W271 Lab3

Question 1

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

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

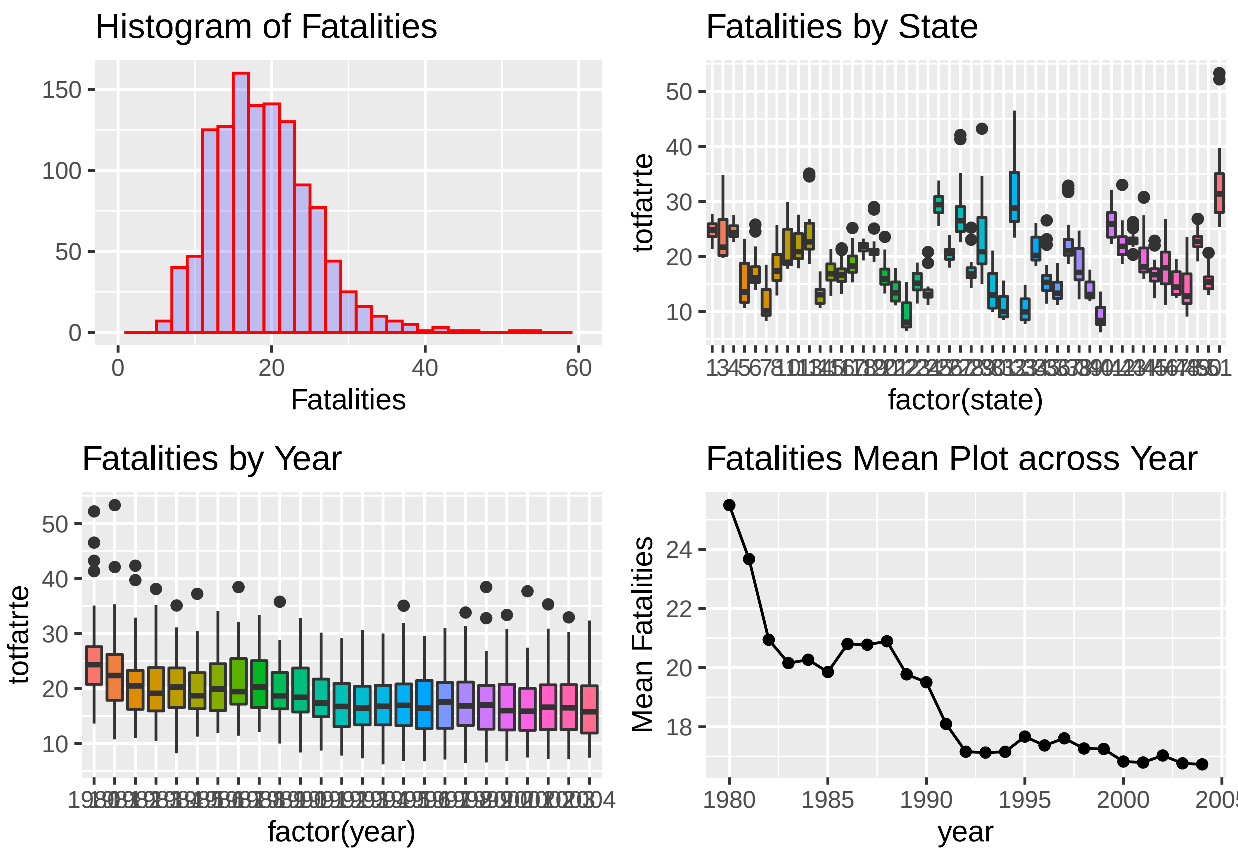
```
table(data$state)
```

```
##
##  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
```

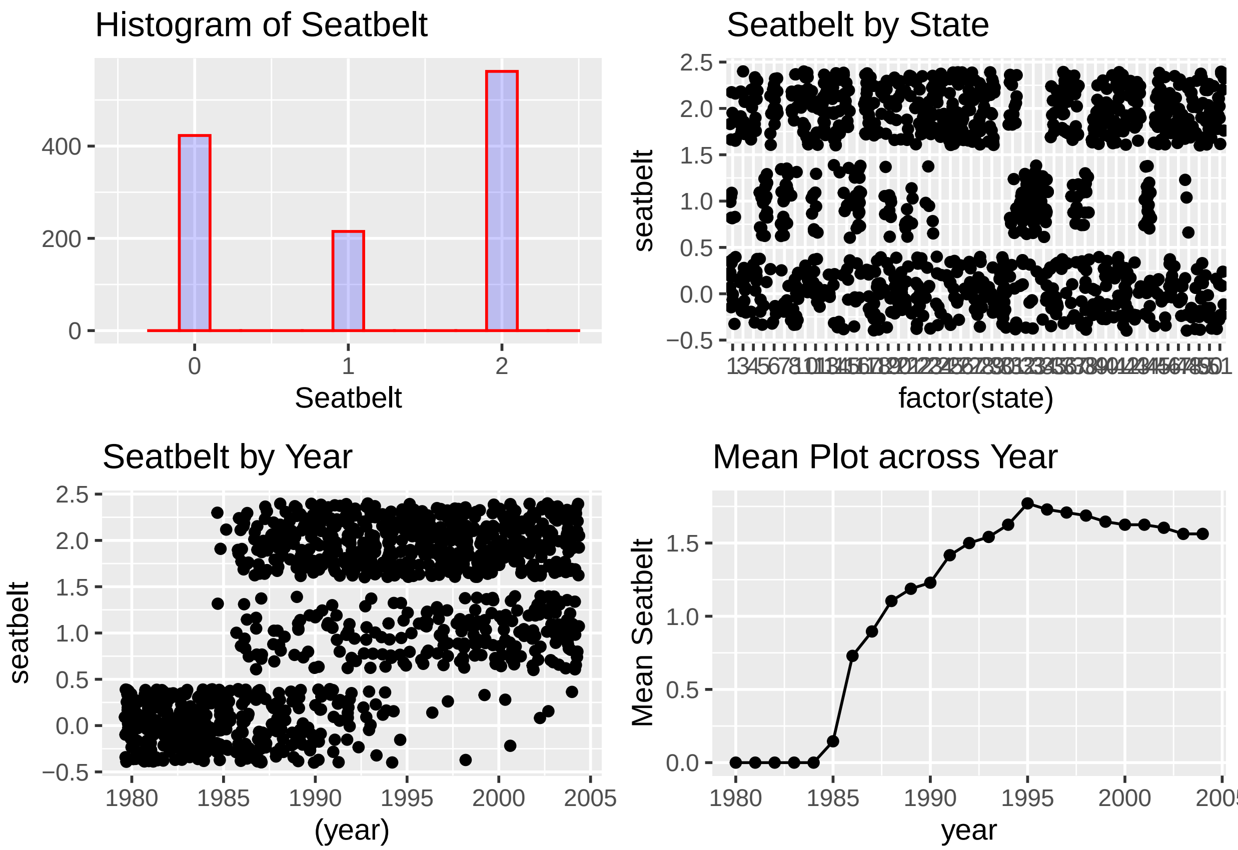
There are no missing data and this is a balanced panel, 25 years observations for each state. We will proceed for subsequent EDA.

```
# dependent variable, totfatrte
p1<-qplot(data$totfatrte,geom="histogram",binwidth =2,main = "Histogram of Fatalities", xlab =
p2 <- ggplot(data, aes(factor(state), totfatrte))+geom_boxplot(aes(fill = factor(state)), show
p3 <- ggplot(data, aes(factor(year), totfatrte))+geom_boxplot(aes(fill = factor(year)), show.le
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(totfatrte))%>%ggplot(aes(x=year, y=mean
grid.arrange(p1,p2,p3,p4,nrow=2)
```

