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

Instructions (Please Read Carefully):

- **Due Date: Sunday 04/19/20 11:59pm**
- 20 page limit
- Do not modify fontsize, margin or line-spacing settings
- One student from each group should submit the lab to their student github repo by the deadline; submission and revisions made after the deadline will not be graded
- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation
- Submit two files:
  1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes in your pdf file
  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted
- Name your files to include all group members names. For example the students' names are Stan Cartman and Kenny Kyle, name your files as follows:
  - `StanCartman_KennyKyle_Lab3.Rmd`
  - `StanCartman_KennyKyle_Lab3.pdf`
- Although it sounds obvious, please write your names on page 1 of your pdf and Rmd files
- For statistical methods that we cover in this course, use the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as `dplyr`, `ggplot2`, etc.
- Your report needs to include:
  - A thorough analysis of the given dataset, which includ examiniation of anomalies, missing values, potential of top and/or bottom code, and other potential anomalies, in each of the variables.
  - A comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Be selective when choosing visuals and tables to illustrate your key points and concise with your explanations (please do not ramble).
  - A proper narrative for each question answered. Make sure that your audience can easily follow the logic of your analysis and the rationale of decisions made in your modeling,

  supported by empirical evidence. Use the insights generated from your EDA step to guide your modeling approach.
  - Clear explanations of all steps used to arrive at a final model, with conclusions that summarize results with respect to the question(s) being asked and key takeaways from the analysis.
- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file.
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity

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.

Exercises:

1. (40%) Load the data. Provide a description of the basic structure of the dataset, as we have done throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive a zero in this exercise.*

```r
load(file = "driving.RData")
driving <- data
sum(is.na(driving))
```

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

```r
nrow(driving)
```

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

```r
colnames(driving)
```

```
##  [1] "year"        "state"       "sl55"        "sl65"
##  [5] "sl70"        "sl75"        "slnone"      "seatbelt"
##  [9] "minage"      "zerotol"     "gdl"         "bac10"
## [13] "bac08"       "perse"       "totfat"      "nghtfat"
## [17] "wkndfat"     "totfatpvm"   "nghtfatpvm"  "wkndfatpvm"
## [21] "statepop"    "totfatrte"   "nghtfatrte"  "wkndfatrte"
## [25] "vehicmiles"  "unem"        "perc14_24"   "sl70plus"
## [29] "sbprim"      "sbsecon"     "d80"         "d81"
## [33] "d82"         "d83"         "d84"         "d85"
## [37] "d86"         "d87"         "d88"         "d89"
## [41] "d90"         "d91"         "d92"         "d93"
## [45] "d94"         "d95"         "d96"         "d97"
## [49] "d98"         "d99"         "d00"         "d01"
## [53] "d02"         "d03"         "d04"         "vehicmilespc"
```

There are no missing values in the dataset and there are 1200 observations. Here the subjects are indicated by the state variable. For each state there are multiple observations, one for each year.

```r
head(driving,10)
```

```
##    year state  sl55  sl65 sl70 sl75 slnone seatbelt minage zerotol gdl
## 1  1980     1 1.000 0.000    0    0      0        0     18       0   0
## 2  1981     1 1.000 0.000    0    0      0        0     18       0   0
## 3  1982     1 1.000 0.000    0    0      0        0     18       0   0
## 4  1983     1 1.000 0.000    0    0      0        0     18       0   0
## 5  1984     1 1.000 0.000    0    0      0        0     18       0   0
## 6  1985     1 1.000 0.000    0    0      0        0     20       0   0
## 7  1986     1 1.000 0.000    0    0      0        0     21       0   0
## 8  1987     1 0.542 0.458    0    0      0        0     21       0   0
## 9  1988     1 0.000 1.000    0    0      0        0     21       0   0
## 10 1989     1 0.000 1.000    0    0      0        0     21       0   0
##    bac10 bac08 perse totfat nghtfat wkndfat totfatpvm nghtfatpvm
## 1      1     0     0    940     422     236     3.200      1.437
## 2      1     0     0    933     434     248     3.350      1.558
## 3      1     0     0    839     376     224     2.810      1.259
## 4      1     0     0    930     397     223     3.000      1.281
## 5      1     0     0    932     421     237     2.830      1.278
## 6      1     0     0    882     358     224     2.510      1.019
## 7      1     0     0   1080     500     279     3.177      1.471
## 8      1     0     0   1111     499     300     2.970      1.334
## 9      1     0     0   1024     423     226     2.580      1.066
## 10     1     0     0   1029     418     247     2.520      1.024
##    wkndfatpvm statepop totfatrte nghtfatrte wkndfatrte vehicmiles unem
## 1       0.803  3893888     24.14      10.84       6.06   29.37500  8.8
## 2       0.890  3918520     24.07      11.08       6.33   27.85200 10.7
## 3       0.750  3925218     21.37       9.58       5.71   29.85765 14.4
## 4       0.719  3934109     23.64      10.09       5.67   31.00000 13.7
## 5       0.720  3951834     23.58      10.65       6.00   32.93286 11.1
## 6       0.637  3972527     22.20       9.01       5.64   35.13944  8.9
## 7       0.821  3991569     27.08      12.53       6.99   33.99371  9.8
## 8       0.802  4015261     27.67      12.43       7.47   37.40741  7.8
## 9       0.569  4023858     25.45      10.51       5.62   39.68992  7.2
## 10      0.605  4030229     25.53      10.37       6.13   40.83333  7.0
##    perc14_24 sl70plus sbprim sbsecon d80 d81 d82 d83 d84 d85 d86 d87 d88
## 1       18.9        0      0       0   1   0   0   0   0   0   0   0   0
## 2       18.7        0      0       0   0   1   0   0   0   0   0   0   0
## 3       18.4        0      0       0   0   0   1   0   0   0   0   0   0
## 4       18.0        0      0       0   0   0   0   1   0   0   0   0   0
## 5       17.6        0      0       0   0   0   0   0   1   0   0   0   0
## 6       17.3        0      0       0   0   0   0   0   0   1   0   0   0
## 7       17.0        0      0       0   0   0   0   0   0   0   1   0   0
## 8       16.6        0      0       0   0   0   0   0   0   0   0   1   0
## 9       16.2        0      0       0   0   0   0   0   0   0   0   0   1
## 10      15.8        0      0       0   0   0   0   0   0   0   0   0   0
##    d89 d90 d91 d92 d93 d94 d95 d96 d97 d98 d99 d00 d01 d02 d03 d04
```

```
## 1     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 2     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 3     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 4     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 5     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 7     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 9     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## 10    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##    vehicmilespc
## 1      7543.874
## 2      7107.785
## 3      7606.622
## 4      7879.802
## 5      8333.562
## 6      8845.614
## 7      8516.377
## 8      9316.308
## 9      9863.649
## 10    10131.764
```

```r
table(driving$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(driving$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
```

