# Econometrics Final Project

# Car Fatalities

1. **Description about the project:**

The data covers vehicle mortality across all the states of the US along with various other macro metrics such as population, government laws varying across states.

2. **Description about the data**
3. **Data summary**

   The data is for the years 1982 to 1988 for all the states of USA. There are in total 336 observations across all the states and years. This is a balanced Panel data. Below is the summary of the columns which we have used in the analysis.

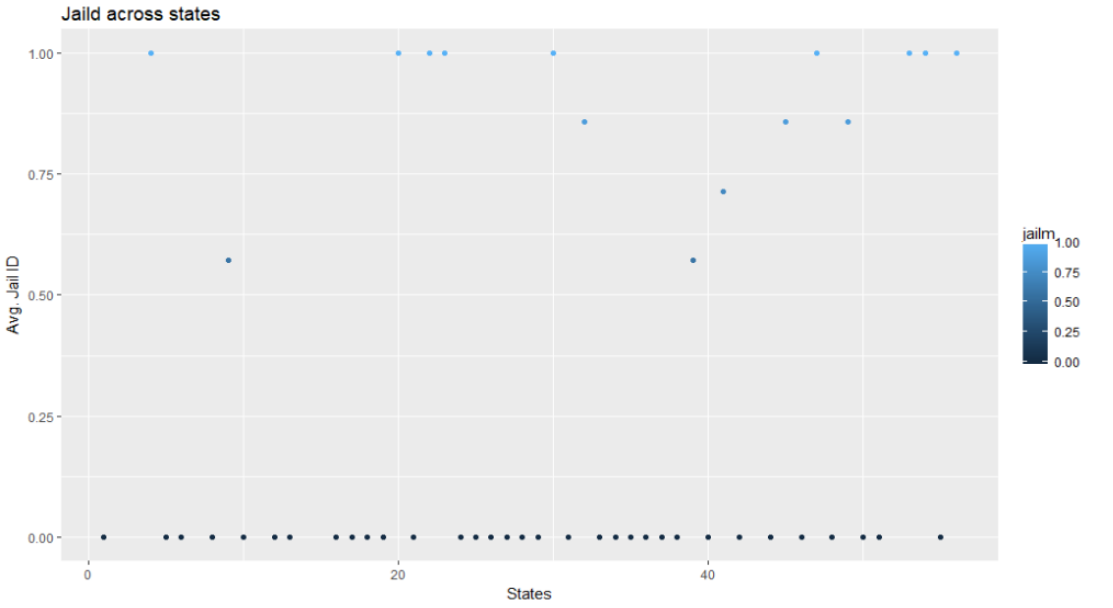
| | state | spircons | unrate | perinc | beertax | mlda | dry | yngdrv | vmiles | jaild | comserd | allmort | mrall | mraidall | pop | miles | gspch |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| nobs | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 | 336.0 |
| NAs | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Minimum | 1.0 | 0.8 | 2.4 | 9513.8 | 0.0 | 18.0 | 0.0 | 0.1 | 4576.3 | 0.0 | 0.0 | 79.0 | 0.0 | 0.0 | 478999.7 | 3993.0 | -0.1 |
| Maximum | 56.0 | 4.9 | 18.0 | 22193.5 | 2.7 | 21.0 | 45.8 | 0.3 | 26148.3 | 1.0 | 1.0 | 5504.0 | 0.0 | 0.0 | 28314028.0 | 241575.0 | 0.1 |
| 1. Quartile | 18.8 | 1.3 | 5.5 | 12085.8 | 0.2 | 20.0 | 0.0 | 0.2 | 7182.5 | 0.0 | 0.0 | 293.8 | 0.0 | 0.0 | 1545251.5 | 11691.5 | 0.0 |
| 3. Quartile | 42.5 | 2.0 | 8.9 | 15175.1 | 0.7 | 21.0 | 2.4 | 0.2 | 8504.0 | 1.0 | 0.0 | 1063.5 | 0.0 | 0.0 | 5751734.9 | 44139.8 | 0.1 |
| Mean | 30.2 | 1.8 | 7.3 | 13880.2 | 0.5 | 20.5 | 4.3 | 0.2 | 7890.8 | 0.3 | 0.2 | 928.7 | 0.0 | 0.0 | 4930271.5 | 37101.5 | 0.0 |
| Median | 30.5 | 1.7 | 7.0 | 13763.1 | 0.4 | 21.0 | 0.1 | 0.2 | 7796.2 | 0.0 | 0.0 | 701.0 | 0.0 | 0.0 | 3310503.3 | 28483.5 | 0.0 |
| Sum | 10143.0 | 589.2 | 2468.5 | 4663742.0 | 172.5 | 6873.1 | 1433.7 | 62.5 | 2651293.2 | 94.0 | 62.0 | 312031.0 | 0.1 | 0.0 | 1656571224.5 | 12466101.0 | 8.5 |
| SE Mean | 0.8 | 0.0 | 0.1 | 122.9 | 0.0 | 0.0 | 0.5 | 0.0 | 80.5 | 0.0 | 0.0 | 51.0 | 0.0 | 0.0 | 276793.2 | 2043.3 | 0.0 |
| LCL Mean | 28.5 | 1.7 | 7.1 | 13638.4 | 0.5 | 20.4 | 3.2 | 0.2 | 7732.4 | 0.2 | 0.1 | 828.4 | 0.0 | 0.0 | 4385799.7 | 33082.2 | 0.0 |
| UCL Mean | 31.8 | 1.8 | 7.6 | 14122.0 | 0.6 | 20.6 | 5.3 | 0.2 | 8049.1 | 0.3 | 0.2 | 1028.9 | 0.0 | 0.0 | 5474743.3 | 41120.8 | 0.0 |
| Variance | 234.4 | 0.5 | 6.4 | 5076217.6 | 0.2 | 0.8 | 90.3 | 0.0 | 2177568.7 | 0.2 | 0.2 | 872452.2 | 0.0 | 0.0 | 25742471181917.2 | 1402829514.3 | 0.0 |
| Stdev | 15.3 | 0.7 | 2.5 | 2253.0 | 0.5 | 0.9 | 9.5 | 0.0 | 1475.7 | 0.4 | 0.4 | 934.1 | 0.0 | 0.0 | 5073703.9 | 37454.4 | 0.0 |
| Skewness | -0.1 | 2.2 | 0.7 | 0.7 | 2.2 | -1.3 | 2.7 | -0.1 | 5.6 | 1.0 | 1.6 | 2.5 | 0.7 | 1.3 | 2.2 | 2.5 | -0.7 |
| Kurtosis | -1.0 | 6.6 | 0.7 | 0.5 | 5.1 | 0.3 | 6.9 | 2.3 | 67.8 | -1.1 | 0.6 | 7.5 | 0.6 | 2.8 | 5.6 | 8.3 | 0.3 |

