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

Instructions (Please Read Carefully):

- **Due 4pm Tuesday August 11 2020**
- 20 page limit (strict)
- 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

```
library(foreign)
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 4.0.2
```

```
library(ggplot2)
library(stats)
library(Hmisc)
library(car)
library(usmap)
```

```
## Warning: package 'usmap' was built under R version 4.0.2
```

```
library(dplyr)
library(gridExtra)
library(stargazer)
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.0.2
```

```
library(data.table)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
library(grid)
```

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:

Part 1

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("driving.RData")

# one row per year per state
head(table(data$year, data$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
##   1980 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1981 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1982 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1983 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1984 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1985 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##
##        30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
##   1980  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1981  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1982  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1983  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1984  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##   1985  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

```
max(data$year)
```

```
## [1] 2004
```

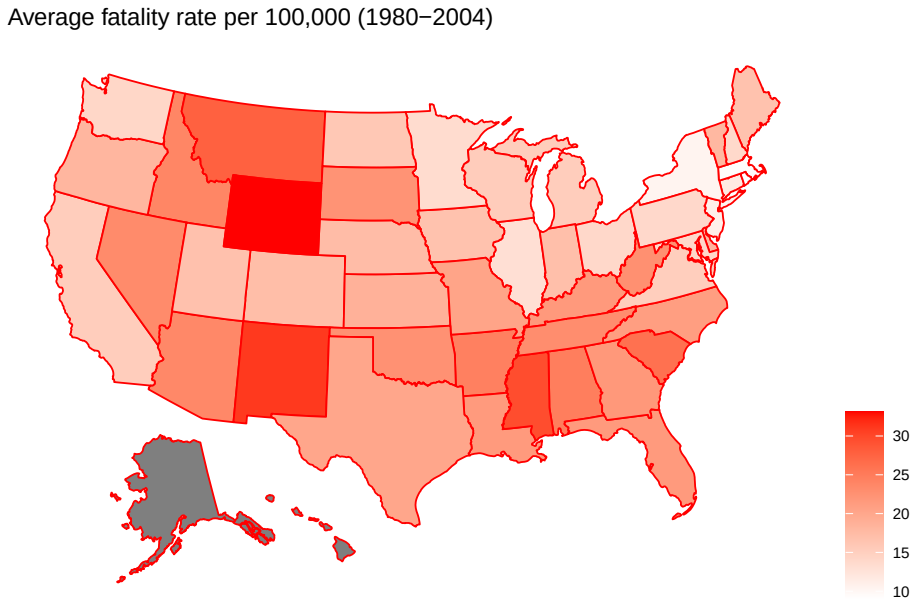
```
fips_map <-read.csv("statecodes.csv")

d_data <- merge(x=data, y=fips_map, by="state",all.x = TRUE) %>% dplyr::select(-state)
d_data <- rename(d_data, c("state"="code"))

d_data <- data.table(d_data)

# average fatality rate per 100,000 across states
state_avg <- d_data %>% group_by(state) %>% summarise(avg_totfatrte=mean(totfatrte))
```