For each state there are 25 observations one for each year. For each year there are 48 observations, one for each state. *This panel is balanced.* There are a total of 1200 observations and since $1200 = 25X48$ we have a balanced panel. d80 through d04 are the indicator variables for each of the time periods.

We cannot use OLS here because we suspect voilation of the independence assumptions. Let us confirm that using the Durbin-Watson test.

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
```

```
## Attaching package: 'zoo'

## The following object is masked from 'package:tsibble':
##
##     index

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dwtest(totfatrte ~ seatbelt + zerotol + slnone, data=driving)
```

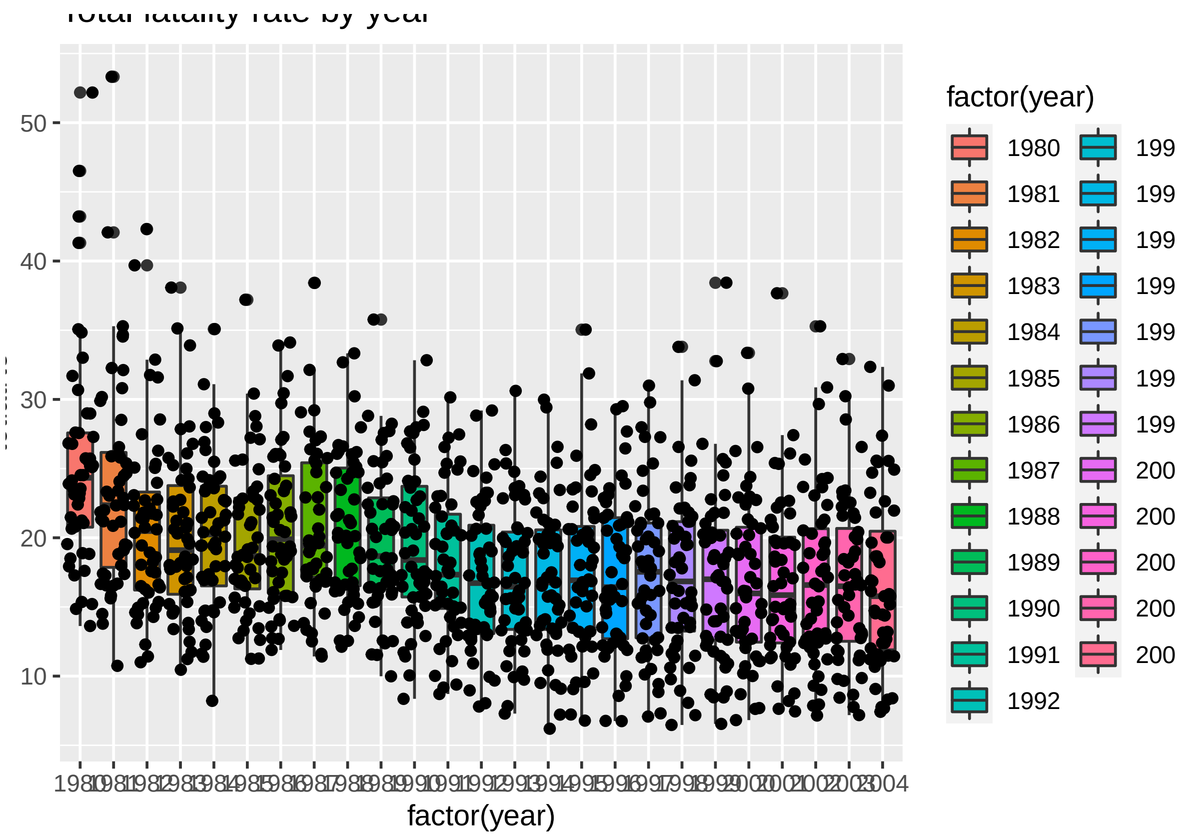
```
##
##  Durbin-Watson test
##
## data:  totfatrte ~ seatbelt + zerotol + slnone
## DW = 0.24975, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Null hypothesis is rejected, this confirms the voilation of the independence assumption.

