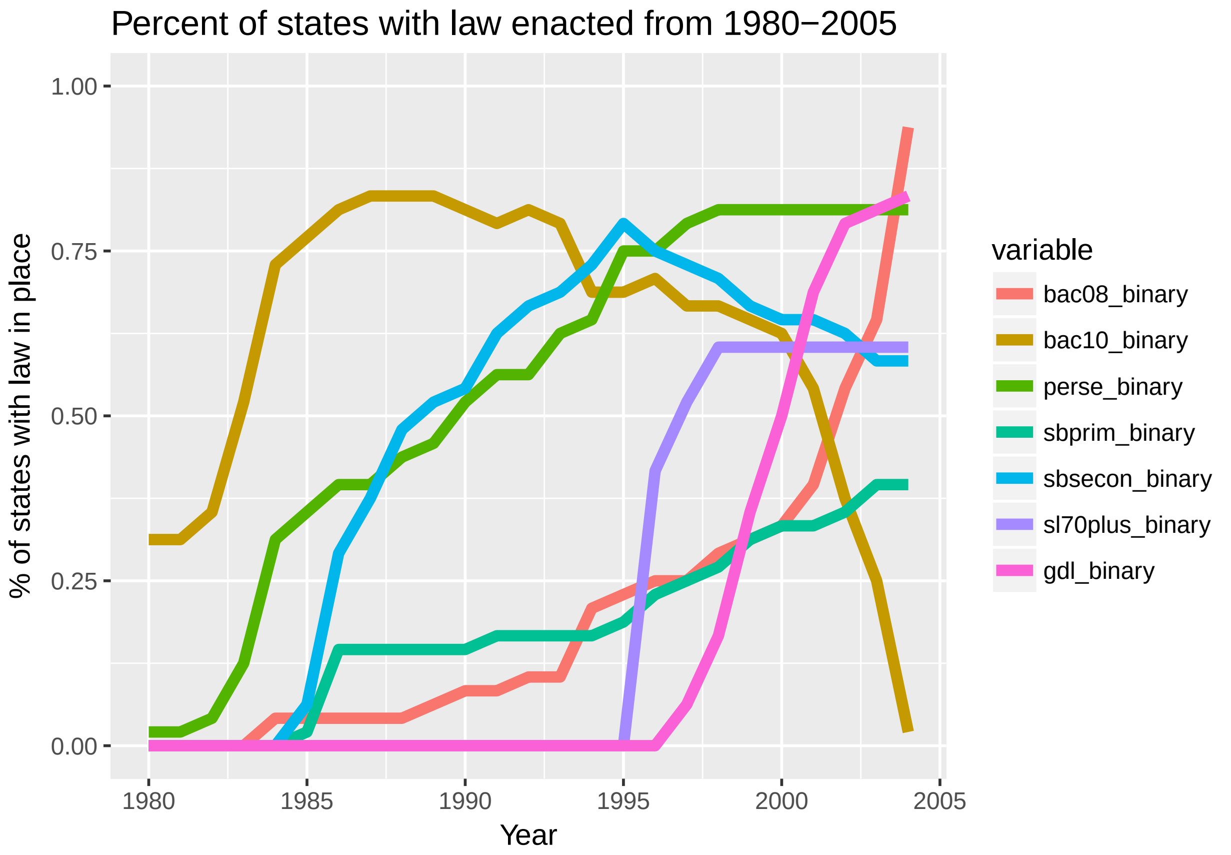
```
##    year bac08
## 12 1991 0.000
## 13 1992 0.000
## 14 1993 0.000
## 15 1994 0.000
## 16 1995 0.417
## 17 1996 1.000
## 18 1997 1.000
## 19 1998 1.000
## 20 1999 1.000
## 21 2000 1.000
```

As shown in the example section of the data above, these fractional values occur between stretches of 0 and 1. There is no additional clarifying information in the dataset or data descriptions. As such, we are moving forward with the assumption that fractional values occur when the law was enacted part way through the year, and the value represents how much of the year the law was in place for.

Keeping these fractional values in our data will force the model to treat them as continuous variables, which is very problematic. We are not interested in the effect of a law if it is present for part of the year; rather, we want to know the binary effect of a law being in place. It is also very problematic to use a continuous variable in a linear model with such an odd and non-normal distribution, given the sparsity of the fractional values. As such, we are moving forward using the rounded values which will turn the fractional values into 0 or 1.

```
#code to binarize all law variables
drivedata$bac08_binary <- ifelse(round(drivedata$bac08)==1, 1, 0)
drivedata$bac10_binary <- ifelse(round(drivedata$bac10)==1, 1, 0)
drivedata$perse_binary <- ifelse(round(drivedata$perse)==1, 1, 0)
drivedata$sbprim_binary <- ifelse(round(drivedata$sbprim)==1, 1, 0)
drivedata$sbsecon_binary <- ifelse(round(drivedata$sbsecon)==1, 1, 0)
drivedata$sl70plus_binary <- ifelse(round(drivedata$sl70plus)==1, 1, 0)
drivedata$gdl_binary <- ifelse(round(drivedata$gdl)==1, 1, 0)
```

```
year_law_agg <- with(drivedata, aggregate(cbind(bac08_binary, bac10_binary, perse_binary, sbprim_binary

year_law_melt <- melt(year_law_agg, id.vars="year")

ggplot(year_law_melt, aes(x=year, y=value, color=variable, group=variable))+
  geom_line(size=2)+ylab("% of states with law in place")+xlab("Year")+
  ggtitle("Percent of states with law enacted from 1980-2005")+
  ylim(c(0,1))
```