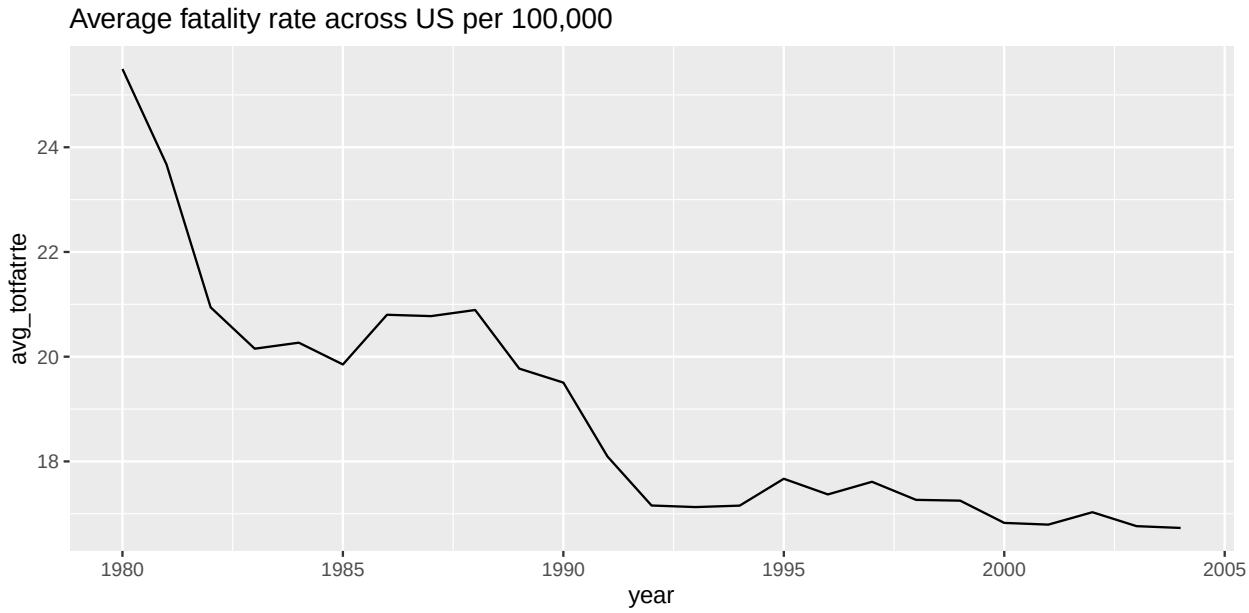
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
state_avg <- rename(state_avg, c("value"="avg_totfatrte"))

p1.1 <- plot_usmap(data = state_avg,  values="value", color = "red") +
  scale_fill_continuous(name="", low="white", high="red") +
  theme(legend.position = "right") +ggtitle("Average fatality rate per 100,000 (1980-2004)")

# average fatality rate per 100,000 across years
year_avg <- d_data %>% group_by(year) %>% summarise(avg_totfatrte=mean(totfatrte))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
p1.2 <- ggplot(data=year_avg, aes(x=year, y=avg_totfatrte)) +
  geom_line()+ ggtitle("Average fatality rate across US per 100,000")

grid.arrange(p1.1, p1.2, nrow=2)
```



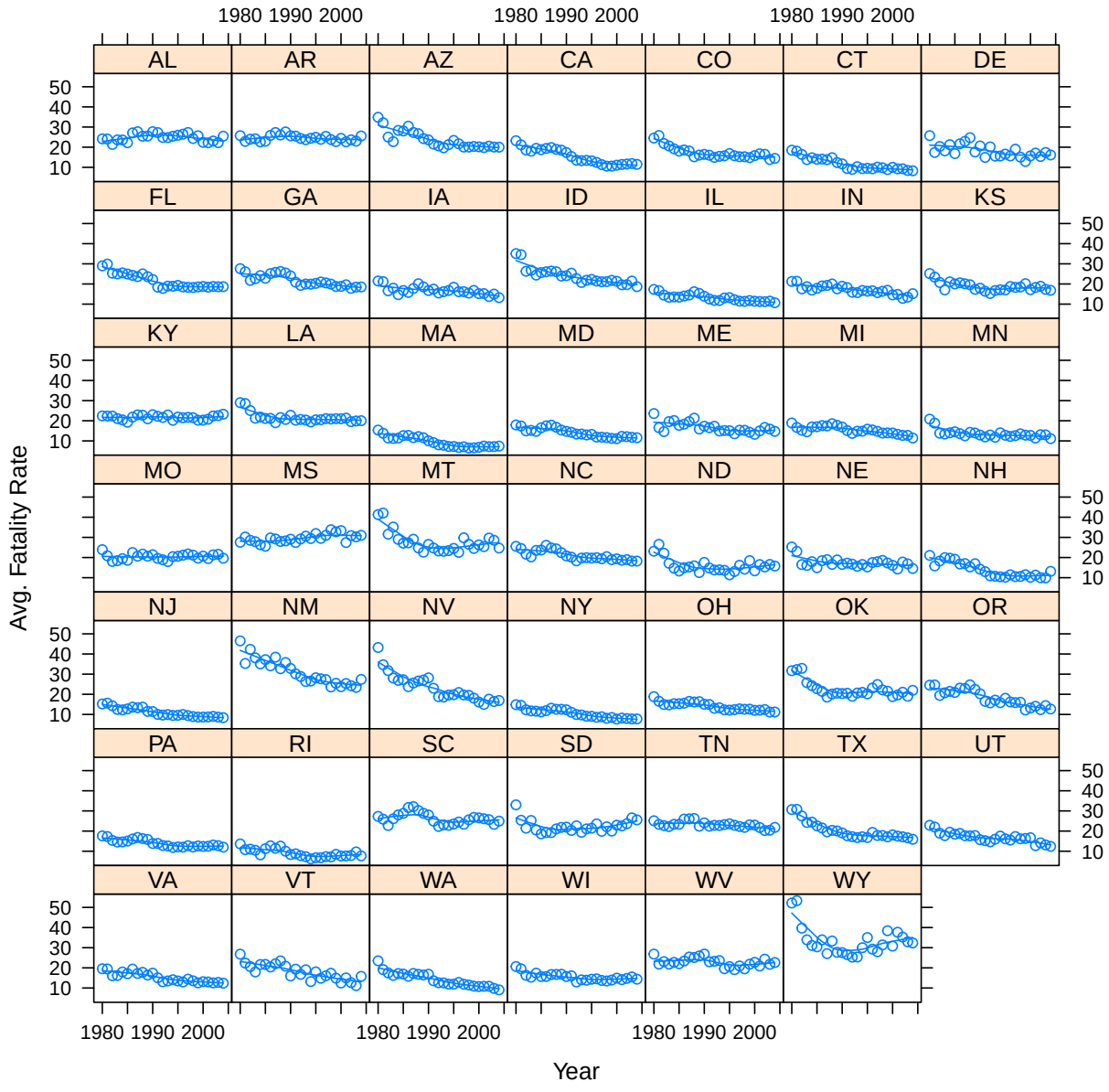


Growth Gurve Analysis

- Note general flat to downward trend with exception of Mississippi
- Nevada and New Mexico drop looks steep