So we will have to use panel methods here. There are 25 panels in this dataset, one for each year of observations.

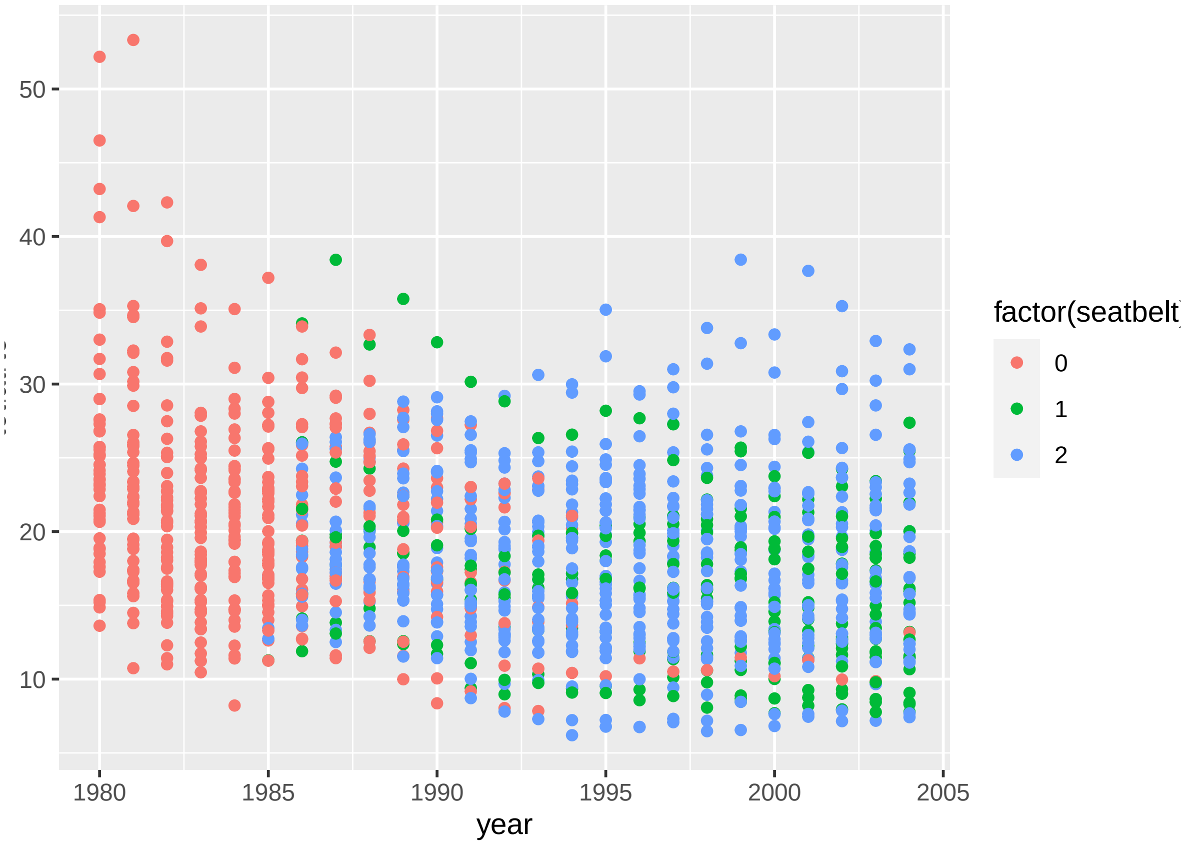
Let's see the response variable (total fatality rate) distribution broken down by panels

```
ggplot(driving, aes(factor(year), totfatrte)) +  geom_boxplot(aes(fill = factor(year))) + geom_
```