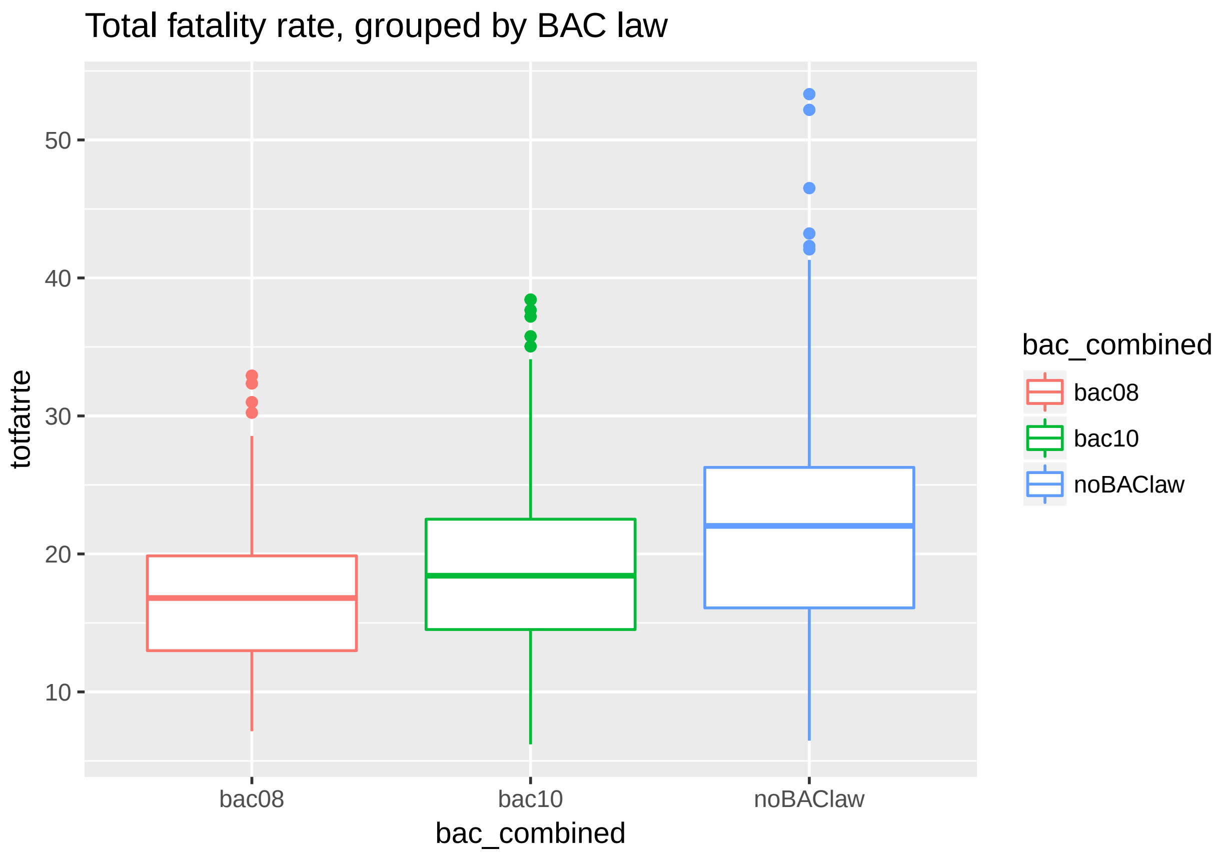


This graph shows the % of states that had a certain law in place in every year.

The blood alcohol content laws are interesting. Clearly in the early 2000s there was a big movement to decrease the legal BAC from .10 to .08 given the opposite directions of those lines. A similar, but weaker, pattern is shown with the seatbelt law being primary or secondary around 1995.

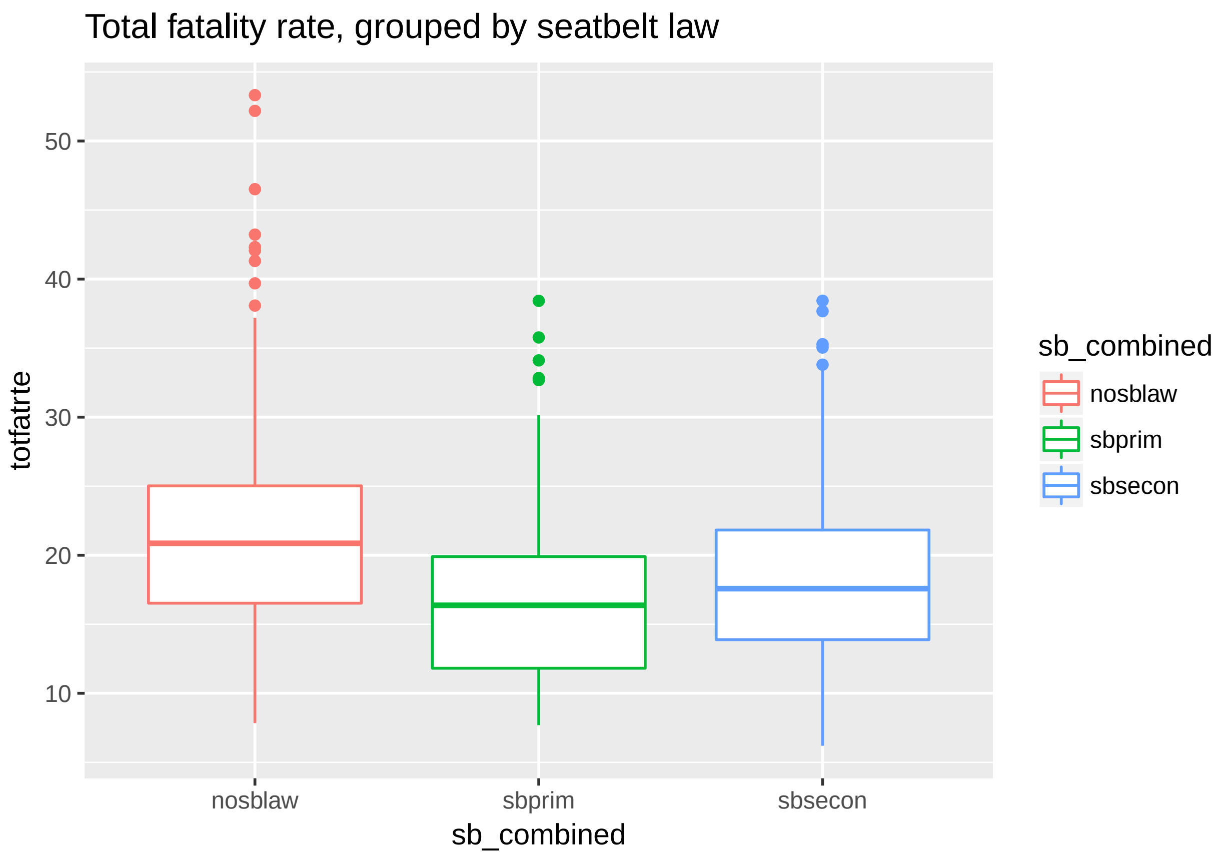
It is also interesting to note that both the graduated drivers license law and the speed limit being 70 or more changes did not happen at all until around 1995, and then fairly rapidly became more of a standard. Still, the speed limit change plateaus around 60%. It seems that about 60% of states adopted the speed limit change very quickly and the remaining states did not.

```
drivedata$bac_combined <- ifelse(drivedata$bac08_binary==1, "bac08",
                                 ifelse(drivedata$bac10_binary==1, "bac10", "noBAClaw"))
ggplot(drivedata, aes(x=bac_combined, y=totfatrte, color=bac_combined, group=bac_combined))+geom_boxplo
  ggtitle("Total fatality rate, grouped by BAC law")
```