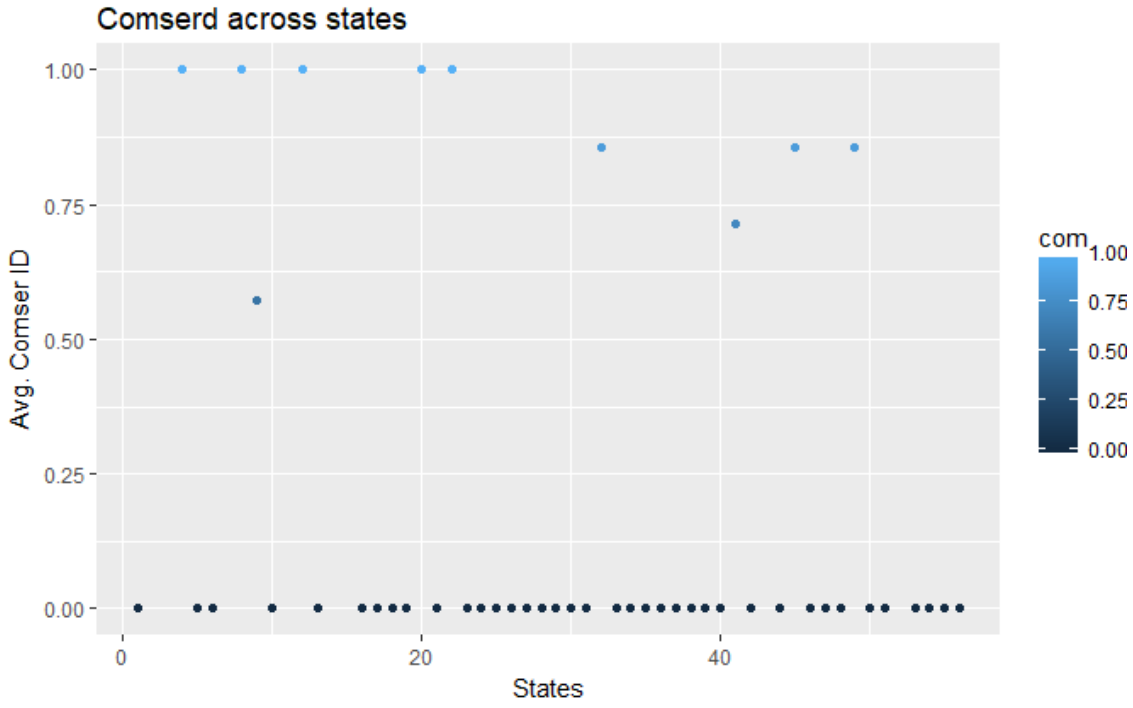
4. **Null data if any and the treatment used for it**

   Columns JailD and Comserd have null values. Both are categorical data.