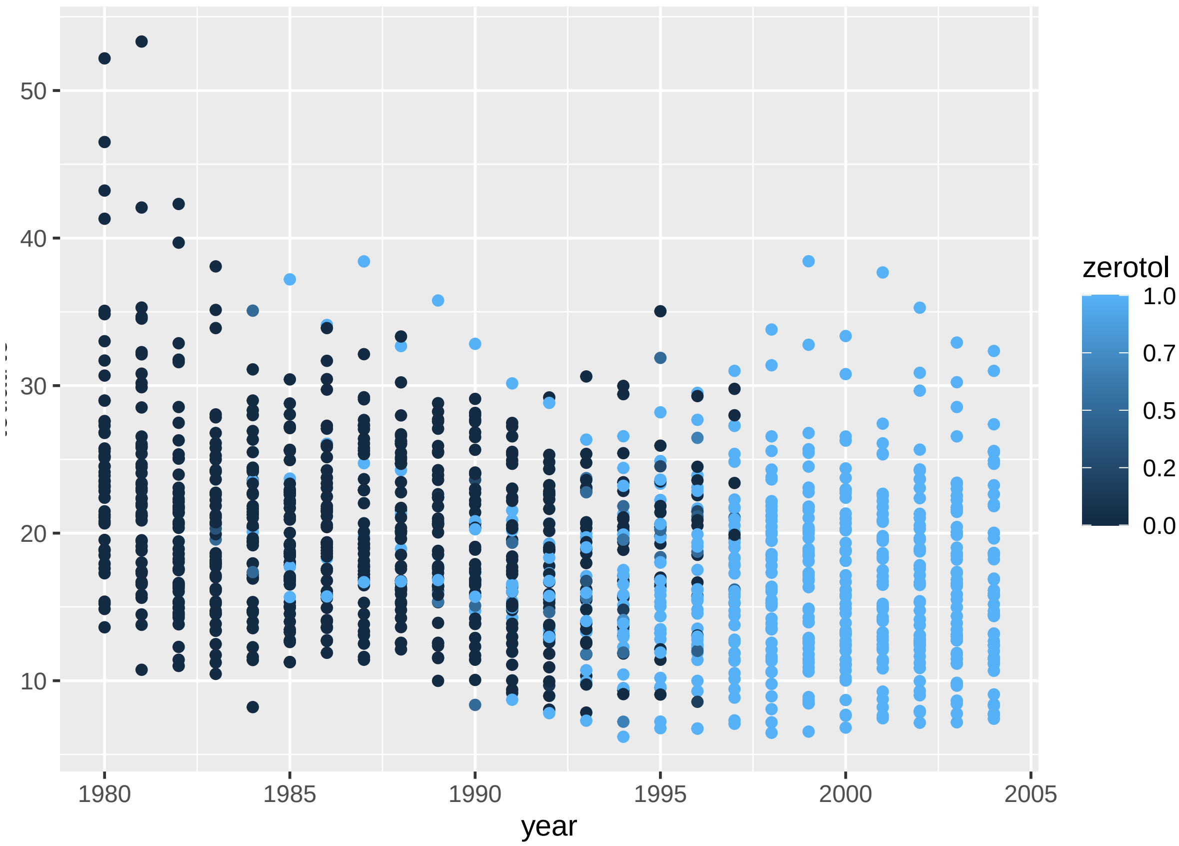
Overall, the total fatality rate has dropped over the years.

```
ggplot(driving, aes(year, totfatrte)) +  geom_point(aes(color = factor(seatbelt)))
```



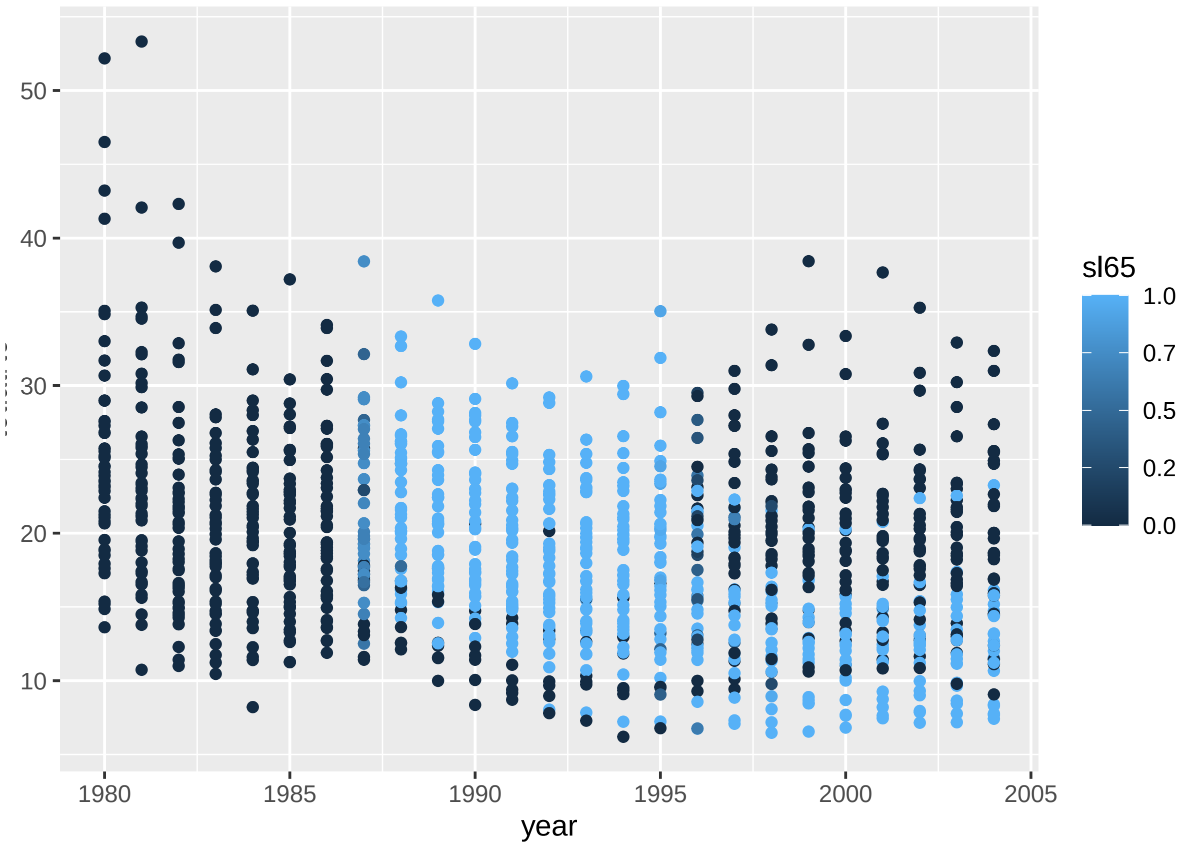
Early years, pre-1985 there seems to have been no seatbelt law. The fatality rate dropped over the years as 1, 2 seatbelt laws were introduced. States started with introduction of seatbelt rule 2 and then majority of them moved to seatbelt rule 1.

```
ggplot(driving, aes(year, totfatrte)) +  geom_point(aes(color = zerotol))
```