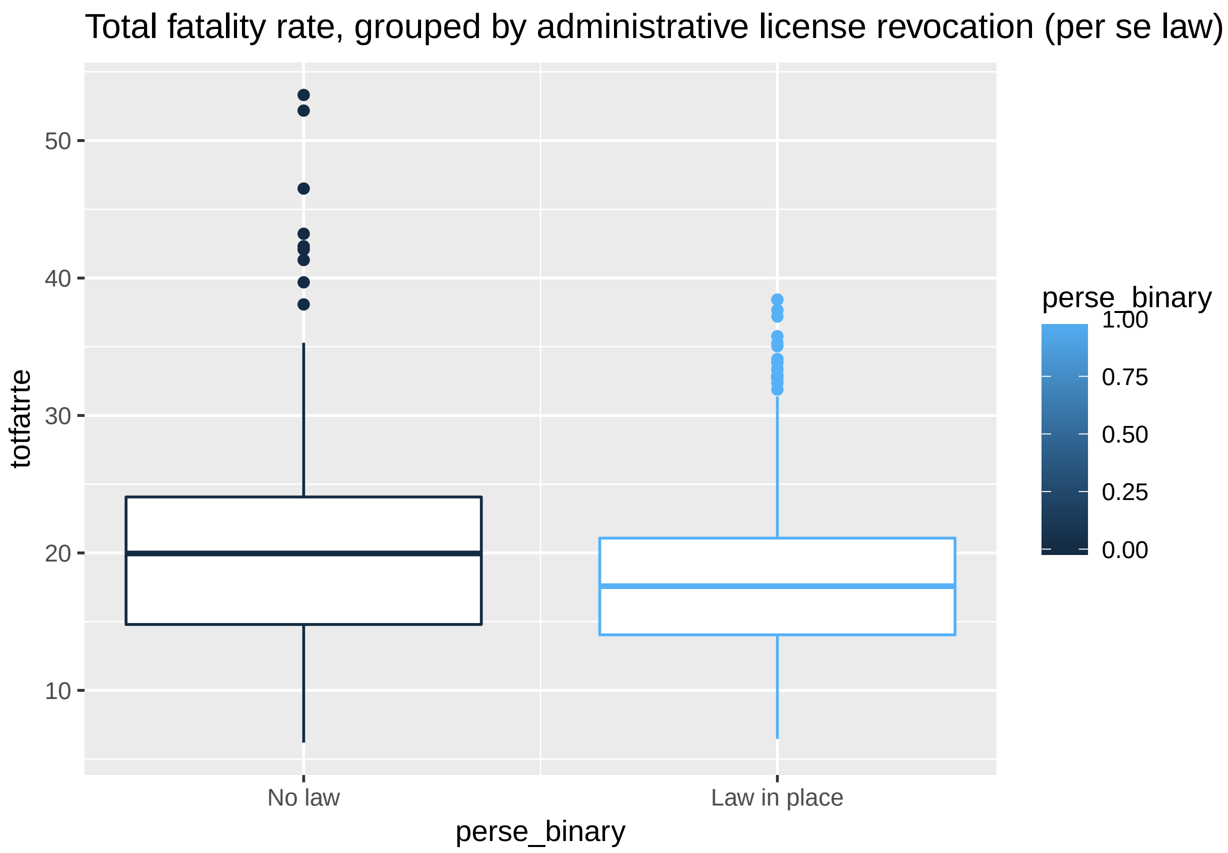
Predictably, in states/years where the BAC limit was .08, there were lower fatality rates than when the limit was .10. States/years where there was no BAC law in place had the highest rates on average.

```
drivedata$sb_combined <- ifelse(drivedata$sbprim_binary==1, "sbprim",
                                ifelse(drivedata$sbsecon_binary==1, "sbsecon", "nosblaw"))
ggplot(drivedata, aes(x=sb_combined, y=totfatrte, color=sb_combined, group=sb_combined))+geom_boxplot()
  ggtitle("Total fatality rate, grouped by seatbelt law")
```

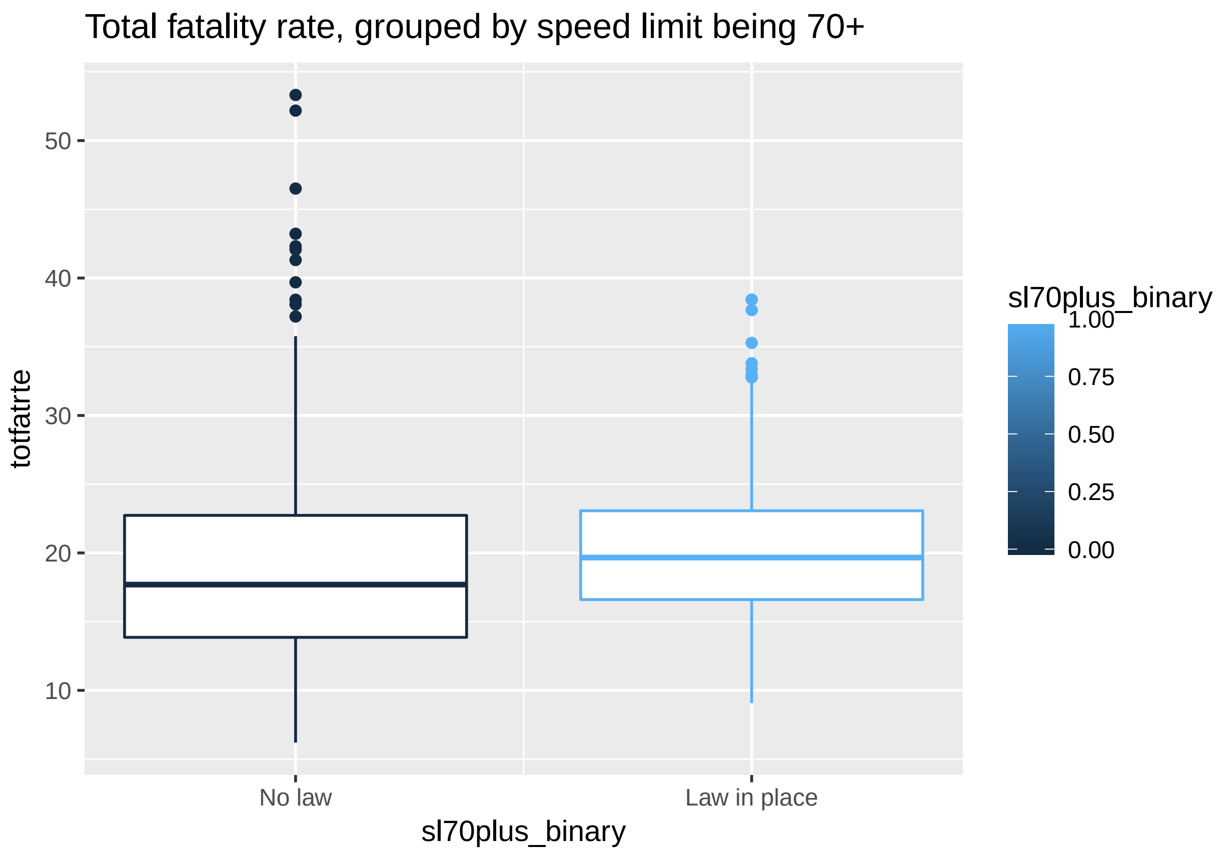


States/years without a seatbelt law have the highest fatality rate. States/years with the primary seatbelt law have lower rates than those with the secondary seatbelt law.