```
xyplot(totfatrte~year | state, data=d_data,
       prepanel = function(x, y) prepanel.loess(x, y, family="gaussian"),
```

```
       xlab = "Year", ylab = "Avg. Fatality Rate",
       panel = function(x, y) {
          panel.xyplot(x, y)
          panel.loess(x,y, family="gaussian") },
       as.table=T)
```




```
# this is hard to read!
#g <- ggplot(data_state, aes(year, totfatrte, colour = as.factor(code)))
#g + geom_line() + ggtitle("Growth Curve by state")
```

Investigation of explanatory variables

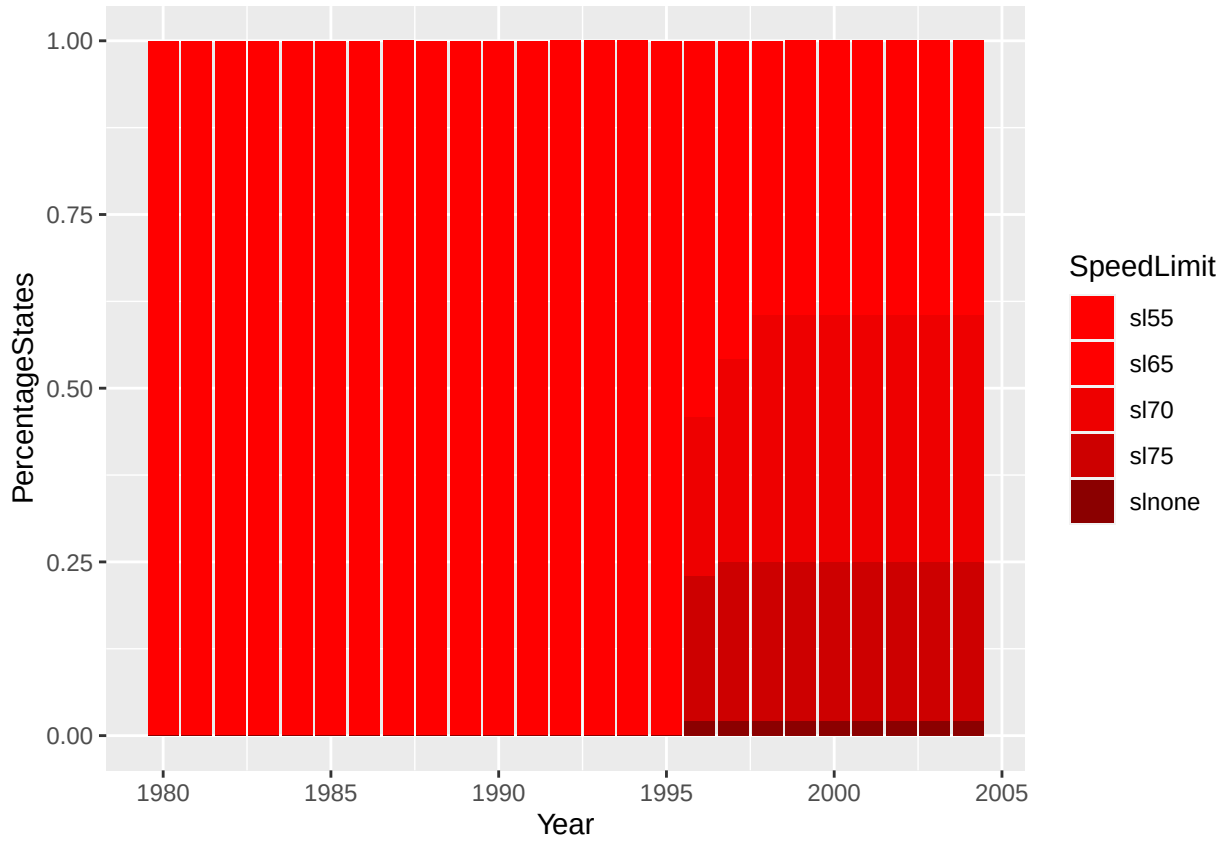
```
# speed limit variables
unique(d_data$sl55)
```