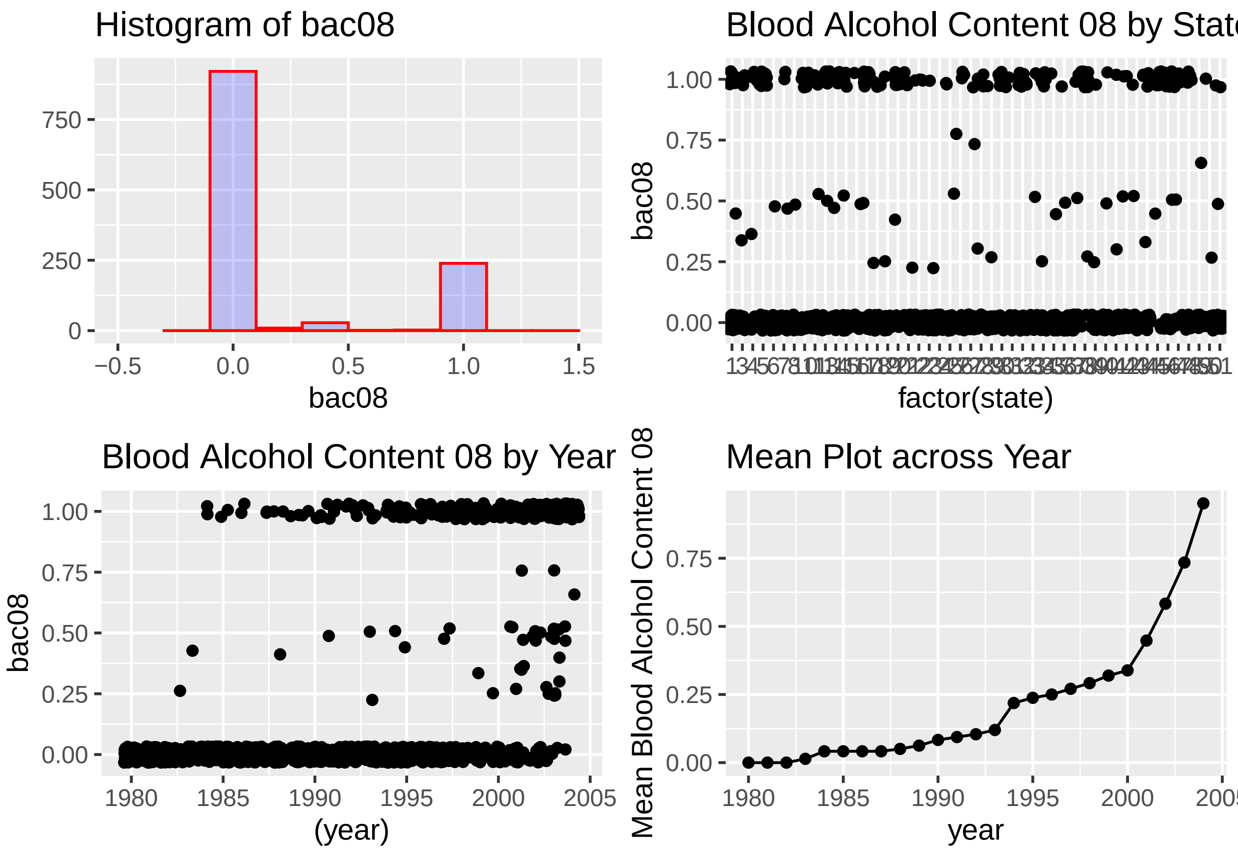
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