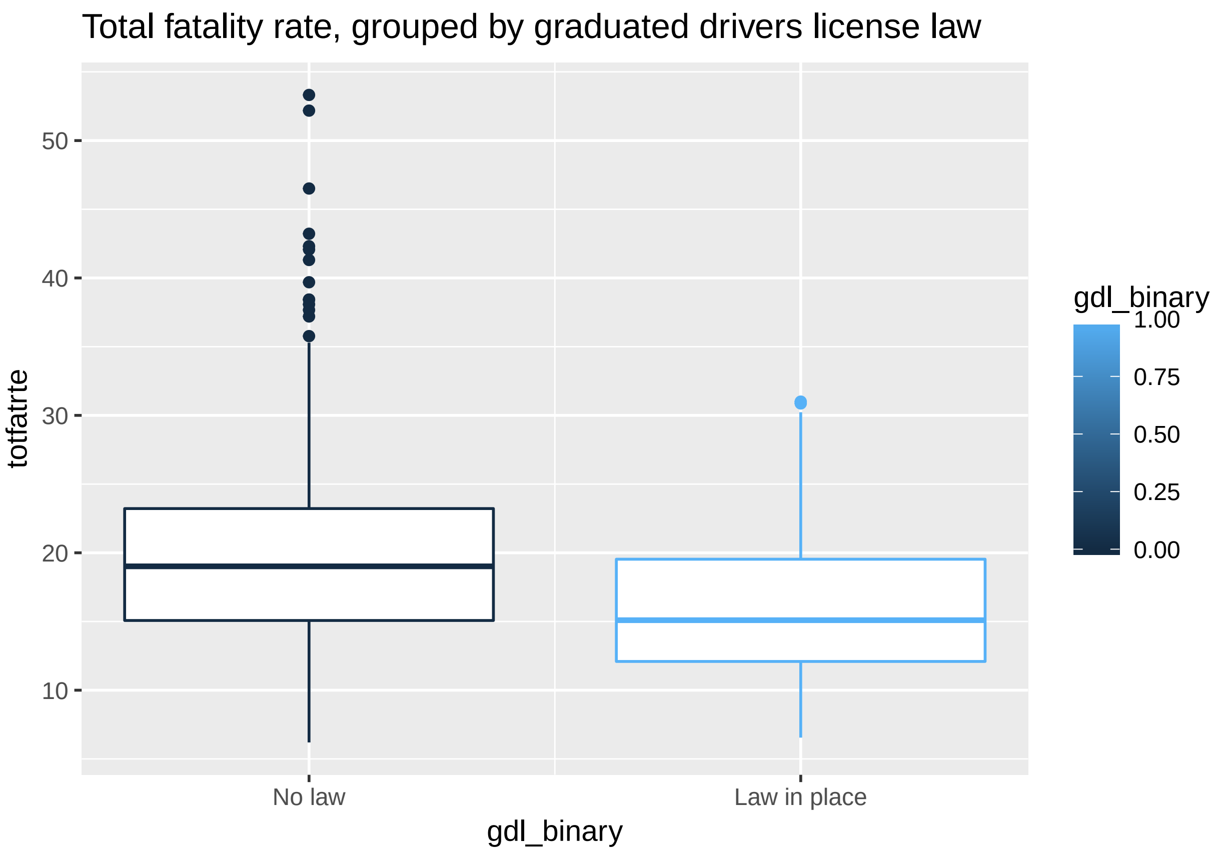
```
#bac, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc

ggplot(drivedata, aes(x=perse_binary, y=totfatrte, color=perse_binary, group=perse_binary))+geom_boxplo
  scale_x_continuous(breaks=c(0,1), labels=c("No law", "Law in place"))+
  ggtitle("Total fatality rate, grouped by administrative license revocation (per se law)")
```



```
ggplot(drivedata, aes(x=sl70plus_binary, y=totfatrte, color=sl70plus_binary, group=sl70plus_binary))+ge
  scale_x_continuous(breaks=c(0,1), labels=c("No law", "Law in place"))+
  ggtitle("Total fatality rate, grouped by speed limit being 70+")
```



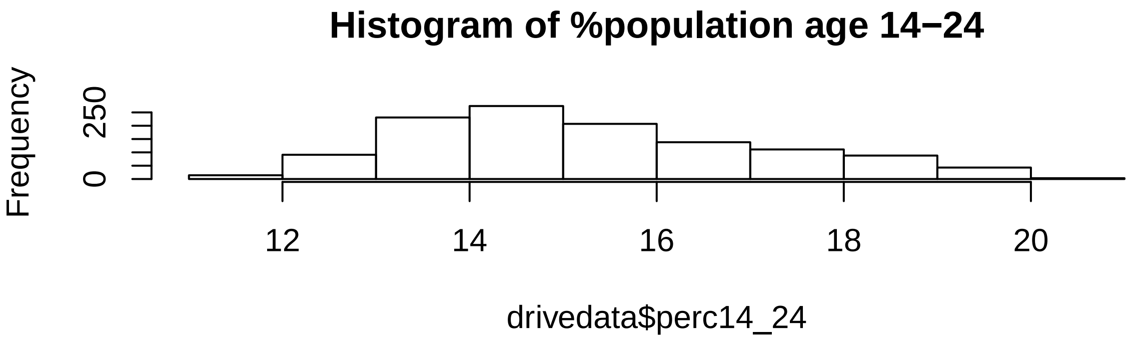
```r
ggplot(drivedata, aes(x=gdl_binary, y=totfatrte, color=gdl_binary, group=gdl_binary))+geom_boxplot()+
  scale_x_continuous(breaks=c(0,1), labels=c("No law", "Law in place"))+
  ggtitle("Total fatality rate, grouped by graduated drivers license law")
```

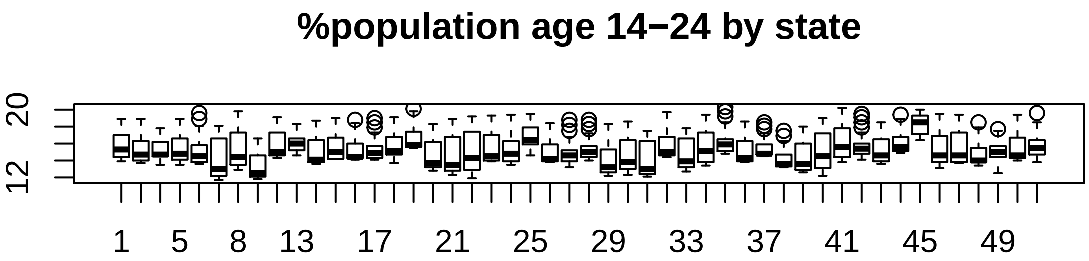


The above boxplots show that fatality rates are lower when the "per se" law and the graduated drivers license law are in place, and higher when the speed limit is above 70. At least based on this plot, it seems the speed limit has the weakest effect. Raising the speed limit would intuitively increase fatality rates due to faster and more reckless driving. However, based on the previous time plot, the speed limit change appeared to happen around the same time as other safe laws were being enacted more often. Including all of these variables into a model should clarify their individual, ceteris paribus effects.