   **Null Data Treatment:**

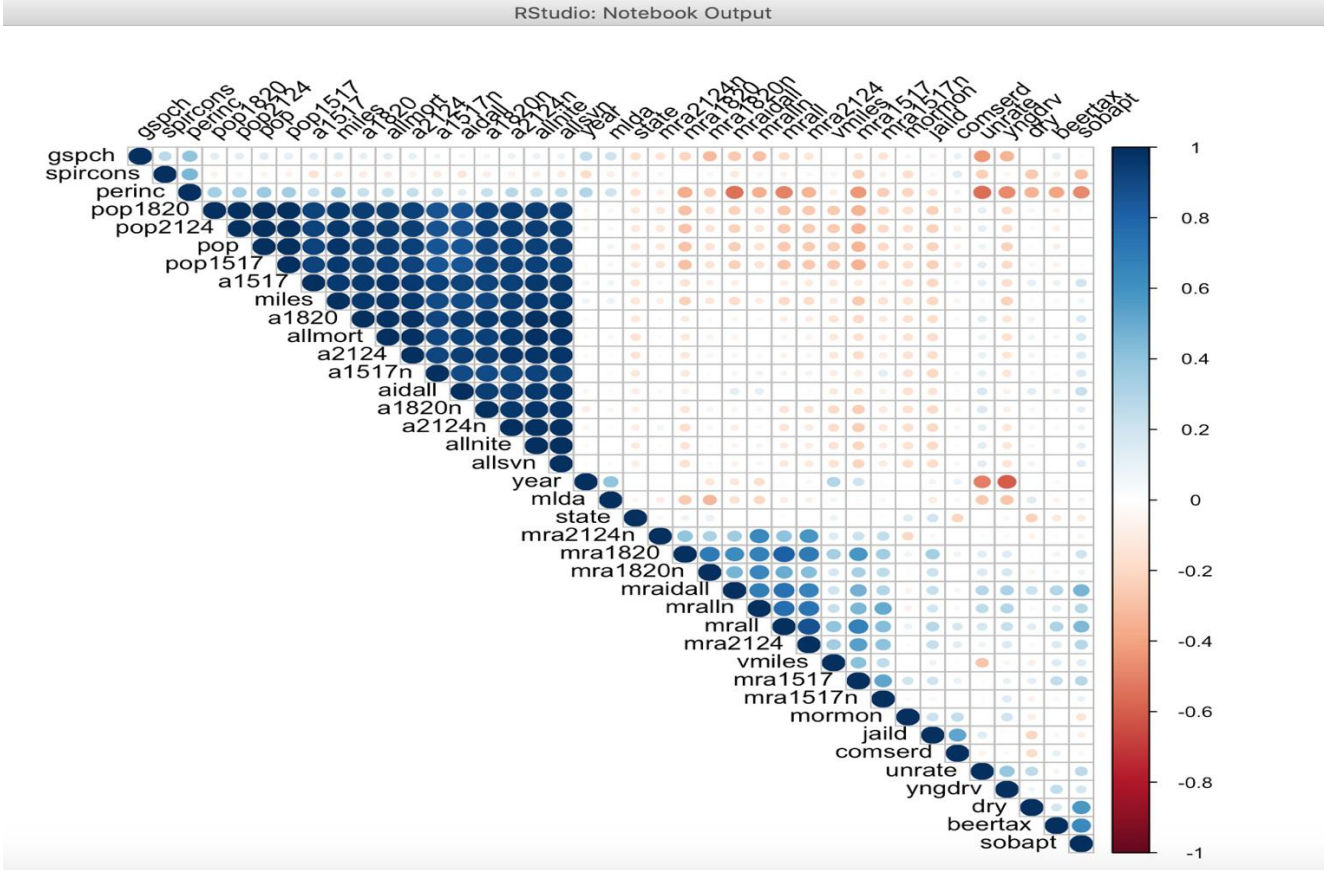
   Since, it is a categorical data, we have taken the mode of JailID and Comserd. On analyzing the data, we found that both Jaild and Comserd is a state characteristic. Both these columns are related to state laws and hence vary across states. Therefore, we took the mode at state level to fill the null values of jaild and comserd. In the below graph, it is evident that noth jaild and comserd are state specific.