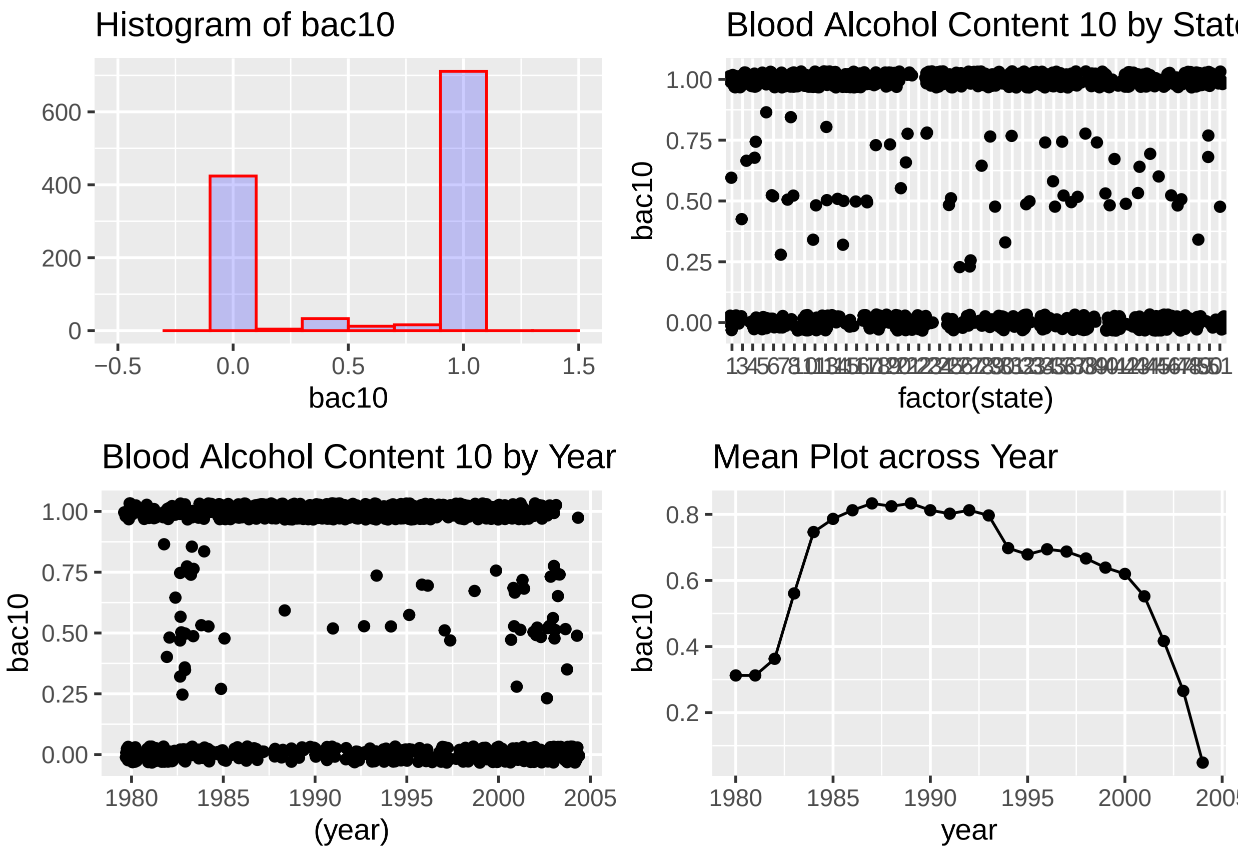

```
# seatbelt
p1<-qplot(data$seatbelt,geom="histogram",binwidth =0.2,main = "Histogram of Seatbelt", xlab = '
p2<-ggplot(data, aes(factor(state), seatbelt))+geom_jitter()+ggtitle("Seatbelt by State") + the
p3<-ggplot(data, aes((year), seatbelt))+geom_jitter() +ggtitle("Seatbelt by Year") + theme(plot
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(seatbelt))%>%ggplot(aes(x=year, y=mean_
grid.arrange(p1,p2,p3,p4,nrow=2)
```




```
# bac08
p1<-qplot(data$bac08,geom="histogram",binwidth =0.2,main = "Histogram of bac08", xlab = "bac08
p2<-ggplot(data, aes(factor(state), bac08))+geom_jitter()+ggtitle("Blood Alcohol Content 08 by
p3<-ggplot(data, aes((year), bac08))+geom_jitter()+ggtitle("Blood Alcohol Content 08 by Year")+
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(bac08))%>%ggplot(aes(x=year, y=mean_gr
grid.arrange(p1,p2,p3,p4,nrow=2)
```

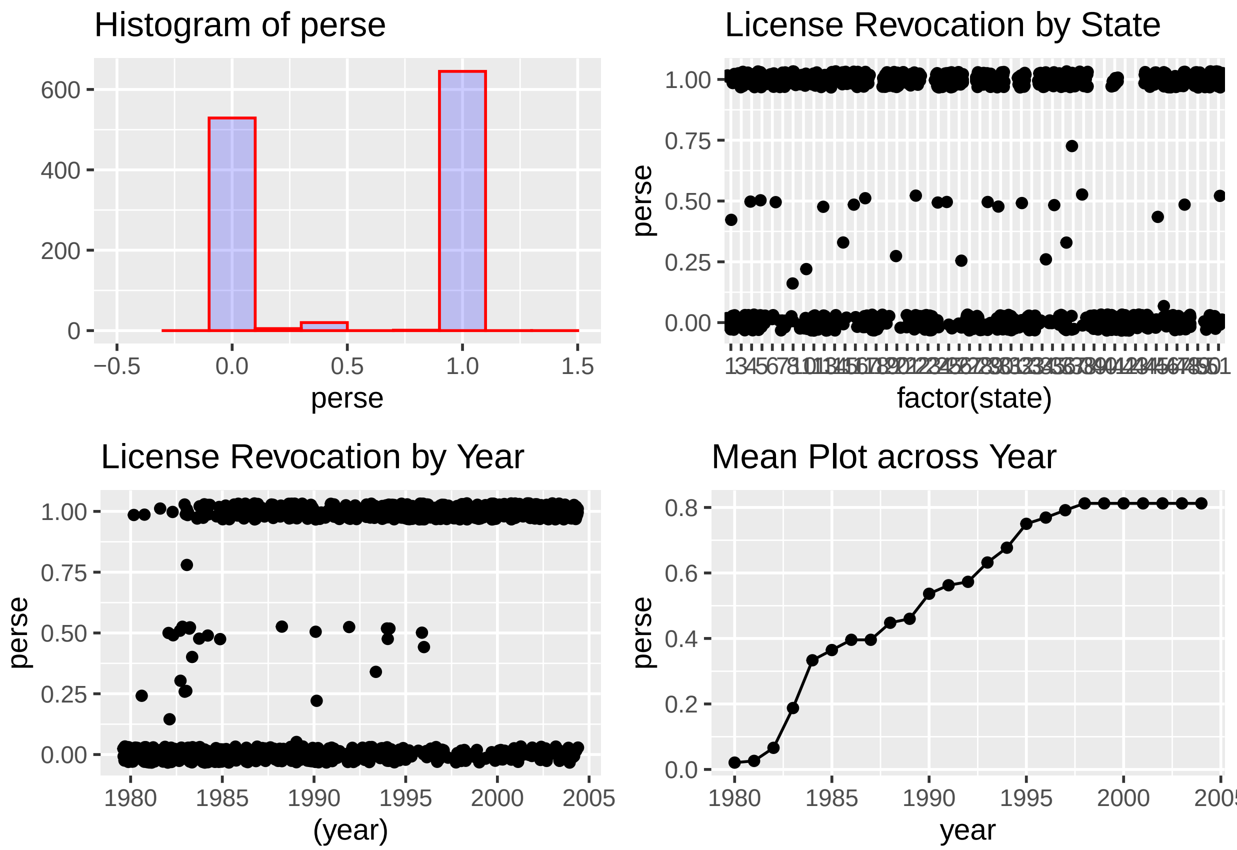



```
# bac10
p1<-qplot(data$bac10,geom="histogram",binwidth =0.2,main = "Histogram of bac10", xlab = "bac10"
p2<-ggplot(data, aes(factor(state), bac10))+geom_jitter()+ggtitle("Blood Alcohol Content 10 by
p3<-ggplot(data, aes((year), bac10))+geom_jitter()+ggtitle("Blood Alcohol Content 10 by Year")+
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(bac10))%>%ggplot(aes(x=year, y=mean_gro
grid.arrange(p1,p2,p3,p4,nrow=2)
```