Other explanatory variables

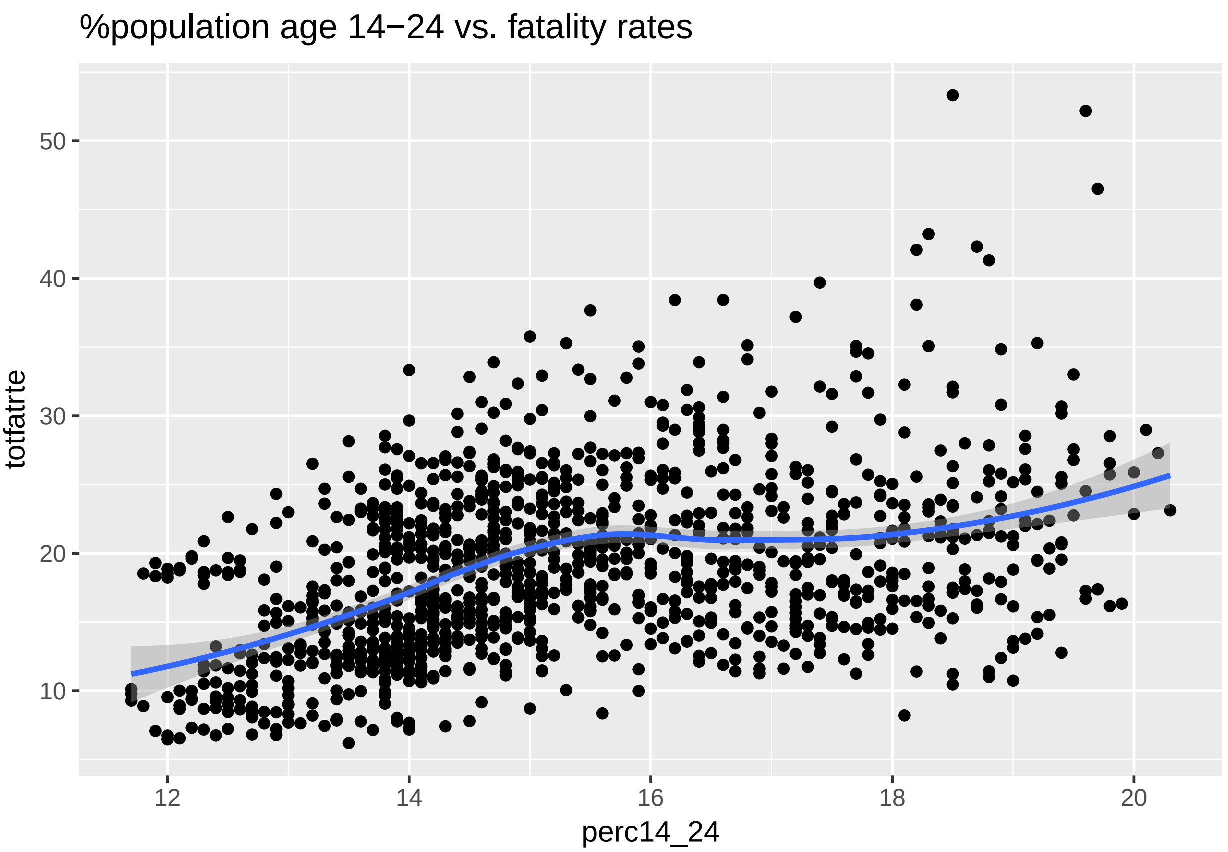
```
par(mfrow=c(2,1))
hist(drivedata$perc14_24, main="Histogram of %population age 14-24")
boxplot(perc14_24~state, data=drivedata, main="%population age 14-24 by state")
```





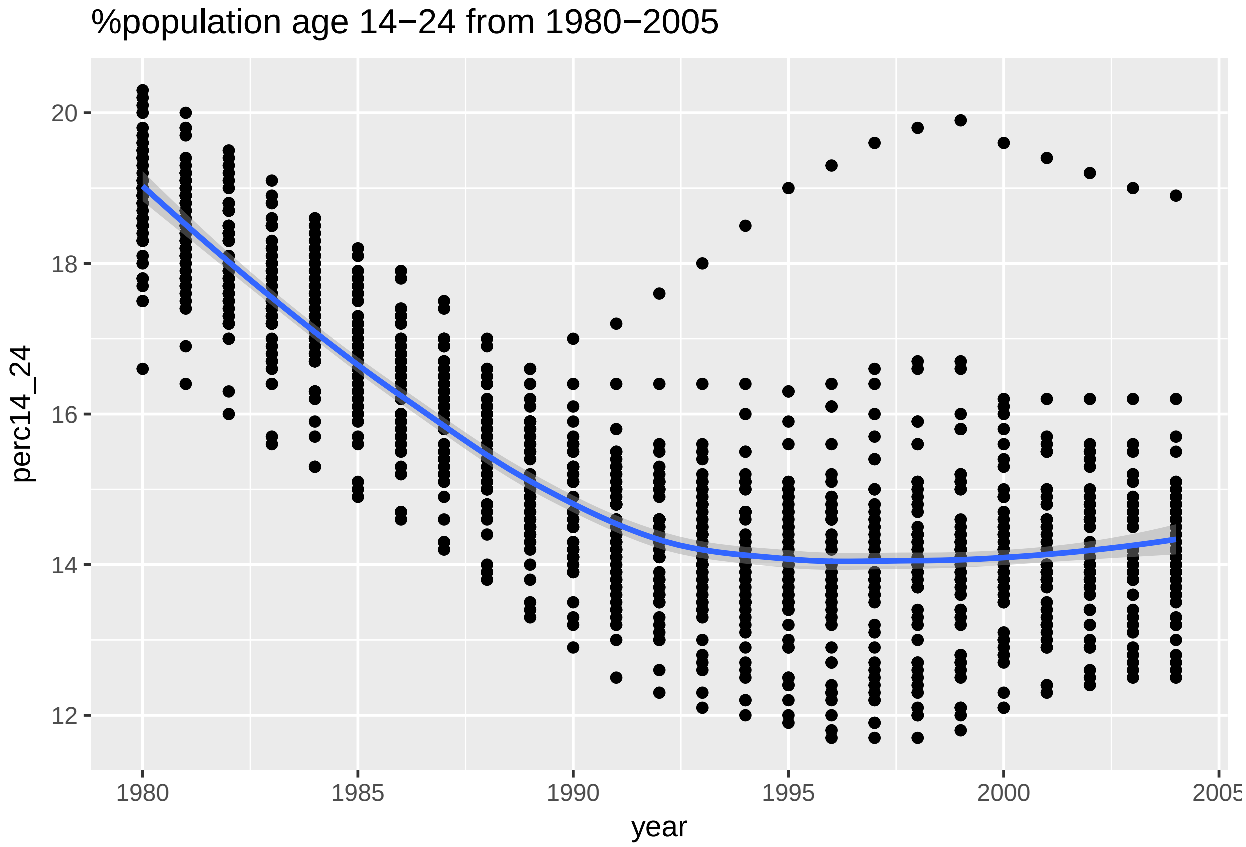
The percent of population across states/years appears to be a normally distributed variable. This variable doesn't appear to be especially different from state to state - there are no states that tend to have an especially high or low % of ages 14-24.

```r
ggplot(drivedata, aes(x=perc14_24, y=totfatrte))+
  geom_point()+geom_smooth(method="loess")+ggtitle("%population age 14-24 vs. fatality rates")
```