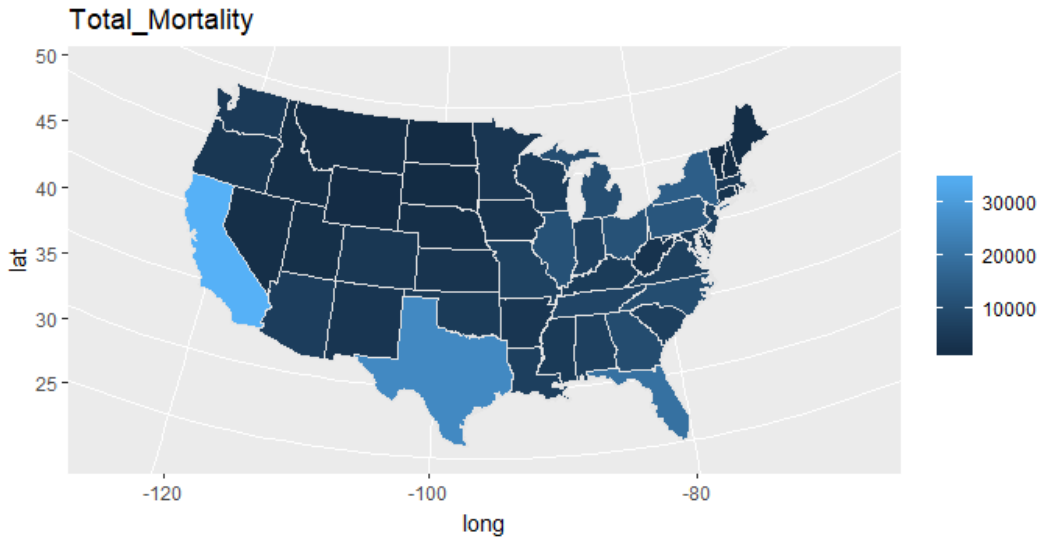



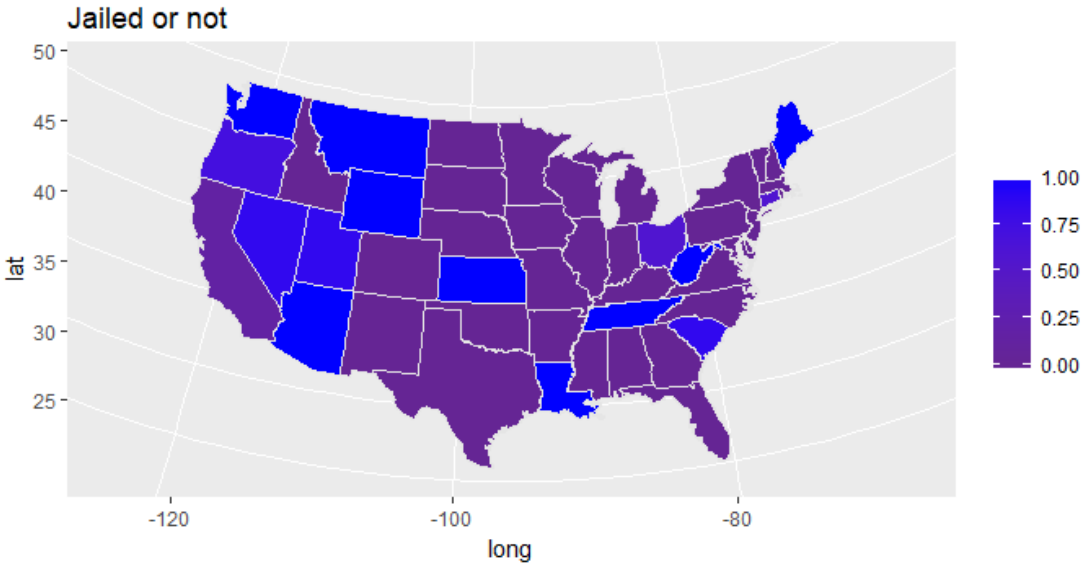

## 5. EDA in a flow

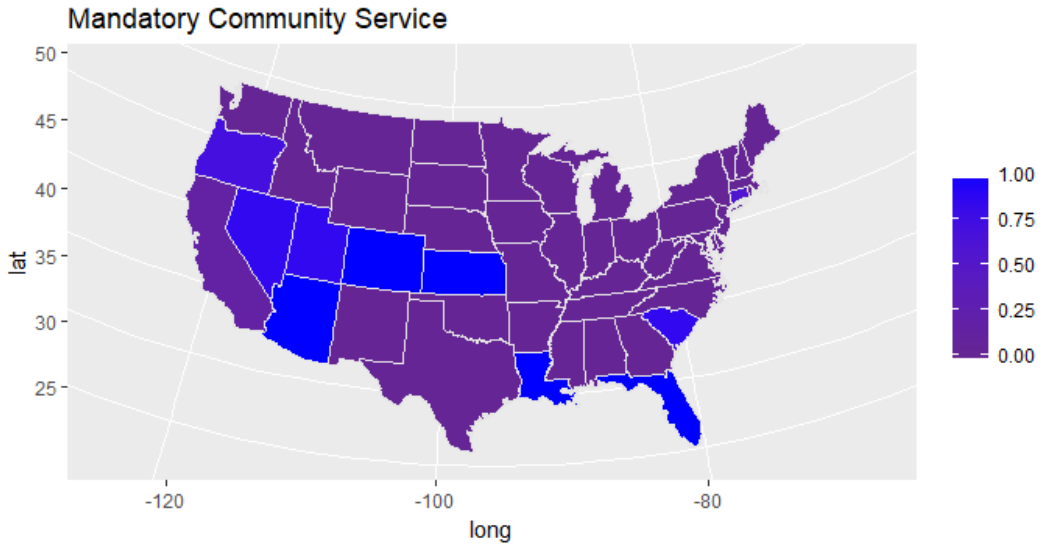
- We first try to understand the correlation between the variable of the data. Since there are 39 different variables, it is more efficient to first target the variables which have some correlation with the dependent variable which is mrall here. If the columns such as a1517n, a2124n do not have a good correlation with