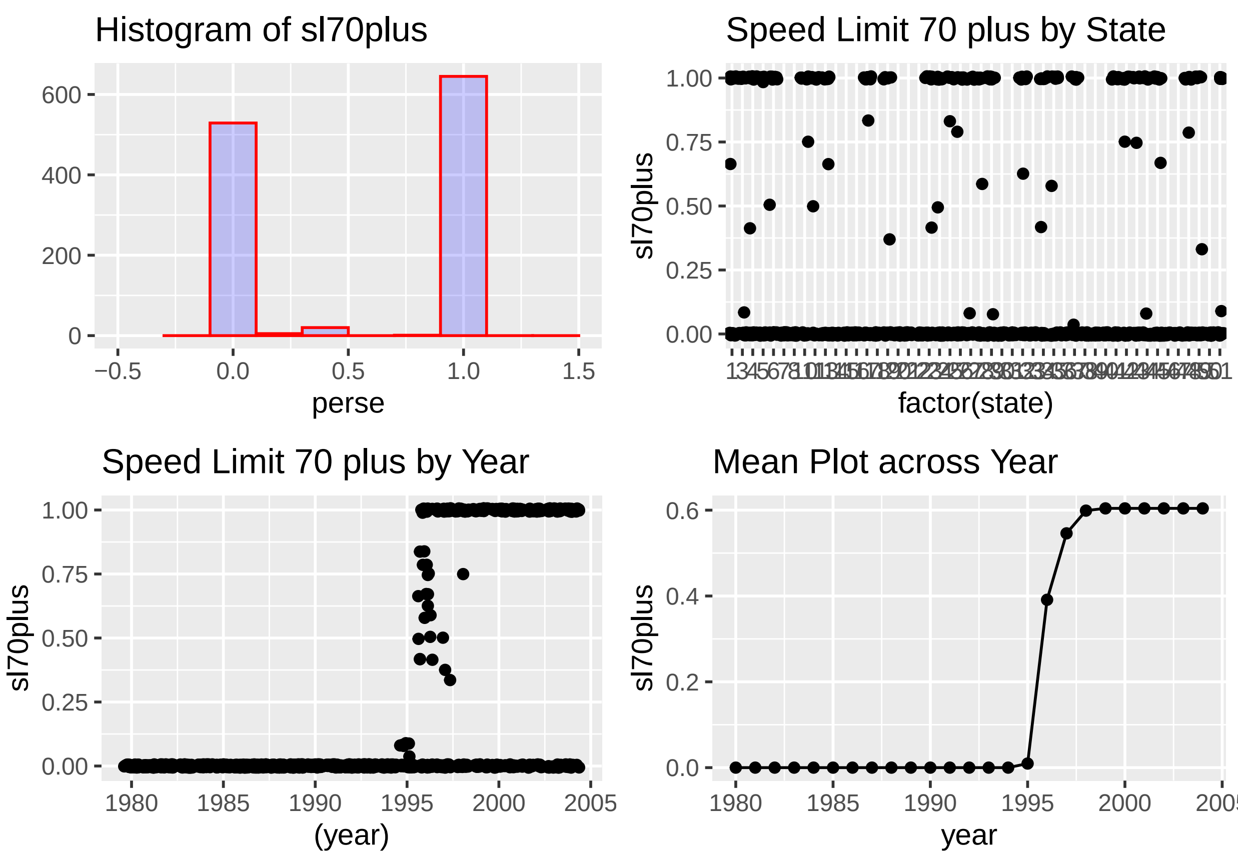



```r
# perse
p1<-qplot(data$perse,geom="histogram",binwidth =0.2,main = "Histogram of perse", xlab = "perse"
p2<-ggplot(data, aes(factor(state), perse))+geom_jitter()+ggtitle("License Revocation by State"
p3<-ggplot(data, aes((year), perse))+geom_jitter()+ggtitle("License Revocation by Year")+theme
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(perse))%>%ggplot(aes(x=year, y=mean_gro
grid.arrange(p1,p2,p3,p4,nrow=2)
```

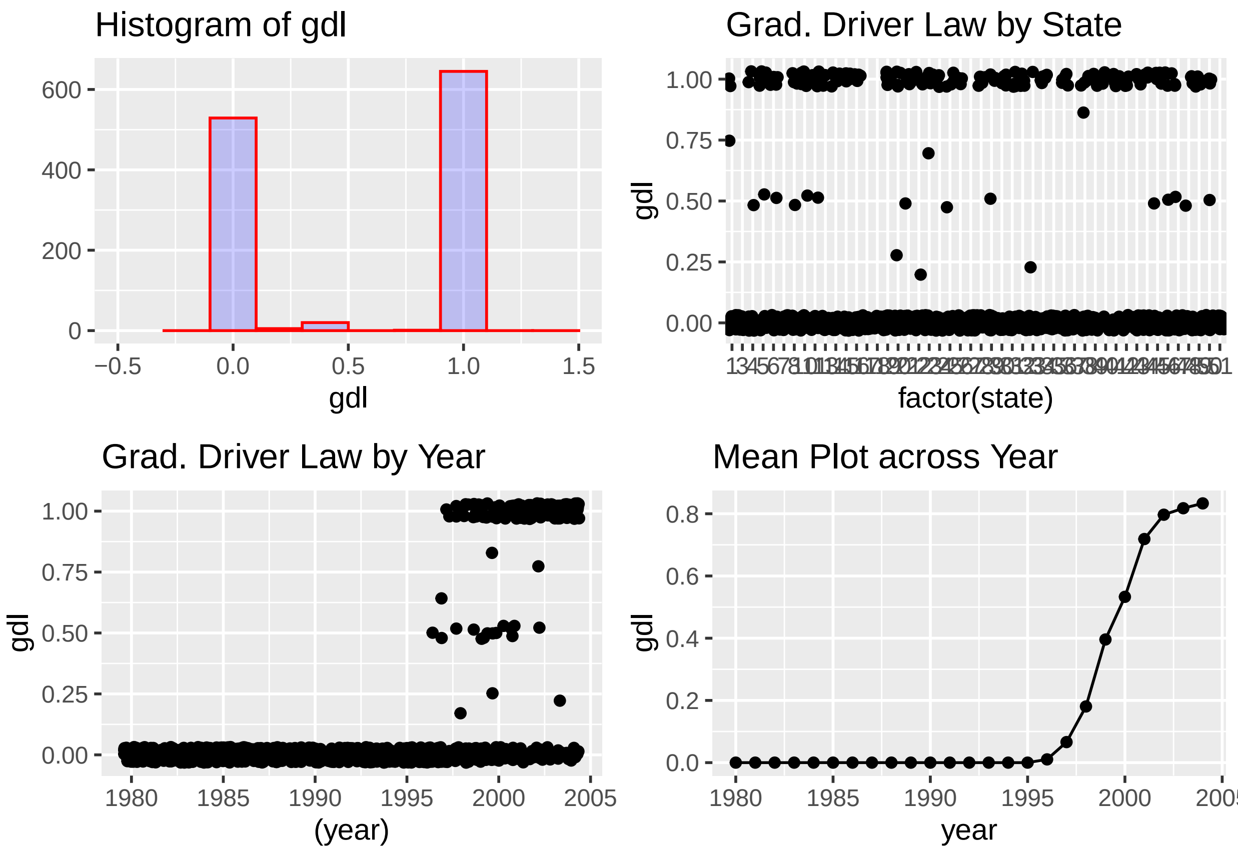
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