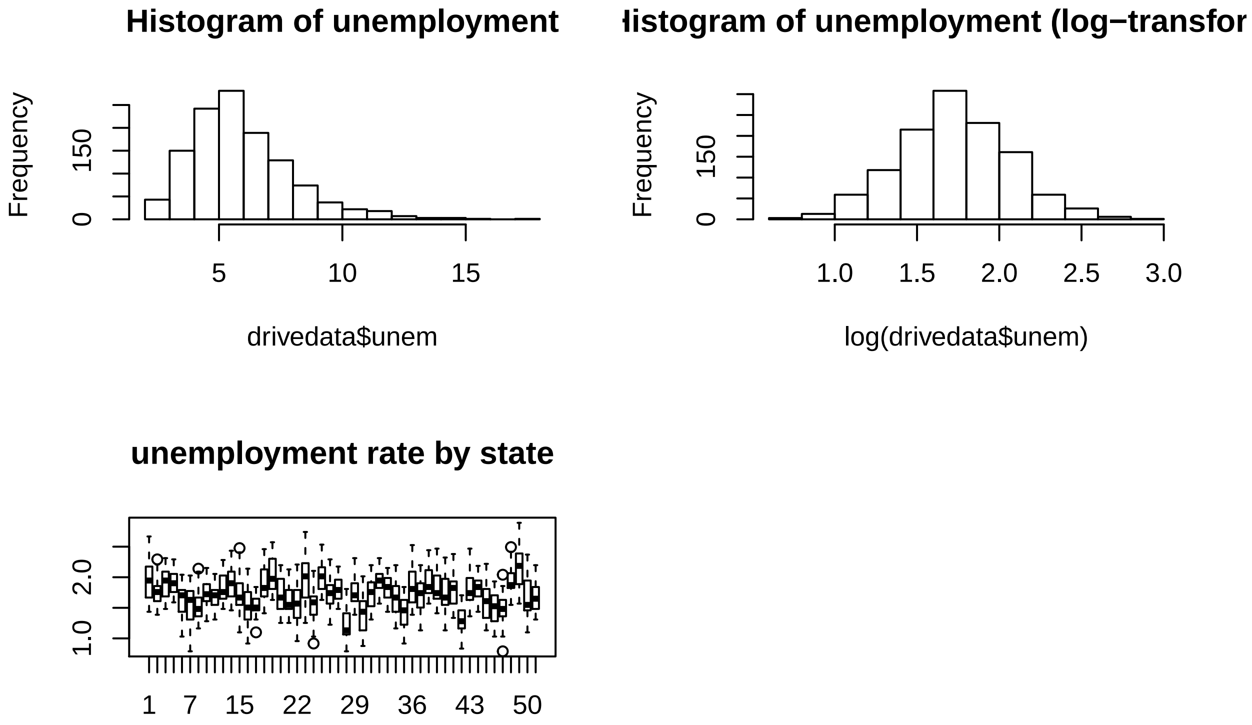
There appears to be a linear relationship where states/years with a higher % of the population between ages 14-24 have higher fatality rates. Perhaps younger drivers are less safe.

```
ggplot(drivedata, aes(x=year, y=perc14_24))+
  geom_point()+geom_smooth(method="loess")+ggtitle("%population age 14-24 from 1980-2005")
```



The % of the population between 14 and 24 very distinctly goes down from 1980 to about 1990, after which it flattens out to around 14%.

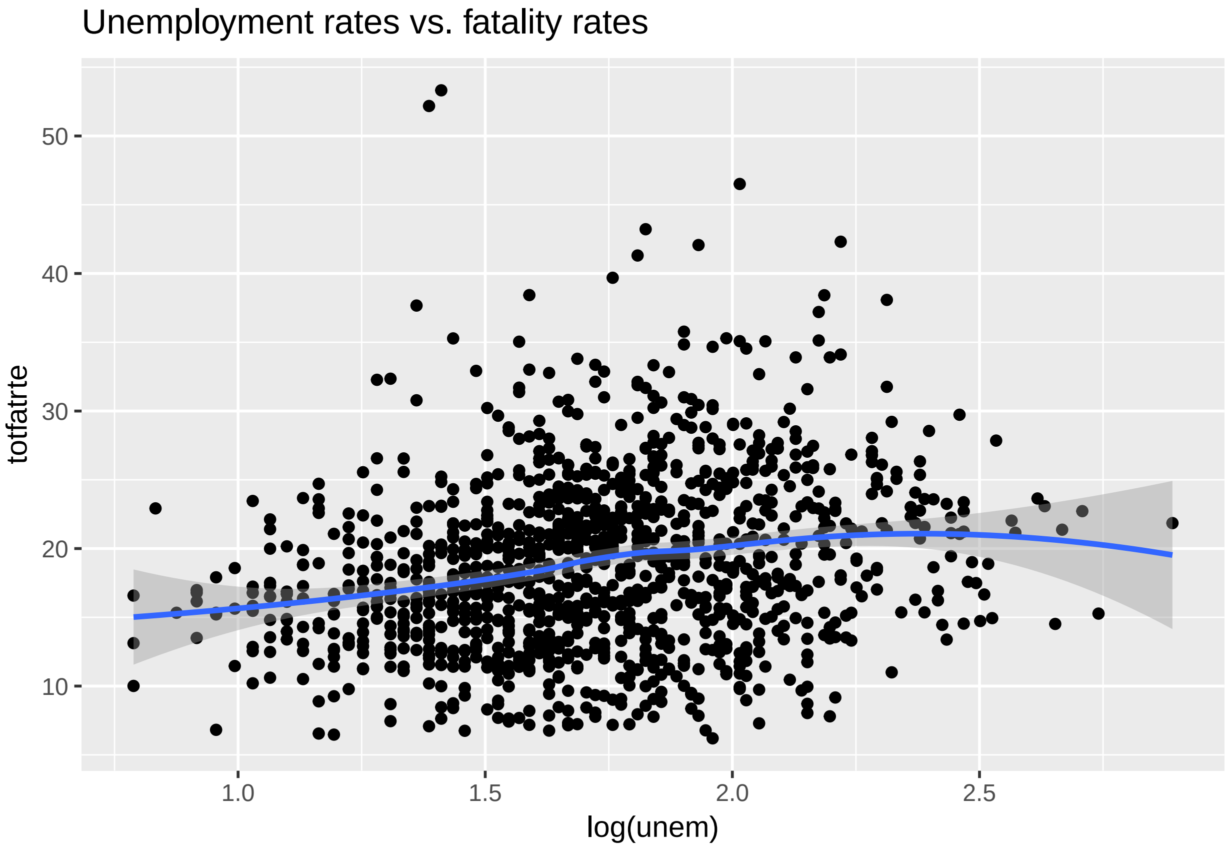
```
par(mfrow=c(2,2))
hist(drivedata$unem, main="Histogram of unemployment")
hist(log(drivedata$unem), main="Histogram of unemployment (log-transformed)")
boxplot(log(unem)~state, data=drivedata, main="unemployment rate by state")
```



The unemployment rate, which is a percent, has a strong positive skew. The normality of the distribution improves when applying a log transformation.