the dependent variable are used in the predictive model, it will unnecessarily reduce the efficiency and the $R^2$ value. Hence, we start with the columns which are correlated with mrall variable such as spircons, perinc, pop, mraidall etc. Below is the correlation graph.



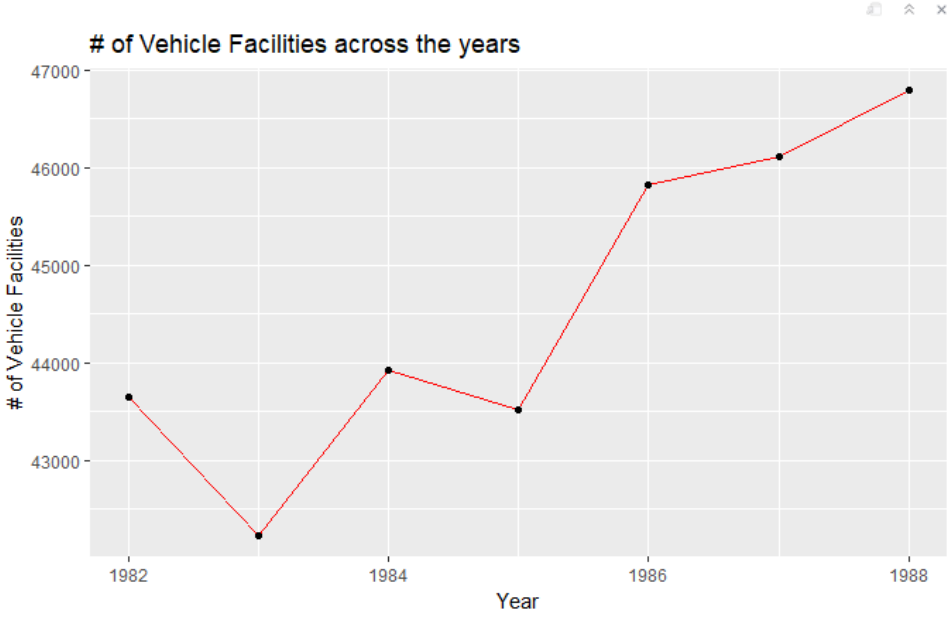

- States like Texas, California and Florida have the highest mortality across the US, but these states do not have the mandatory jail sentence imposed.