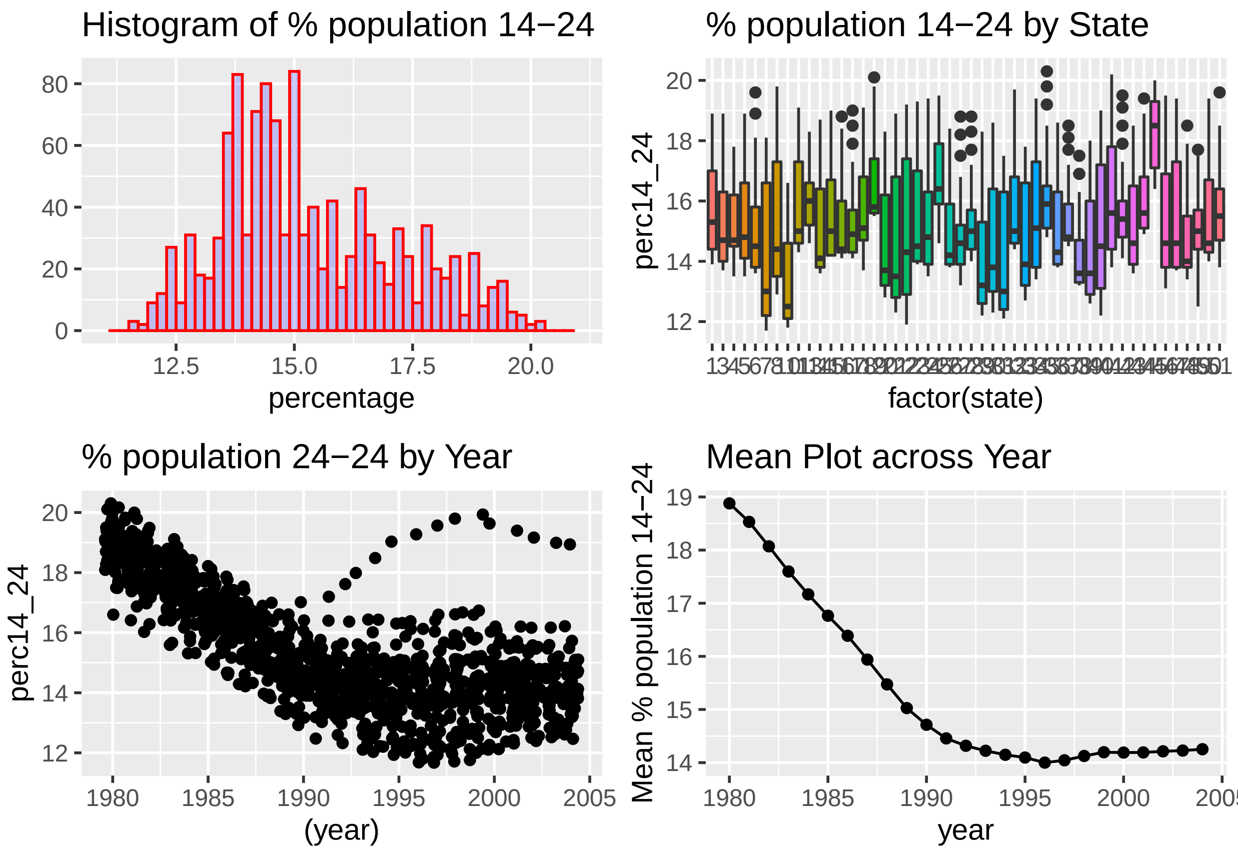

```
# sl70plus
p1<-qplot(data$perse,geom="histogram",binwidth =0.2,main = "Histogram of sl70plus", xlab = "pe
p2<-ggplot(data, aes(factor(state), sl70plus))+geom_jitter()+ggtitle("Speed Limit 70 plus by S
p3<-ggplot(data, aes((year), sl70plus))+geom_jitter()+ggtitle("Speed Limit 70 plus by Year")+t
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(sl70plus))%>%ggplot(aes(x=year, y=mean_
grid.arrange(p1,p2,p3,p4,nrow=2)
```