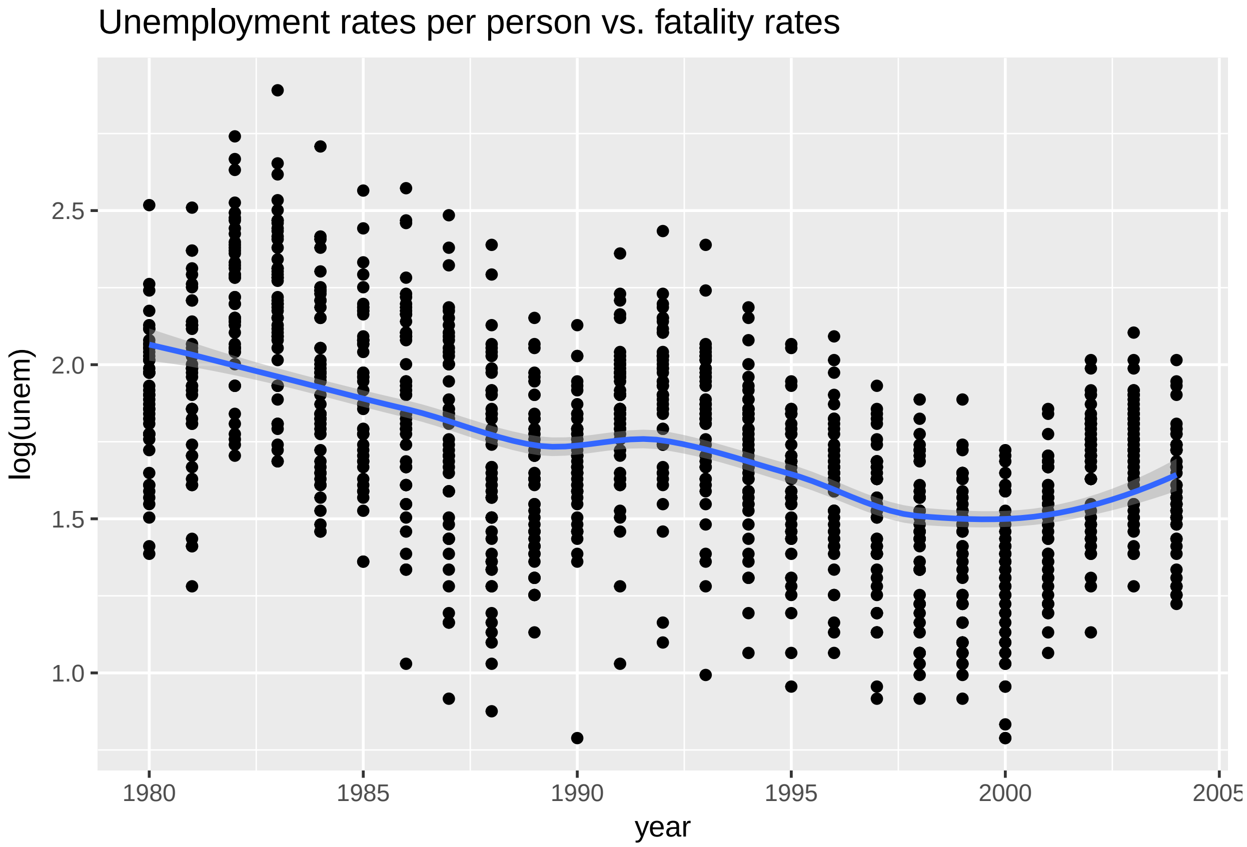
Some states appear to have distinctly higher or lower unemployment rates.

```
ggplot(drivedata, aes(x=log(unem), y=totfatrte))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Unemployment rates vs. fatality rates")
```



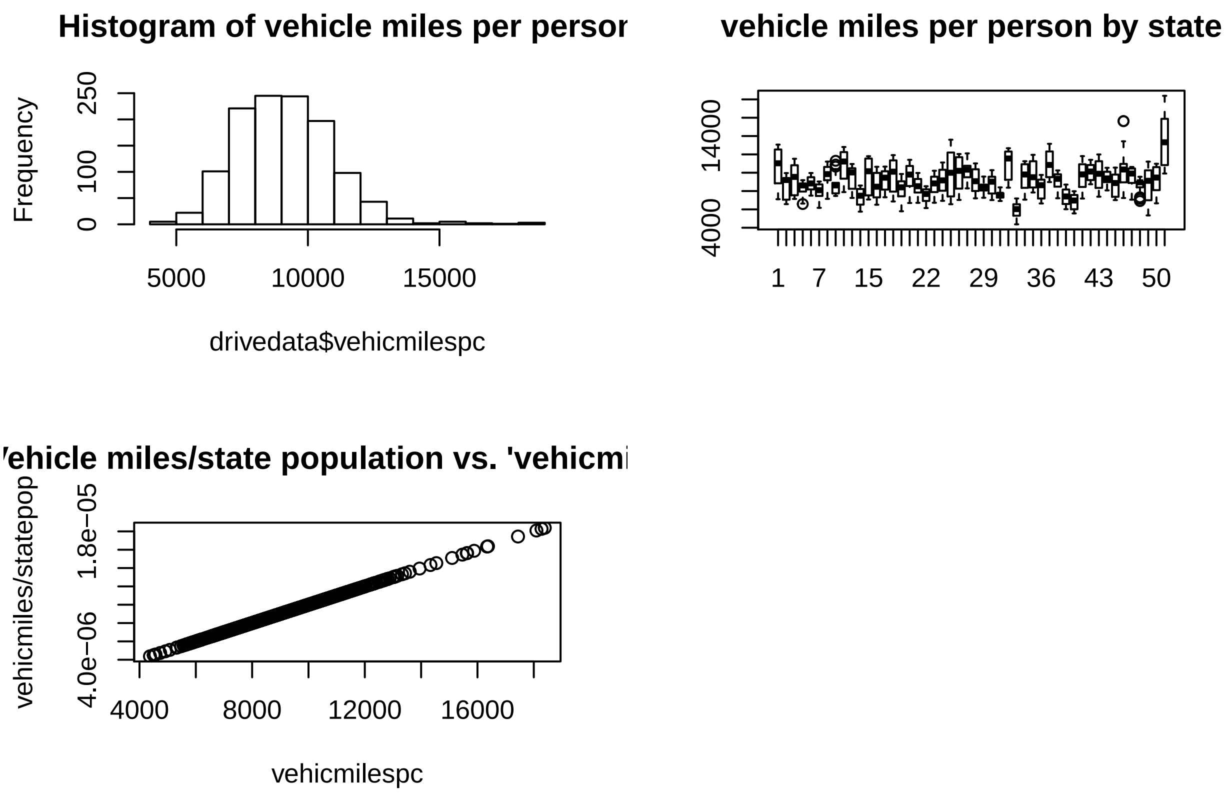
There does not appear to be any discernible relationship between unemployment rates and driving fatality rates.

```
ggplot(drivedata, aes(x=year, y=log(unem)))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Unemployment rates per person vs. fatality rates")
```