```
##  [1] 1.000 0.542 0.000 0.250 0.333 0.750 0.044 0.083 0.417 0.458 0.500 0.011
## [13] 0.917 0.292 0.049 0.583 0.375
```

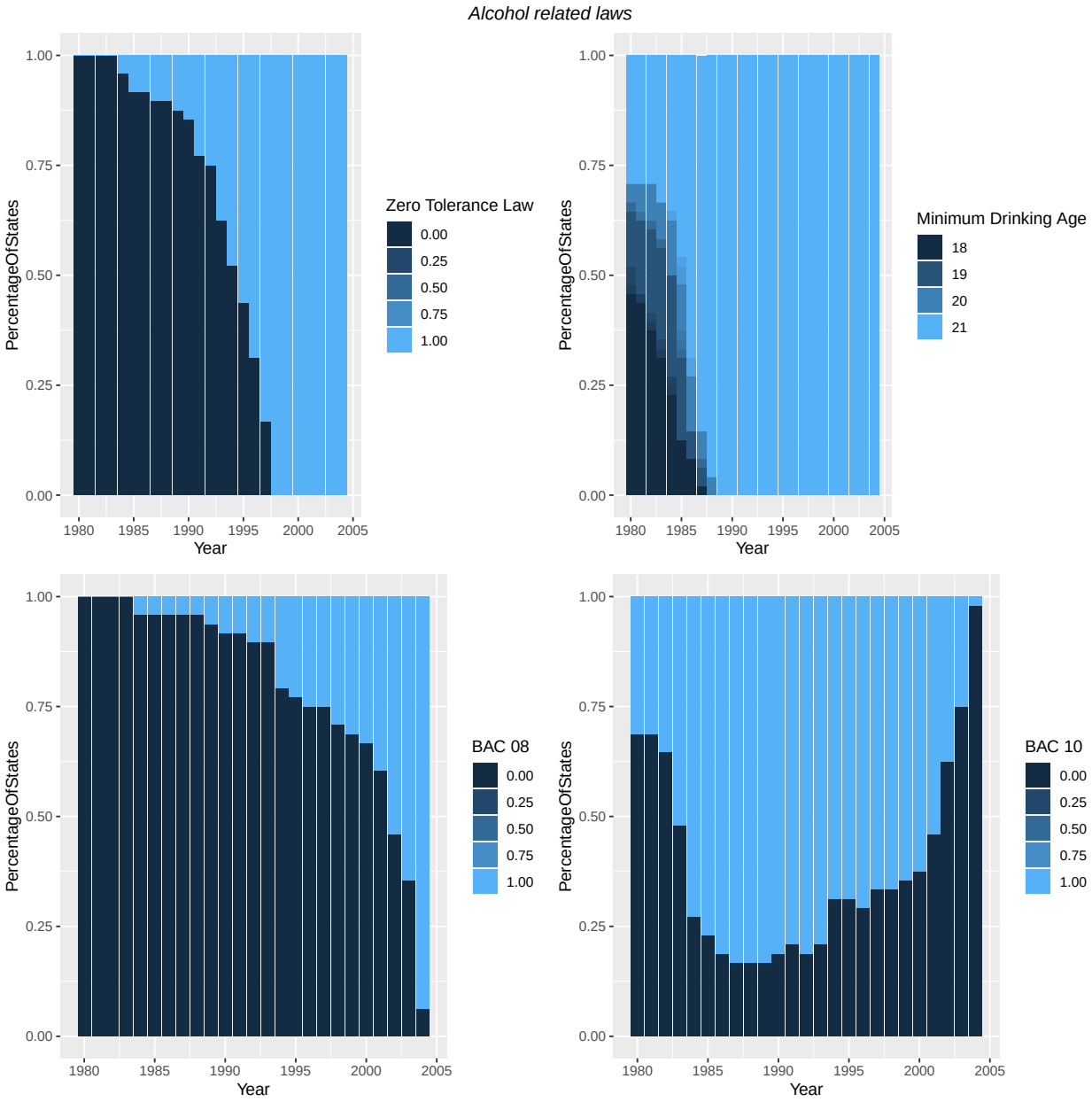
```
# we see that speed limits are not binary. Its a ratio that possibly represents what month in
# We may want to represent this as binary based on which speed limit was more prevalent that y
d_data[sl55==0.5 & sl65==0.5, c("sl55", "sl65"):=list(0,1) ]
d_data[sl65==0.5 & sl70==0.5, c("sl65", "sl70"):=list(0,1) ]
d_data[sl65==0.5 & sl75==0.5, c("sl65", "sl75"):=list(0,1) ]