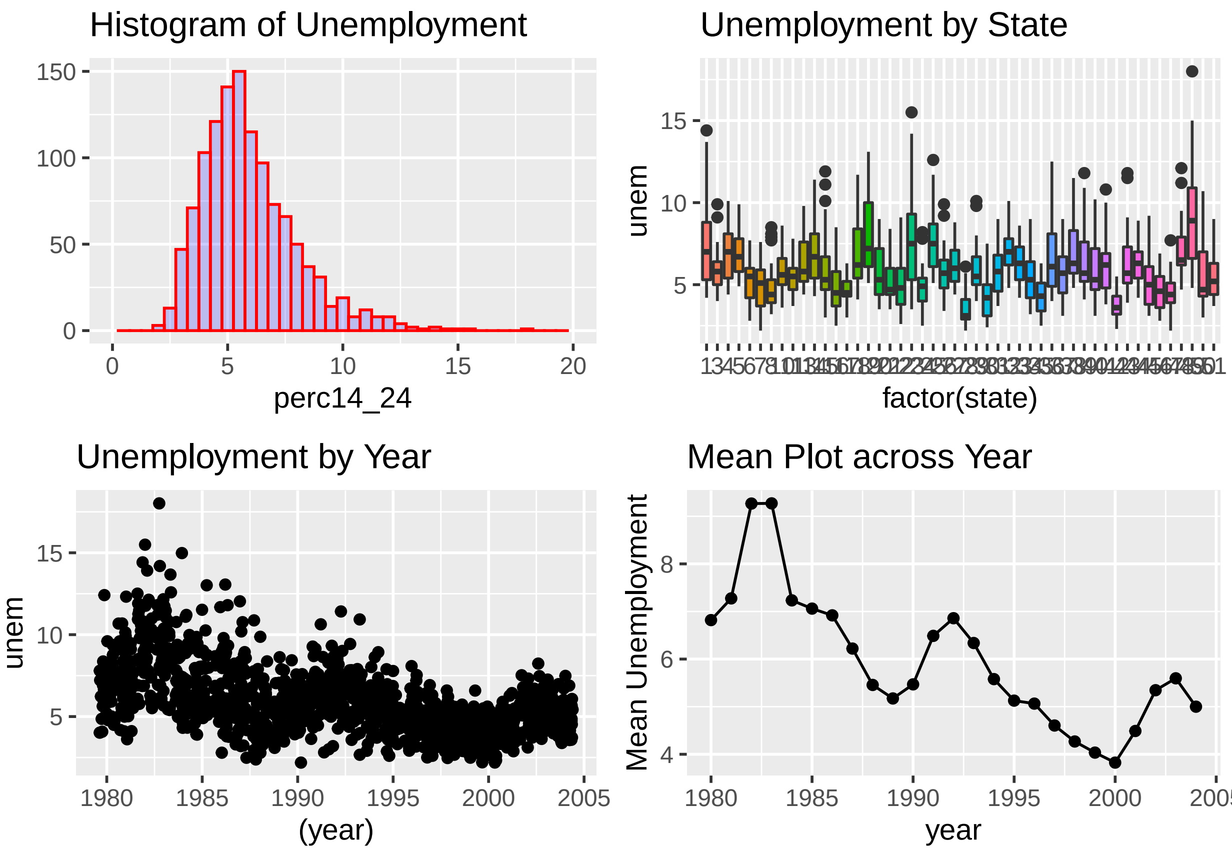

```
# gdl
p1<-qplot(data$perse,geom="histogram",binwidth =0.2,main = "Histogram of gdl", xlab = "gdl",fil
p2<-ggplot(data, aes(factor(state), gdl))+geom_jitter()+ggtitle("Grad. Driver Law by State")+t
p3<-ggplot(data, aes((year), gdl))+geom_jitter()+ggtitle("Grad. Driver Law by Year")+theme(plo
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(gdl))%>%ggplot(aes(x=year, y=mean_grou
grid.arrange(p1,p2,p3,p4,nrow=2)
```



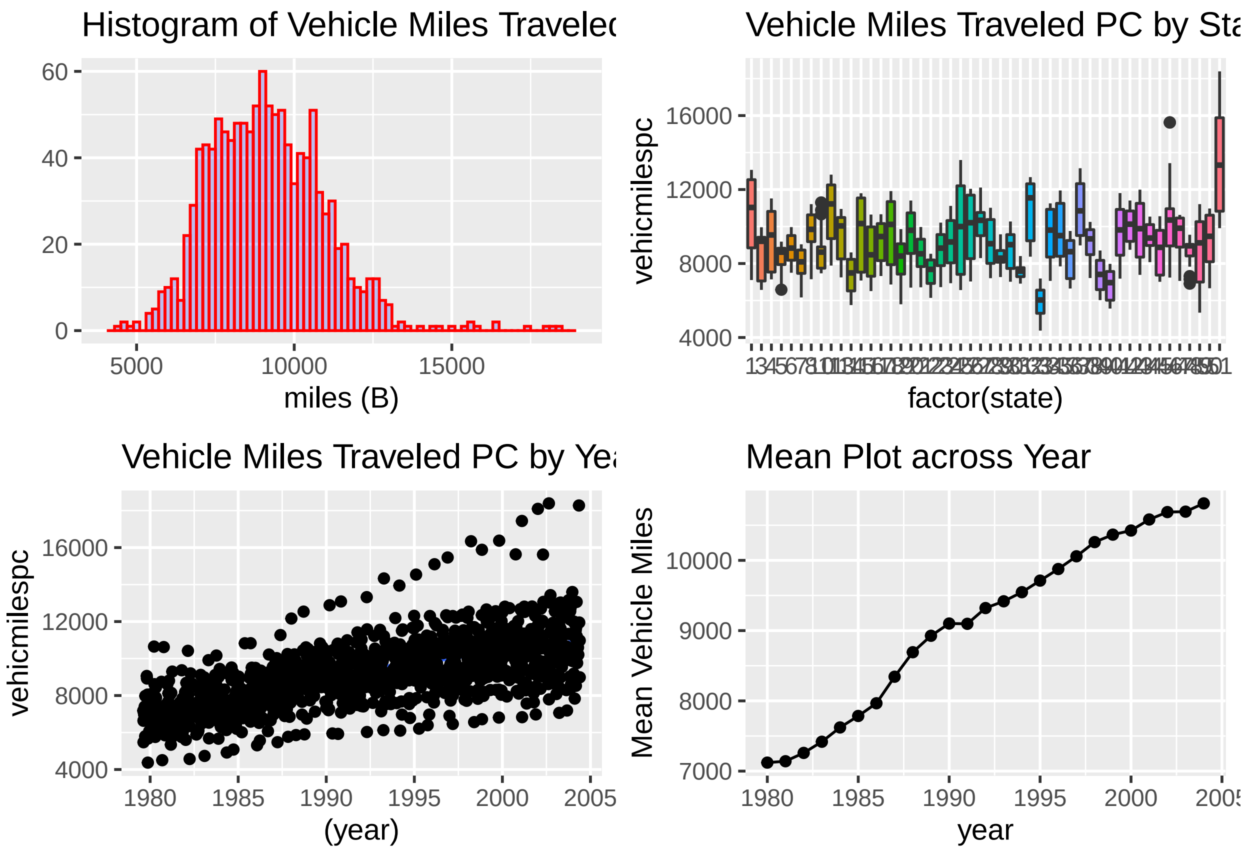

```
p1<-qplot(data$perc14_24,geom="histogram",binwidth = 0.2,main = "Histogram of % population 14-2
p2<-ggplot(data, aes(factor(state), perc14_24))+geom_boxplot(aes(fill = factor(state)),show.leg
p3<-ggplot(data, aes((year), perc14_24))+geom_jitter()+ggtitle("% population 24-24 by Year")+tl
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(perc14_24))%>%ggplot(aes(x=year, y=mear
grid.arrange(p1,p2,p3,p4,nrow=2)
```




```
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=
p3<-ggplot(data, aes((year), unem))+geom_jitter()+ggtitle("Unemployment by Year")+theme(plot.ti
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(unem))%>%ggplot(aes(x=year, y=mean_grou
grid.arrange(p1,p2,p3,p4,nrow=2)
```




```r
# vehicmilespc
p1<-qplot(data$vehicmilespc,geom="histogram",binwidth = 200,main = "Histogram of Vehicle Miles
p2<-ggplot(data, aes(factor(state),vehicmilespc))+geom_boxplot(aes(fill=factor(state)),show.leg
p3<-ggplot(data, aes((year), vehicmilespc))+geom_smooth(method='gam',formula=y~s(x,bs="cs"))+ge
p4<-data %>% group_by(year)%>%summarise(mean_group=mean(vehicmilespc))%>%ggplot(aes(x=year, y=m
grid.arrange(p1,p2,p3,p4,nrow=2)
```