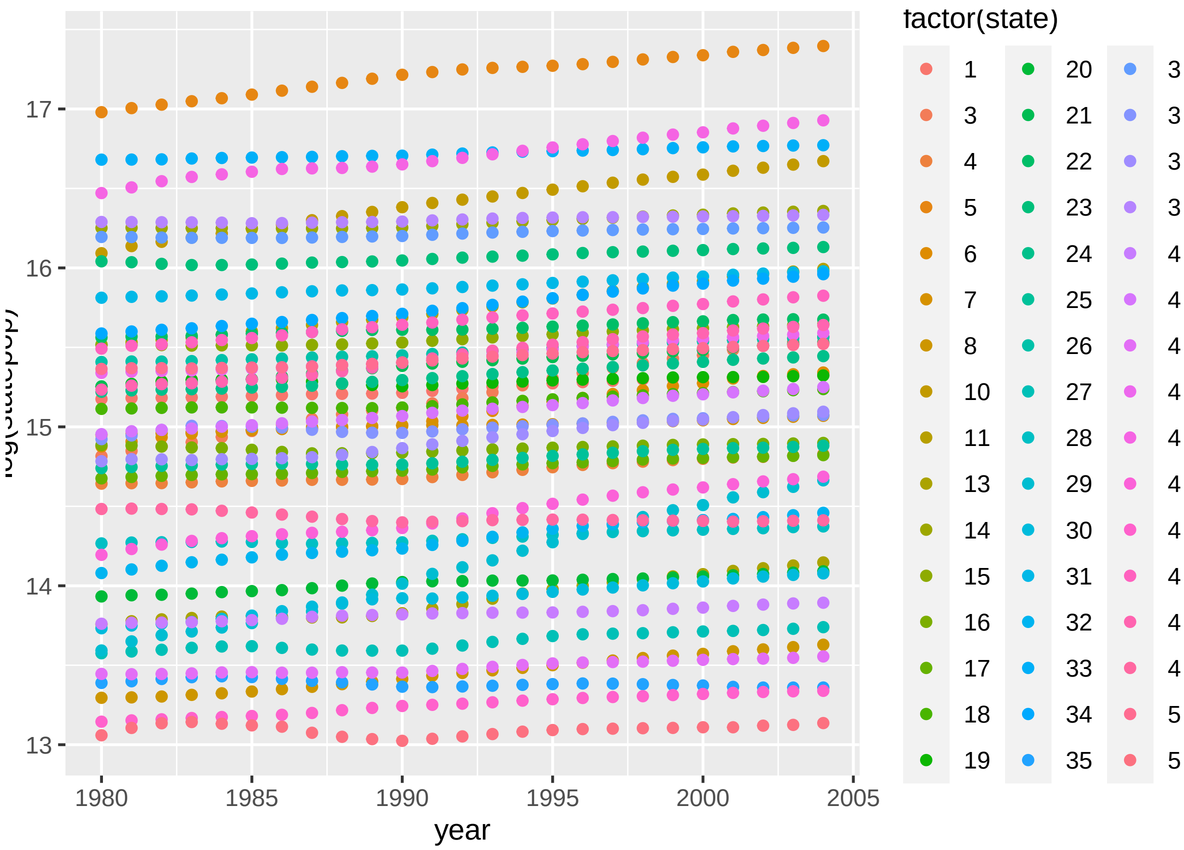
Most of the states moved to a zero tolerance policy over the years. In fact all the states seems to have introduced theh zero tolerance policy.

```
ggplot(driving, aes(year, totfatrte)) +  geom_point(aes(color = sl65))
```