for (sp in c("sl55","sl65","sl70","sl75", "slnone")){
  d_data[get(sp) >0.5, eval(quote(sp)):= 1]
  d_data[get(sp) <0.5, eval(quote(sp)):= 0]
}
```

How did speed limits change over the years

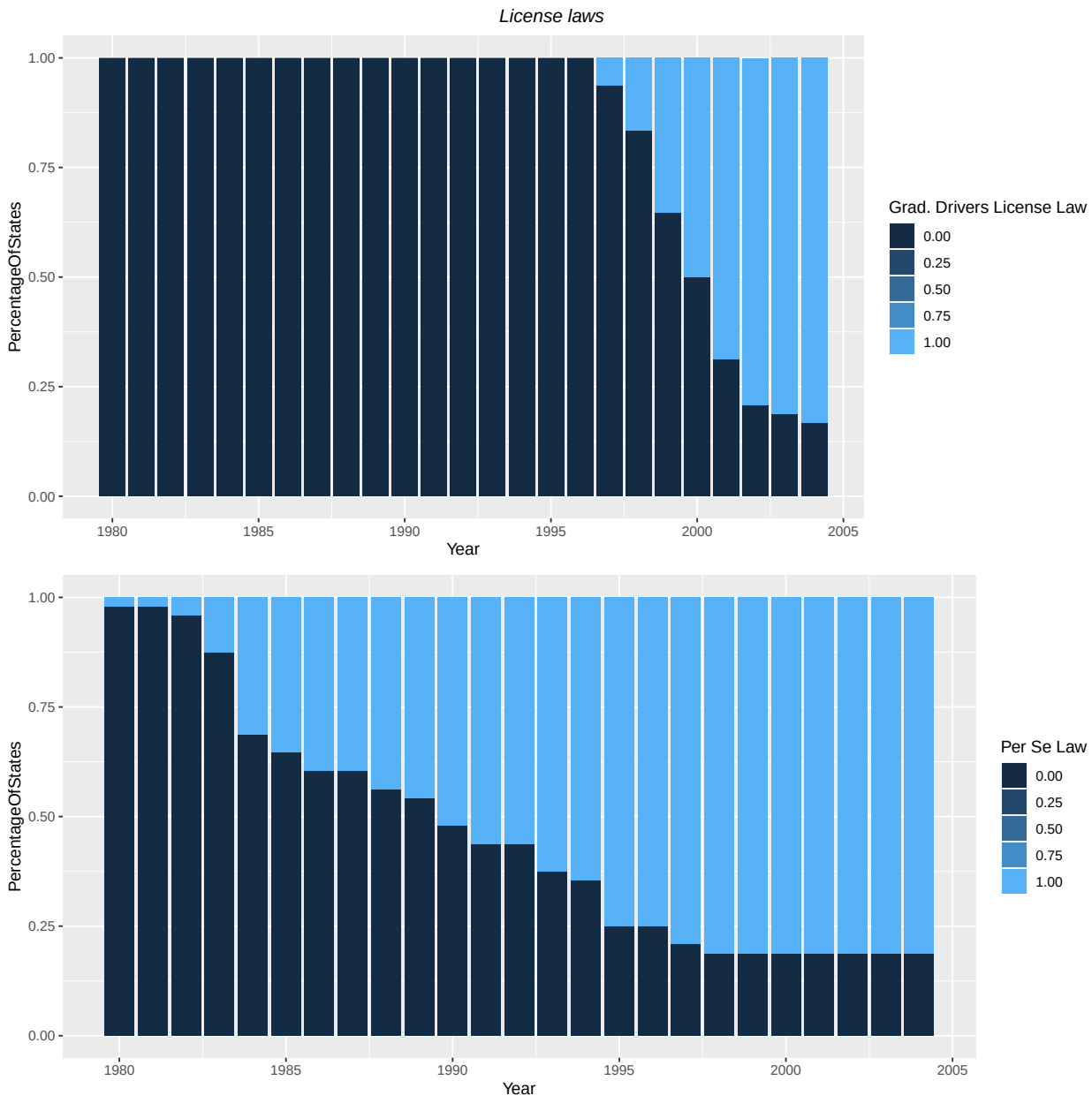
```
d2 <- d_data[, c("year", "state", "sl55","sl65","sl70","sl75", "slnone")]
d3 <- d2 %>%
  gather(key="speed_lim", value="Value", "sl55","sl65","sl70","sl75","slnone") %>%
  dplyr::filter(Value==1) %>%
  dplyr::select(-Value)

d3$year <- factor(d3$year)
d3$speed_lim <- factor(d3$speed_lim)

# breakdown of speed limits counts over the years
xt <- table(d3$speed_lim, d3$year) %>% reshape2::melt() %>% rename(SpeedLimit=Var1, Year=Var2,