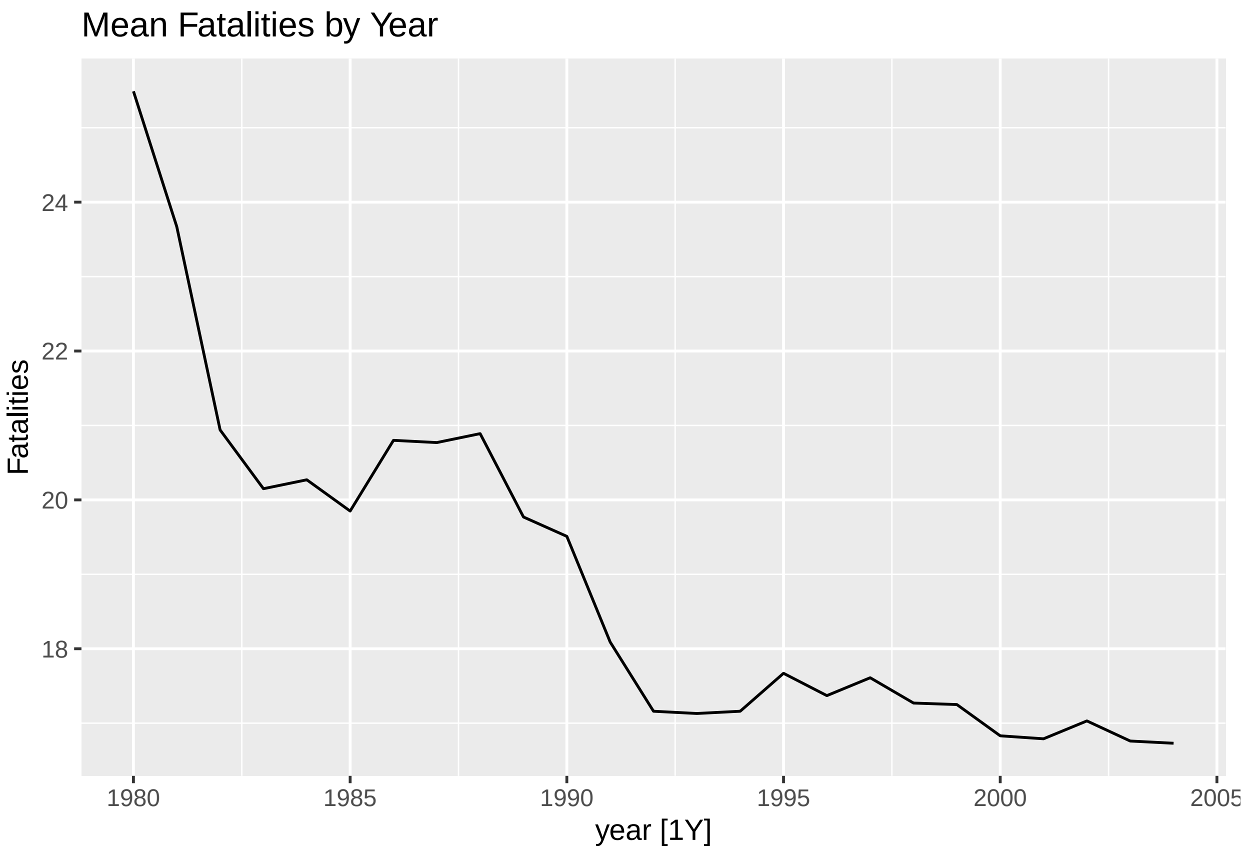
Question 2

Variable *totfatrte* is defined as total number of fatalities in 100,000 population.

```
byYear.mean <- aggregate(data, by=list(data$year), FUN=mean)
mean.totfatrte.df = round(data.frame(year=1980:2004, mean.totfatrte=byYear.mean$totfatrte), 2)
t(mean.totfatrte.df)
```

```
##                   [,1]    [,2]    [,3]    [,4]    [,5]    [,6]   [,7]    [,8]
## year           1980.00 1981.00 1982.00 1983.00 1984.00 1985.00 1986.0 1987.00
## mean.totfatrte   25.49   23.67   20.94   20.15   20.27   19.85   20.8   20.77
##                   [,9]   [,10]   [,11]   [,12]   [,13]   [,14]   [,15]   [,16]
## year           1988.00 1989.00 1990.00 1991.00 1992.00 1993.00 1994.00 1995.00
## mean.totfatrte   20.89   19.77   19.51   18.09   17.16   17.13   17.16   17.67
##                  [,17]   [,18]   [,19]   [,20]   [,21]   [,22]   [,23]   [,24]
## year           1996.00 1997.00 1998.00 1999.00 2000.00 2001.00 2002.00 2003.00
## mean.totfatrte   17.37   17.61   17.27   17.25   16.83   16.79   17.03   16.76
##                  [,25]
## year           2004.00
## mean.totfatrte   16.73
```

```r
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.

```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 show that total fatalities is decreasing over time and it is statistically significant. Driving became safer over time.