Most of the states seemed to have moved to speed limit 65 policy which correlates storngly with the drop in fatality rate. However several states seemed to have moved away to a possibly lower limit (55 mph) post 1996.

```
ggplot(driving, aes(year, log(statepop))) +  geom_point(aes(color = factor(state)))
```



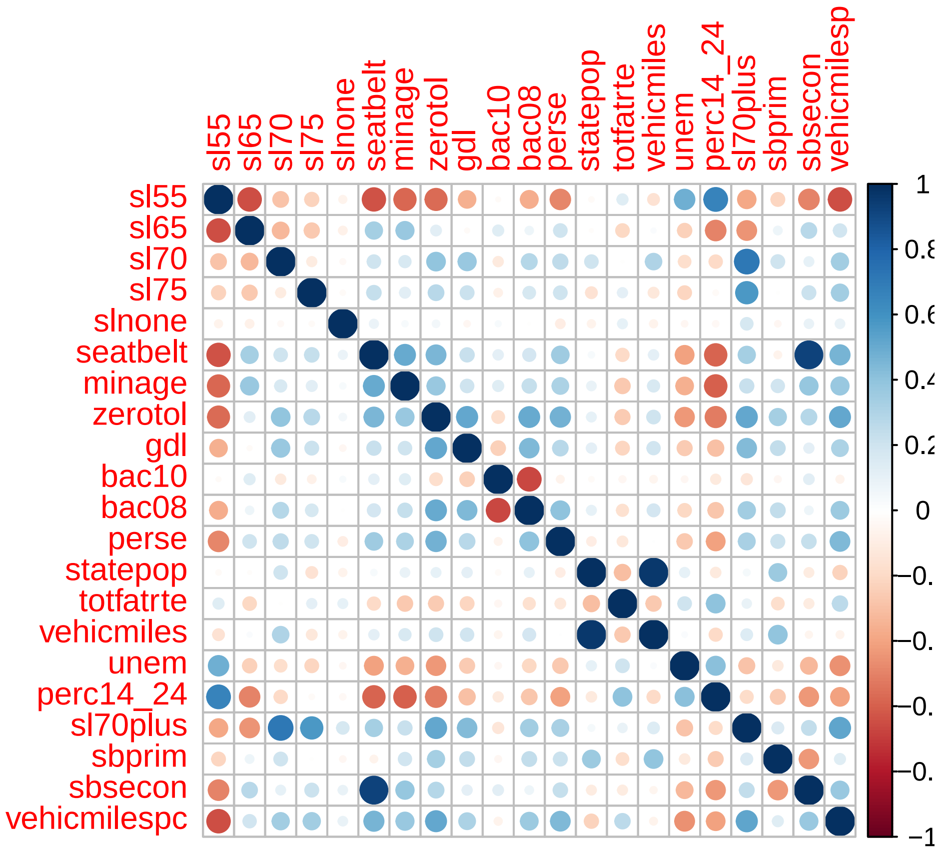
TODO: find the variables that are time invariant and the ones which are time varying. Perform specific EDA on those varibles.

We could take a look at the panel data after inserting a structure into the dataset. The indices here are the "state" and the "year".
We will have to select the appropriate explanatory variables to look at the correlation between those variables and the fatality rate. We would drop some of variables which are likely to have a higher correlation such as "fatality rate" with "weekend fatality rate" and "night fatality rate" etc. from the dataset used for the correlation matrix.

```
library(plm)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
driving.panel <- pdata.frame(driving, c("state","year"), drop.index = TRUE)
M <- cor(driving.panel[,c(1:12,19:20,23:28,54)])
corrplot(M, method='circle')
```



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.

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.

Running OLS only on data for one panel only, the equation can be written as below

Fixed year effect

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

δ_t the change that is common to every city in year t. It estimates the common change in the fatality rate in year t relative to the base / reference year 1980. We assume that this change is common across all cities for a given year t.

```
driving.ols <- lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 +
summary(driving.ols)
```

```
##
## 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)
##
## 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
```

From the above coefficients, we can say that for the year 2000, the fatality rate has dropped by 8.66 units as compared to the year 1980. Except for the year 1981, the coefficients for all the year dummies are significant. From this we can also say that the fatality rate has been dropping over the years. Driving did seem to become safer over the years.

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.)