# Stacked + percent
ggplot(xt, aes(fill=SpeedLimit, y=PercentageStates, x=Year))+
  geom_bar(position="fill", stat = "identity")+
  scale_fill_manual(values = c("red","red1","red2","red3","red4"))
```



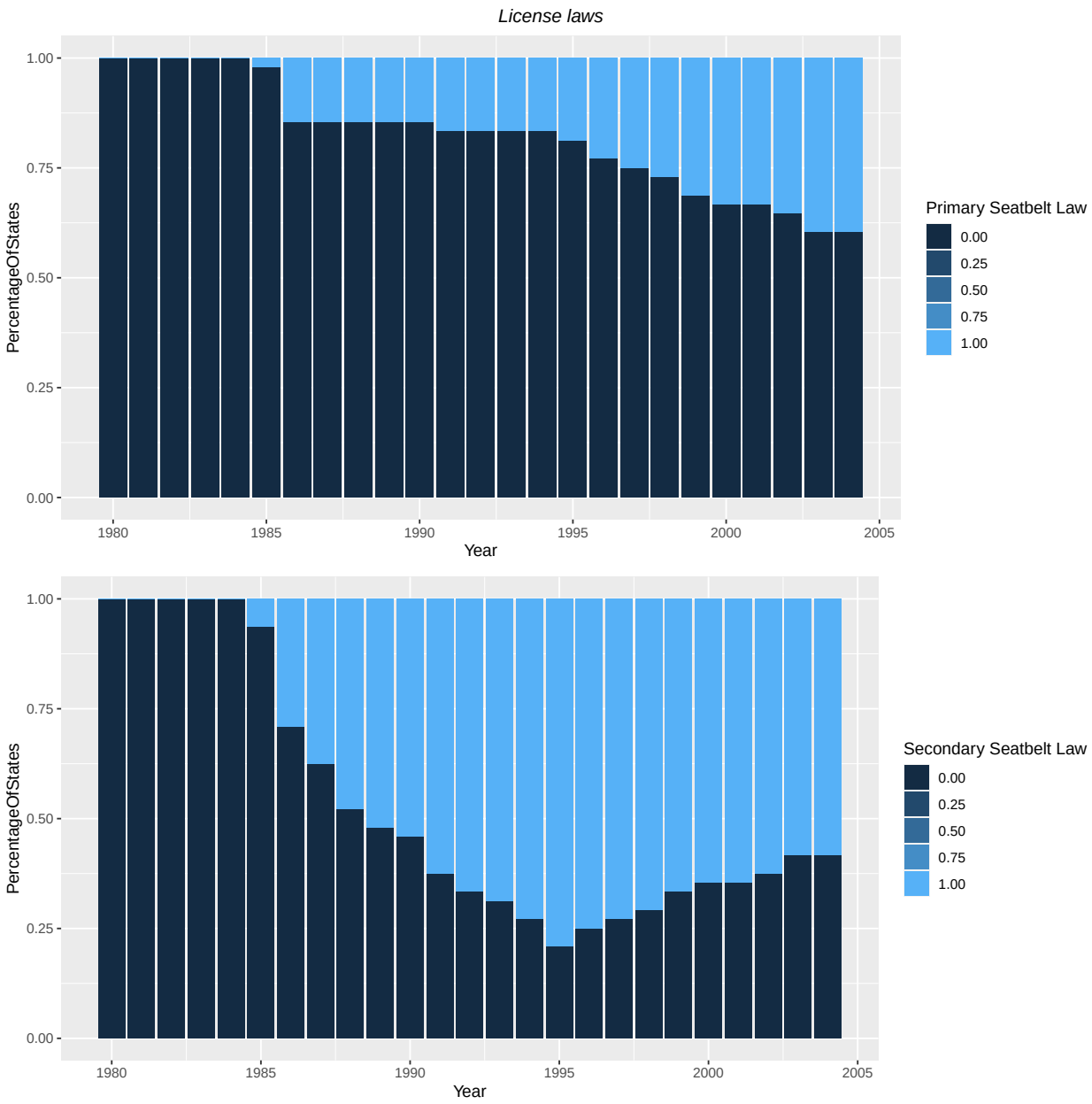

```
for (sp in c("zerotol","gdl","bac10", "bac08", "perse", "sbprim", "sbsecon", "sl70plus")){
  d_data[get(sp) >0.5, eval(quote(sp)):= 1]
  d_data[get(sp) <=0.5, eval(quote(sp)):= 0]
}


genplot <- function(df, col, legend){
  df2 <- table(df[[col]], df$year) %>% reshape2::melt()
  colnames(df2) <- c(col, 'Year', 'PercentageOfStates')
  df2.plot <- ggplot(df2, aes(fill=get(col), y=PercentageOfStates, x=Year))+
                geom_bar(position="fill", stat = "identity") +
              guides(fill=guide_legend(title=legend))

  return (df2.plot)
}


minage.plot <- genplot(d_data, "minage", "Minimum Drinking Age")
zerotol.plot <- genplot(d_data, "zerotol", "Zero Tolerance Law")
bac10.plot <- genplot(d_data, "bac10", "BAC 10")
bac08.plot <- genplot(d_data, "bac08", "BAC 08")

grid.arrange(zerotol.plot, minage.plot, bac08.plot, bac10.plot,top=textGrob("Alcohol related la
```