Question 3

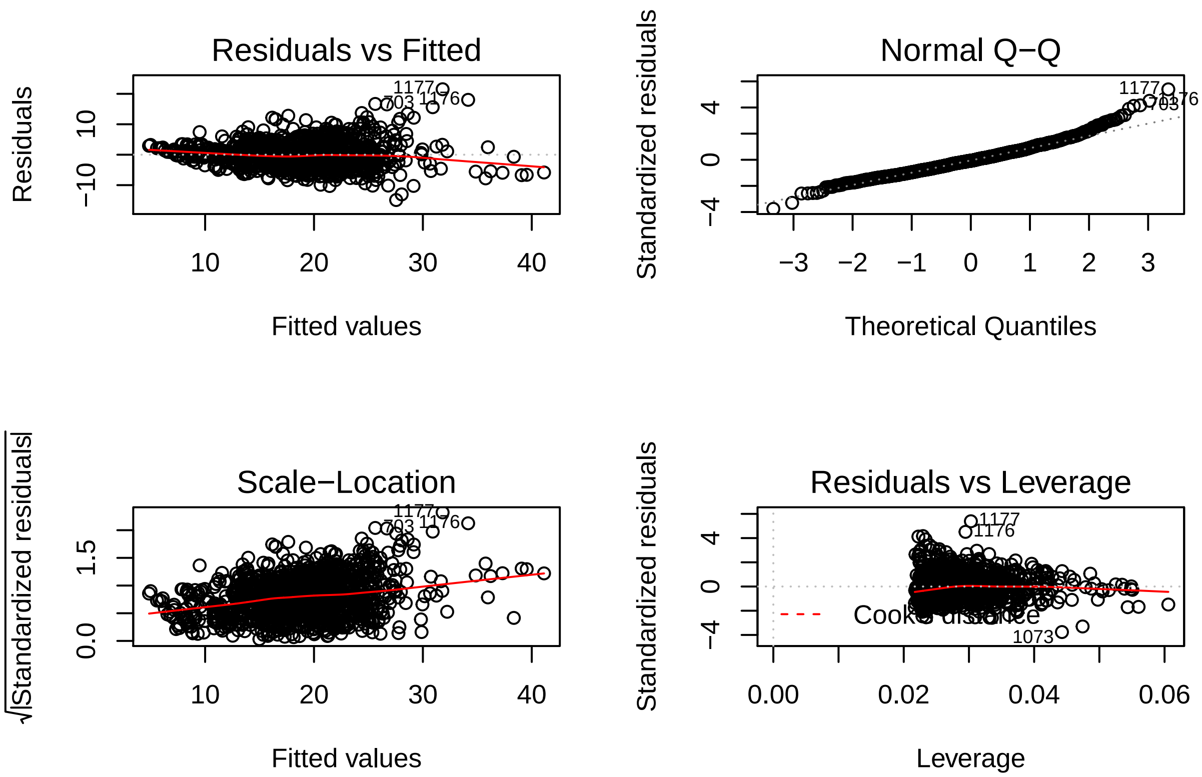
Variables bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl are supposed to be binary variables. But due to the fact that some states implemented the law in middle of year, some of the these variables have values between 0 and 1. For correct modeling of binary variables, we need all values to be 0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

```
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.lm2 <- lm(totfatrte ~ factor(year)+bac08+bac10+perse+sbprim+sbsecon+sl70plus+gdl+perc14_24
summary(fit.lm2)
```

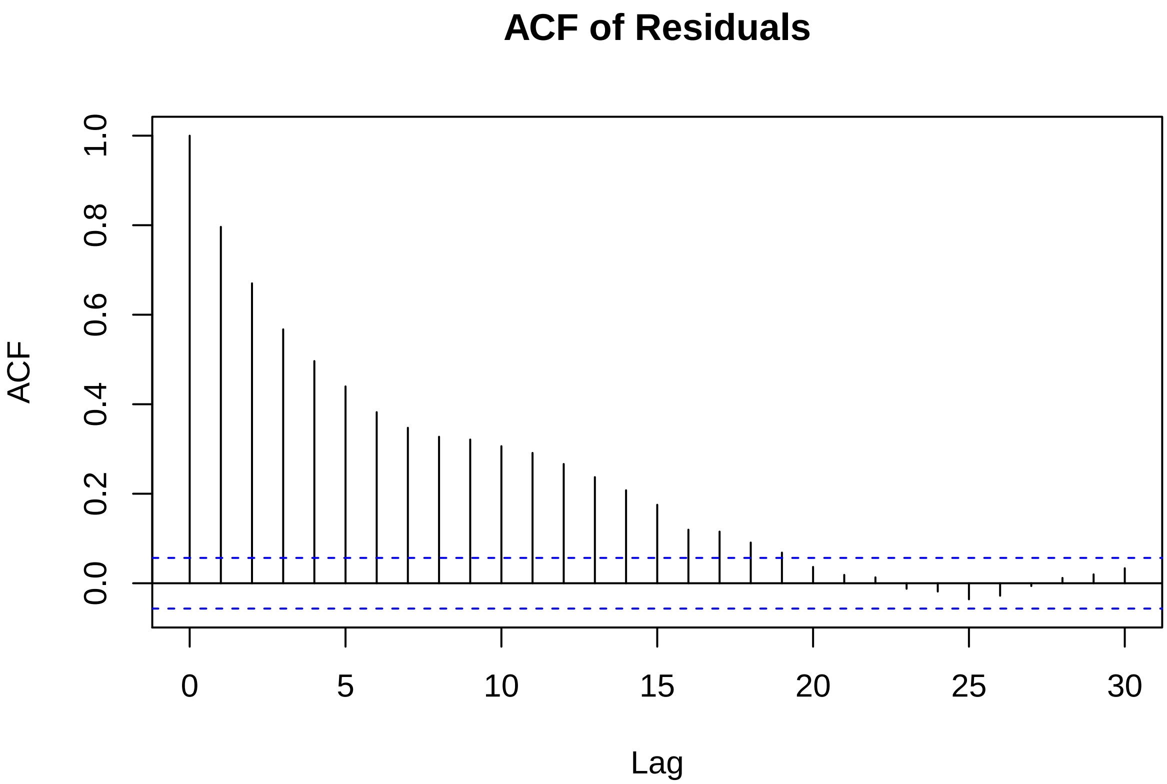
```
##
## Call:
## lm(formula = totfatrte ~ factor(year) + bac08 + bac10 + perse +
##     sbprim + sbsecon + sl70plus + gdl + perc14_24 + unem + vehicmilespc,
##     data = data.round)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8962  -2.7265  -0.3033   2.3323  21.5064
##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

```
par(mfrow=c(2,2));plot(fit.lm2);
```




```
par(mfrow=c(1,1));acf(fit.lm2$residuals, main="ACF of Residuals");
```



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 has impact on 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 not empirical evidence that primary seat belt law has impact on fatalities.

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 is unobserved fixed effects. Serial correlations and heteroskedasticity on residuals suggest the test statistics in pooled OLS result are not valid.

Question 4

```
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+vehicmi
                 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 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 see that 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+vehicm
                  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, we can reject null hypothesis that random effect model is preferred. Fixed Effect model should be chosen for our analysis.

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.

Question 7

Estimators are not efficient. All statistical inference are not valid. If unobserved effect is uncorrelated with all explanatory variables, estimators are consistent, otherwise estimators are not consistent.