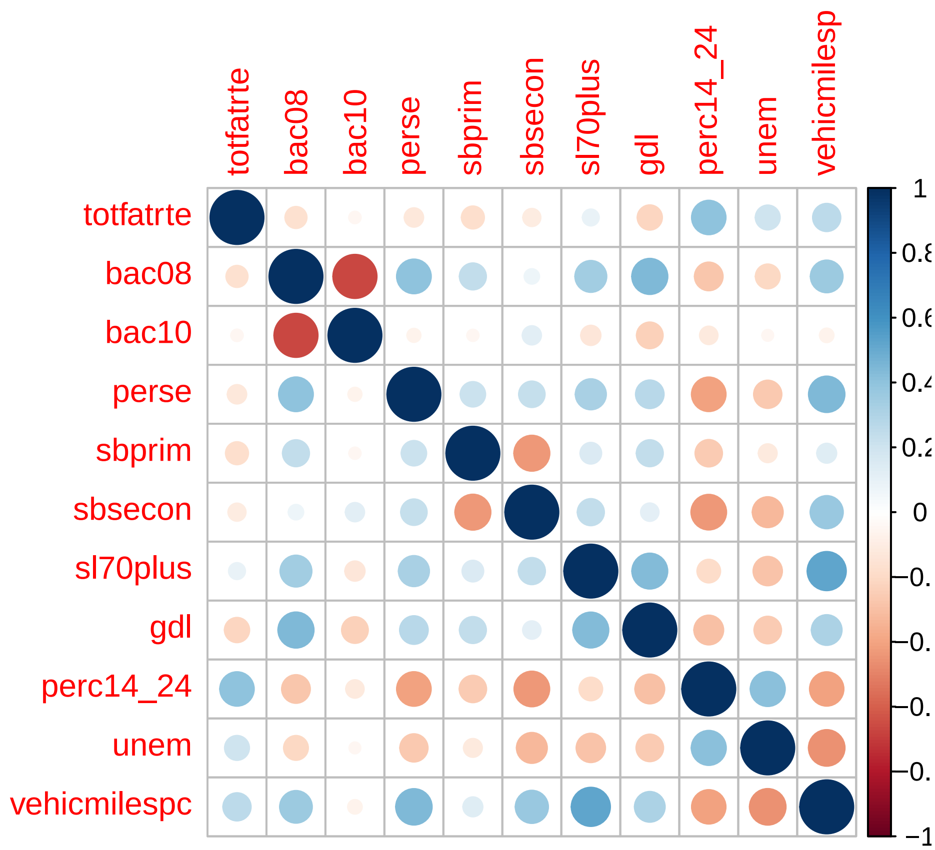
```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
df_3 <- driving[, c('totfatrte','bac08','bac10','perse','sbprim','sbsecon','sl70plus','gdl','pe
M <- cor(df_3)
corrplot(M, method='circle')
```



The correlation plot shows that the fatality rate has a high correlation with the *perc_14_24*.

```
driving.ols.pool <- lm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 +
summary(driving.ols.pool)
```

```
##
## 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)
##
## 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
```

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. All the bac08, bac10 and sbprim are all significant. They also have a strong negative relationship with the fatality rate. The perse however is not very significant. However, speed limit of 70-plus did turn out significiant. It fact it indicates that controlling for all other factors, a unit change in *sl70plus* increases the fatality rate by 3.34 units.

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?

```
driving.panel <- pdata.frame(driving, c("state","year"), drop.index = TRUE)
df_4.panel <- driving.panel[, c('d81', 'd82', 'd83', 'd84', 'd85', 'd86', 'd87', 'd88', 'd89',
fatrte.plm <- plm(totfatrte ~ d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91 
summary(fatrte.plm)
```

```
## 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 = df_4.panel, 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
```

Just like the pooled OLS, the coefficients for the parameters bac08 and bac10are very significant. They also have a very strong negative relationship with fatality rate. However the coefficients indicate that the effect of the variables are less when compared with what was reported in the pooled OLS. e.g. for bac08, the fatality rate drops by 1.4 units for every unit increase in bac08 level, however with the pooled OLS, the indication is that for every unit increase in bac08, the fatality rate drops by 2.49 units. The fixed effect model is much more conservative when compared to the pooled OLS. The *perse* variable turns out to be significant in the fixed effect model. The effect of the variable reported higher when compared to the effect reported in the pooled OLS.