Total_Mortality
50
45
40
35
30
25
lat
-120
-100
-80
long
30000
20000
10000

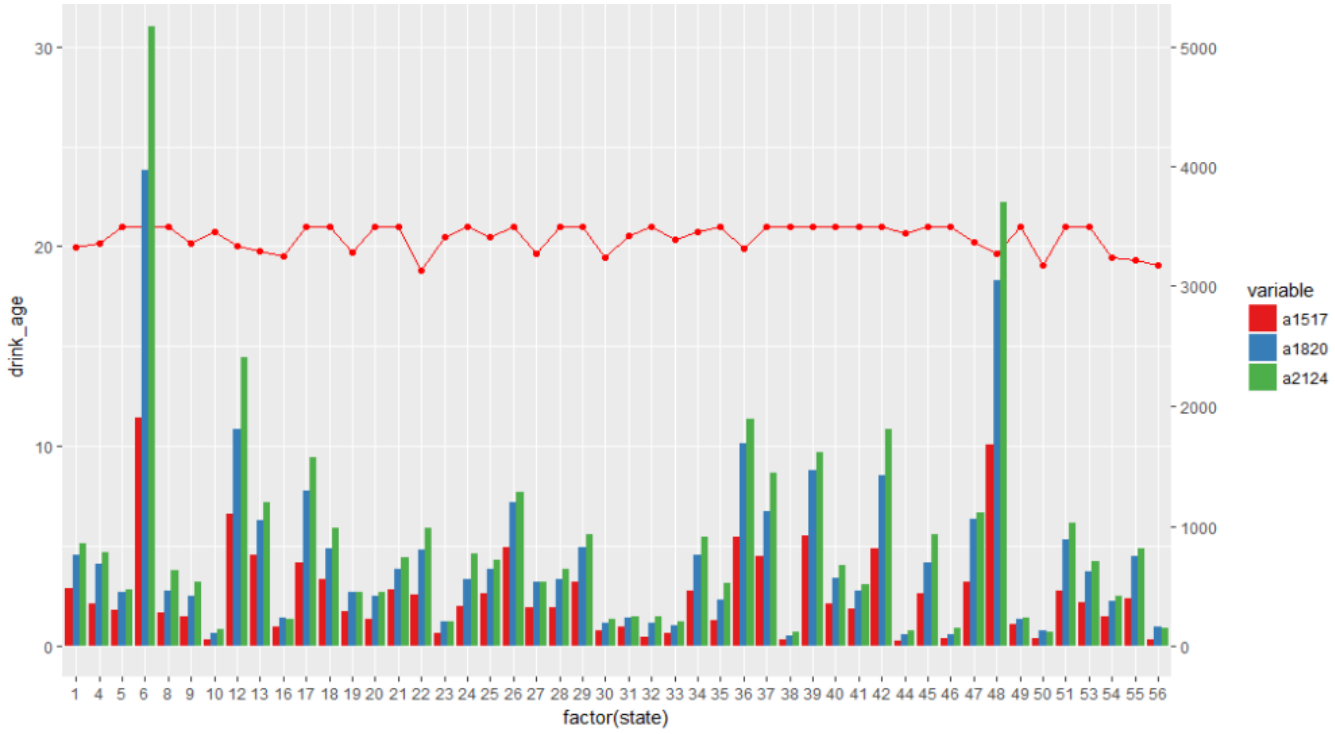

Jailed or not
50
45
40
35
30
25
lat
-120
-100
-80
long
1.00
0.75
0.50
0.25
0.00



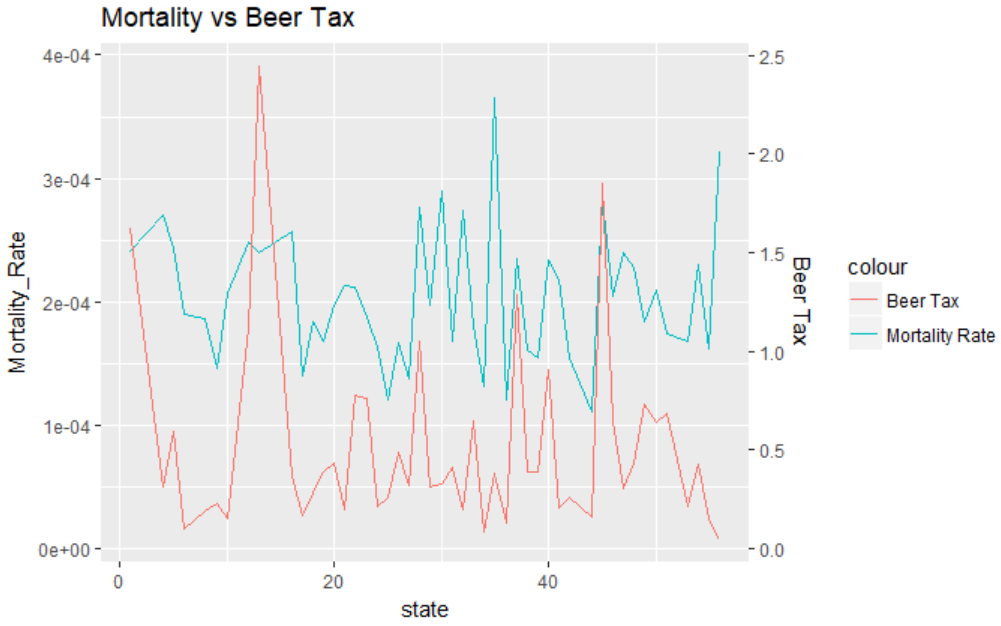

- The vehicle mortality keeps on increasing year over year. The vehicle mortality has increased by 7.21% from 1982 to 1988.