```
# License laws
gdl.plot <- genplot(d_data, "gdl", "Grad. Drivers License Law")
perse.plot <- genplot(d_data, "perse", "Per Se Law")

grid.arrange(gdl.plot, perse.plot, top=textGrob("License laws",gp=gpar(fontsize=12,font=3)))
```




```
# seatbelt laws
sbprim.plot <- genplot(d_data, "sbprim", "Primary Seatbelt Law")
sbsecon.plot <- genplot(d_data, "sbsecon",  "Secondary Seatbelt Law")

grid.arrange(sbprim.plot, sbsecon.plot, top=textGrob("License laws",gp=gpar(fontsize=12,font=3)
```




```
### Factors that probably dont change much over time for each state

# unemployment rate (convert to 0-1 scale?)
#hist(d_data$unem)

# percent population aged 14 through 24
#hist(d_data$perc14_24)

# normalize?
#hist(d_data$vehicmilespc)

# state population
```

```
#summary(d_data$statepop)
#hist(d_data$statepop)
```

Part 2

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.

What is the average of this variable in each of the years in the time period covered in this dataset

```
# yearly average nationwide
year_avg
```

```
## # A tibble: 25 x 2
##     year avg_totfatrte
##    <int>         <dbl>
##  1  1980          25.5
##  2  1981          23.7
##  3  1982          20.9
##  4  1983          20.2
##  5  1984          20.3
##  6  1985          19.9
##  7  1986          20.8
##  8  1987          20.8
##  9  1988          20.9
## 10  1989          19.8
## # ... with 15 more rows
```

Regression model and explanation

This model gives us the time effect on crime rate. The intercept in this case is the average *totfatrte* across all states in the omitted year 2004. Each of the coefficients `d80`, `d81`…`d04` is the average increase in *totfatrte* relative to the base year 2004.

```
stargazer(lm(totfatrte~d80+d81+d82+d83+d84+d85+d86+d87+d88+d89+d90+d91+d92+d93+d94+d95+d96+d97+
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              totfatrte
## -----------------------------------------------
## d80                           8.766***
##                                (1.226)
```

```
##
## d81                          6.941***
##                               (1.226)
##
## d82                          4.214***
##                               (1.226)
##
## d83                          3.424***
##                               (1.226)
##
## d84                          3.539***
##                               (1.226)
##
## d85                           3.123**
##                               (1.226)
##
## d86                          4.071***
##                               (1.226)
##
## d87                          4.046***
##                               (1.226)
##
## d88                          4.163***
##                               (1.226)
##
## d89                           3.043**
##                               (1.226)
##
## d90                           2.776**
##                               (1.226)
##
## d91                            1.366
##                               (1.226)
##
## d92                            0.429
##                               (1.226)
##
## d93                            0.399
##                               (1.226)
##
## d94                            0.426
##                               (1.226)
##
## d95                            0.940
##                               (1.226)
##
## d96                            0.640
##                               (1.226)
```

```
##
## d97                              0.882
##                                 (1.226)
##
## d98                              0.536
##                                 (1.226)
##
## d99                              0.521
##                                 (1.226)
##
## d00                              0.097
##                                 (1.226)
##
## d01                              0.064
##                                 (1.226)
##
## d02                              0.301
##                                 (1.226)
##
## d03                              0.035
##                                 (1.226)
##
## Constant                       16.729***
##                                 (0.867)
##
## ------------------------------------------------
## Observations                     1,200
## R2                               0.128
## Adjusted R2                      0.110
## Residual Std. Error      6.008 (df = 1175)
## F Statistic          7.164*** (df = 24; 1175)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Did driving become safer

This model clearly highlights that driving has become safer between the years 1980 through 1990 since the differences are significant at the < 0.05 level up until year 1990

Part 3

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.)

Variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl and perc14_24* are yes-no indicator dummies, with the caveat that they can be fractional as noted in the problem statement. The fractional values will ideally need to be changed to 0 or 1 based on whether the variable is < 0.5 or $>= 0.5$. Note that we need to do special handling of the edge case of 0.5 in two categories.

Transformation is needed for variables `unem` and `perc14_24`

Part 4

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?

Part 5

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

Part 6

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.

#Part 7

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?