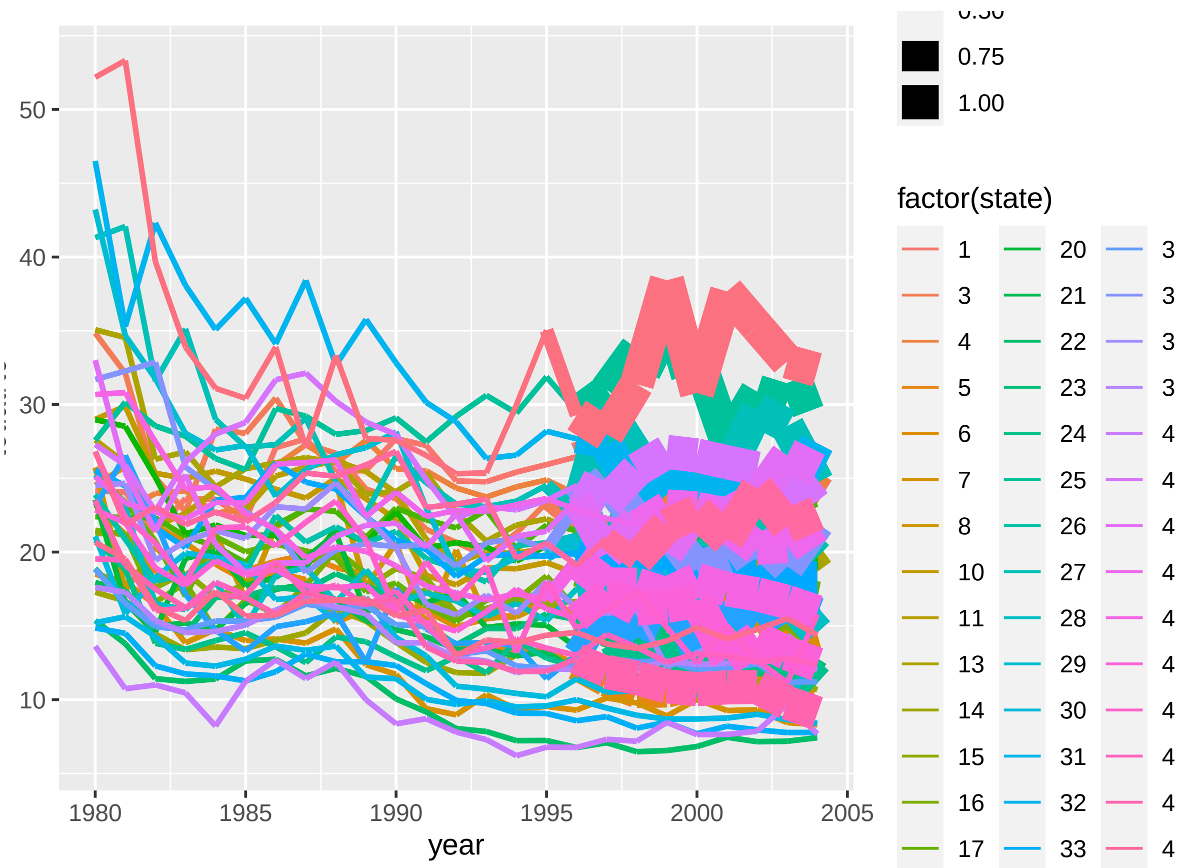
5. (5%) Would you perfer to use a random effects model instead of the fixed effects model you built in *Exercise 4*? Please explain.

```r
driving.panel <- pdata.frame(driving, c("state","year"), drop.index = TRUE)
df_4.panel <- driving.panel[, c('d81', 'd82', 'd83', 'd84', 'd85', 'd86', 'd87', 'd88', 'd89',
fatrte.plm <- plm(totfatrte ~  d81 + d82 + d83 + d84 + d85 + d86 + d87 + d88 + d89 + d90 + d91
summary(fatrte.plm)
```

```
## Oneway (individual) effect Between 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 = df_4.panel, model = "between")
##
## Balanced Panel: n = 48, T = 25, N = 1200
## Observations used in estimation: 48
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -6.34853 -1.29750 -0.24859  1.87280  4.49502
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)  -1.1591e+01  9.9902e+00 -1.1602 0.2533950
## bac08        -2.4650e+00  2.6984e+00 -0.9135 0.3668886
## bac10        -1.9318e+00  2.2638e+00 -0.8533 0.3989618
## perse         1.5811e-01  1.4353e+00  0.1102 0.9128815
## sbprim       -1.3912e+00  2.9251e+00 -0.4756 0.6371565
## sbsecon      -1.7252e+00  3.1144e+00 -0.5539 0.5829500
## sl70plus      1.1145e+01  3.0755e+00  3.6237 0.0008676 ***
## gdl          -2.9461e+00  4.5422e+00 -0.6486 0.5205996
## perc14_24    -1.8133e-01  5.9151e-01 -0.3066 0.7608998
## unem          1.4555e+00  3.8436e-01  3.7868 0.0005433 ***
## vehicmilespc  2.7972e-03  4.2742e-04  6.5444 1.155e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1459.1
## Residual Sum of Squares: 312.88
## R-Squared:      0.78557
## Adj. R-Squared: 0.72761
## F-statistic: 13.5548 on 10 and 37 DF, p-value: 1.3818e-09
```

The random effects model does not allow for arbitrary correlations between a_i (i.e. unobserved fixed effects) and x_{itj}. The random effects model has ignored the coefficients for the year dummies. It has reported only 3 variables as significant. The *sl70plus*, *unem* and *vehicmilespc*. Is it reasonable to assume that the *bac08* or *bac10* are uncorrelated to the unobserved fixed effects? e.g. there could be many reasons why people drive drunk other than just the fact that they may be young or unemployed. It may also depend on the type of jobs people are involved in, in their respective cities, or the culture of the cities etc. These unobserved fixed effects may have correlation with *bac08* and or *bac10*. Similarly, *unem* may have strong correlation with unobserved city effects, due to the geographical features or the city, presence of industries that rely on natural resources such as oil and gas, mines etc. may be limited that effects the unemployment rates of the city.

```
driving <- driving %>% mutate(sl70plus_bin = ifelse(sl70plus > 0.5, 1, 0))
ggplot(driving, aes(x=year, y=totfatrte, size=sl70plus, color=factor(state))) + geom_line()
```



The 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 imporant explanatory variable for the fatality rate, however it comes up as very significant in the RE model.

All the above provides strong evidence that a fixed effect model might be more appropriate for this dataset.

6. (5%) Suppose that *vehicmilespc*, the number of miles driven per capita, increases by $1,000$. Using the FE estimates, what is the estimated effect on *totfatrte*? Please interpret the estimate.

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