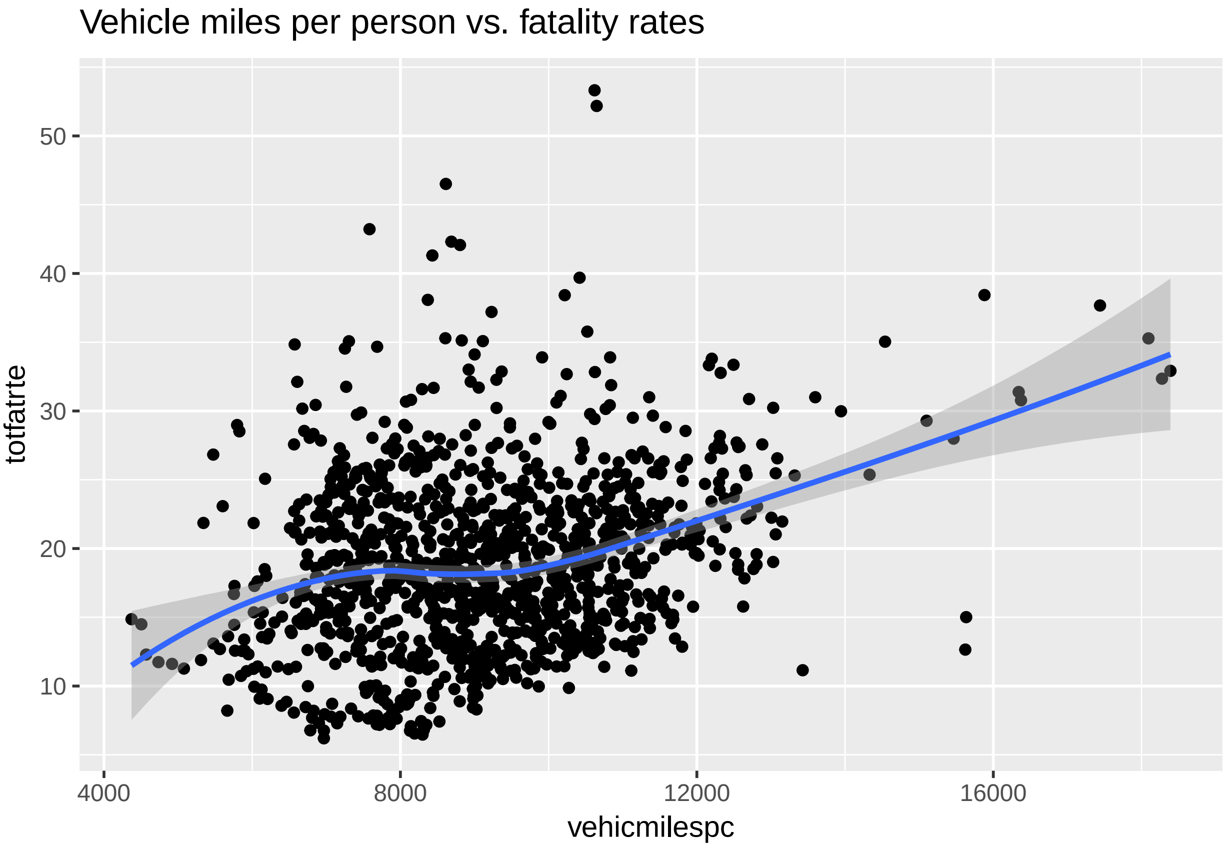
Unemployment apears to have gone down over time, with a slight uptick in 1990.

```
par(mfrow=c(2,2))
hist(drivedata$vehicmilespc, main="Histogram of vehicle miles per person")
boxplot(vehicmilespc~state, data=drivedata, main="vehicle miles per person by state")
plot(vehicmiles/statepop~vehicmilespc, data=drivedata, main="Vehicle miles/state population vs. 'vehicm
```



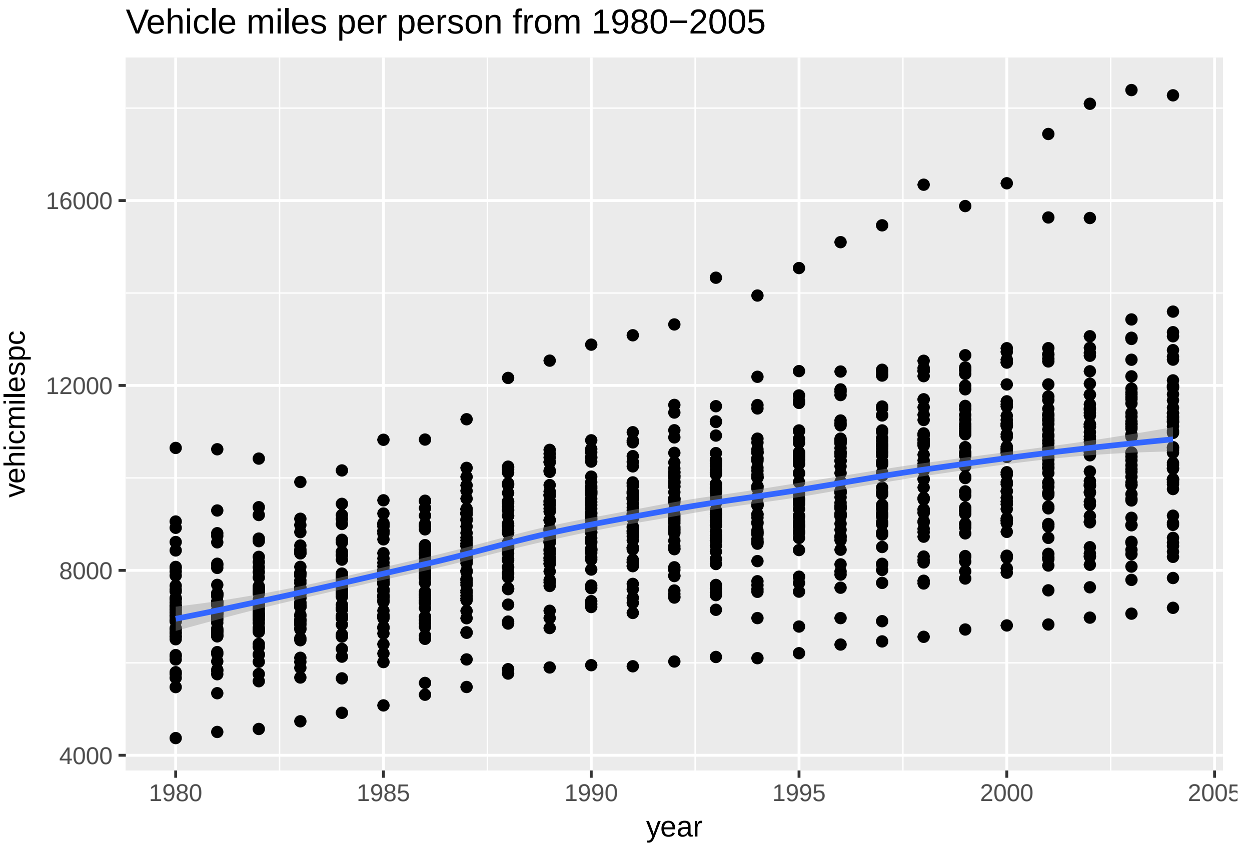
Although there is no description given for how "vehicmilespc" was calculated, we have shown (in the third plot above) that it is the number of vehicle miles travelled (in billions) divided by the state population. Despite this attempt at correction, it is clear from the boxplots that some states tend to have a higher or lower value for this variable.

```
ggplot(drivedata, aes(x=vehicmilespc, y=totfatrte))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Vehicle miles per person vs. fatality rates")
```



Fatality rates appears to be positively correlated with vehicle miles, although these relationship may be driven by a relatively small number of data points with large influence.

```
ggplot(drivedata, aes(x=year, y=vehicmilespc))+
  geom_point()+geom_smooth(method="loess")+ggtitle("Vehicle miles per person from 1980-2005")
```



The number of vehicle miles per person appears to steadily increase over time across states in a linear fashion.