- In most of the states, the number of vehicle fatalities of people of ages 21-24 is comparatively higher than for ages 18-20. But in the states where the Minimum Legal Drinking Age (years) is low, the number of vehicle fatalities for the age 18-20 is almost equal to or more than the number of vehicle fatalities of people of ages 21-24.

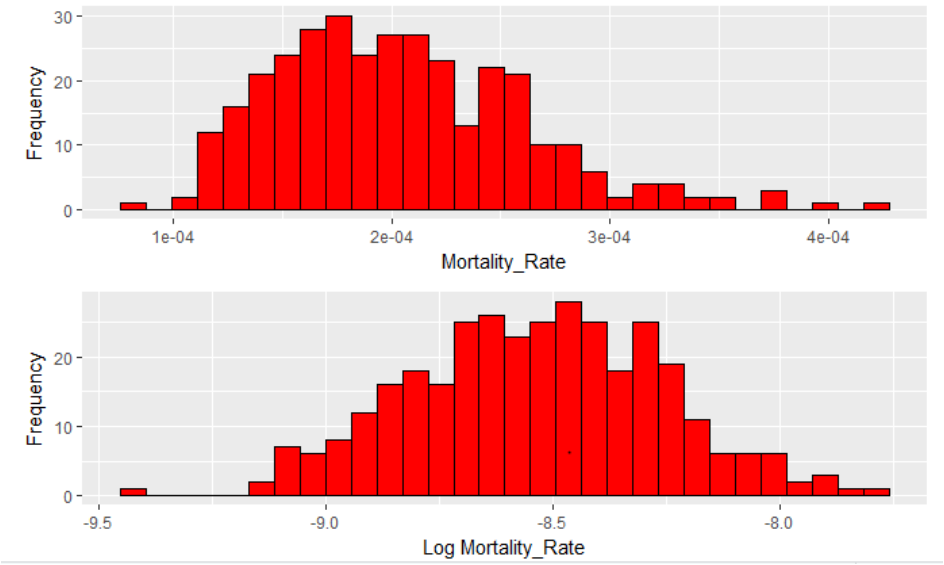



- Vehicle mortality is inversely proportional to beer tax.




**6. SKUness of data**

Vehicle mortality is rightly skewed. It does not follow a normal distribution. Hence, we have taken log while carrying out the predicting models.




7. **External data to get states from FIPS including the link**
   The data set car_fatality has data in state level, but the state names are not there, instead there is FIPS code. In order to get state names for using in creating maps in R, we have used the FIPS to state mapping from United States Censes Bureau. Below is the link for the mapping file.
   https://www.census.gov/geo/reference/ansi_statetables.html

8. **Small note about panel data**
9. **Running models and why not choosing or choosing–**
   Our objective is to predict the vehicle mortality rate. Being a panel data, we took a step by step approach. We first converted the data into a panel data with state id as the individual and the year column as year.
   Step 1: Based on our understanding of the data and the correlation plot, we first regressed a pooled OLS using, individual fixed effects and individual and time fixed effects. The dependent variable is mrall and the explanatory variables were spircons, unrate, perinc, beertax, mlda, dry, yngdrv, jaild, comserd, mraidall, vmiles.

   **Pooled OLS:**
   Here, the relationship between beer_tax and mrall is positive. This is due to the fact that pooled OLs does not take into account the unobserved heterogeneity. There is some omitted variable, which is highly correlated to beer_tax and is causing the covariance between beer_tax and the error term to be non-zero, and hence having an upward bias on the beta value of beer_tax. Same is the case with jaild. The estimator of jaild suggests that the states with mandatory jail sentence have 2.15% more mortality rate than the states where there is no mandatory jail sentence. This is also upwardly biased.

```
Call:
plm(formula = log(mrall) ~ spircons + unrate + perinc + beertax +
    mlda + dry + yngdrv + jaild + comserd + mraidall + vmiles,
    data = data1, model = "pooling")

Balanced Panel: n = 48, T = 7, N = 336

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.8733901 -0.0724623  0.0038625  0.0929127  0.4062250

Coefficients:
               Estimate  Std. Error  t-value  Pr(>|t|)
(Intercept) -9.1218e+00  2.7872e-01 -32.7282 < 2.2e-16 ***
spircons     4.0153e-02  1.4822e-02   2.7090  0.007107 **
unrate       2.1947e-03  4.6142e-03   0.4756  0.634649
perinc      -3.1626e-05  6.5790e-06  -4.8071 2.348e-06 ***
beertax      2.8877e-02  2.0707e-02   1.3946  0.164094
mlda         1.0024e-02  1.0463e-02   0.9580  0.338752
dry          4.3136e-04  1.0572e-03   0.4080  0.683524
yngdrv      -5.6549e-01  4.1833e-01  -1.3518  0.177391
jaild        2.1575e-02  2.4947e-02   0.8648  0.387785
comserd      1.0846e-01  2.7432e-02   3.9537 9.454e-05 ***
mraidall     5.7029e+03  4.1793e+02  13.6454 < 2.2e-16 ***
vmiles       5.3255e-05  6.5100e-06   8.1805 6.505e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    25.52
Residual Sum of Squares: 7.8034
R-Squared:      0.69423
Adj. R-Squared: 0.68385
F-statistic: 66.874 on 11 and 324 DF, p-value: < 2.22e-16
```

**Individual Fixed Effects:**

Here, we can see that both beer_tax and jaild have negative effects on vehicle mortality rate indicating that the unobserved heterogeneity has been taken care off.

```
Oneway (individual) effect Within Model

Call:
plm(formula = log(mrall) ~ spircons + unrate + perinc + beertax +
    mlda + dry + yngdrv + jaild + comserd + mraidall + vmiles,
    data = data1, effect = "individual", model = "within",
    index = c("state", "year"))

Balanced Panel: n = 48, T = 7, N = 336

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.2628252 -0.0306159  0.0017729  0.0295694  0.2441817

Coefficients:
            Estimate  Std. Error t-value  Pr(>|t|)
spircons  2.5301e-01  4.3102e-02  5.8700 1.243e-08 ***
unrate   -1.1504e-02  3.9995e-03 -2.8762  0.004337 **
perinc    4.0387e-05  8.8276e-06  4.5751 7.188e-06 ***
beertax  -9.1905e-02  7.0201e-02 -1.3092  0.191565
mlda      5.3521e-03  7.4573e-03  0.7177  0.473544
dry       1.0285e-02  5.4572e-03  1.8847  0.060522 .
yngdrv   -9.7282e-02  3.1511e-01 -0.3087  0.757761
jaild    -6.3817e-02  5.0970e-02 -1.2520  0.211608
comserd   4.7212e-02  5.7924e-02  0.8151  0.415732
mraidall  2.4705e+03  3.1313e+02  7.8898 7.055e-14 ***
vmiles    7.0141e-06  3.7017e-06  1.8948  0.059155 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2.2297
Residual Sum of Squares: 1.208
R-Squared:      0.4582
Adj. R-Squared: 0.34475
F-statistic: 21.2959 on 11 and 277 DF, p-value: < 2.22e-16
```

Individual and Time fixed effects:

Here, we suspect that the variation is not only within individuals but also across time. Since, the mortality is increasing over time, hence we ran a model with both individual and timed fixed effects.

```
Twoways effects Within Model

Call:
plm(formula = log(mrall) ~ spircons + unrate + perinc + beertax +
    mlda + dry + yngdrv + jaild + comserd + mraidall, data = data1,
    effect = "twoways", model = "within", index = c("state",
        "year"))

Balanced Panel: n = 48, T = 7, N = 336

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.2502392 -0.0330407  0.0012325  0.0345102  0.2067303

Coefficients:
            Estimate  Std. Error t-value  Pr(>|t|)
spircons  2.9888e-01  5.0358e-02  5.9352 8.911e-09 ***
unrate   -1.9619e-02  4.6310e-03 -4.2365 3.111e-05 ***
perinc    3.2924e-05  8.9927e-06  3.6612 0.0003015 ***
beertax  -8.7521e-02  6.8402e-02 -1.2795 0.2018100
mlda      6.6858e-04  7.3573e-03  0.0909 0.9276600
dry       8.2526e-03  5.2841e-03  1.5618 0.1195030
yngdrv   -3.9902e-02  3.5980e-01 -0.1109 0.9117771
jaild    -3.1074e-02  5.0023e-02 -0.6212 0.5349889
comserd   2.6447e-02  5.6368e-02  0.4692 0.6393172
mraidall  2.0636e+03  3.1437e+02  6.5642 2.638e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2.1077
Residual Sum of Squares: 1.1057
R-Squared:      0.47542
Adj. R-Squared: 0.35392
F-statistic: 24.6515 on 10 and 272 DF, p-value: < 2.22e-16
```

In order to test for if years have an effect on the variation of mortality, we conducted and F test between models Individual Fixed Effects and Individual Fixed Effects with time as dummy variables. Based on the above regression results, we are using just pircons, unrate, perinc, dry, mraidall, vmiles, beertax columns.
Since the data is spanned out for years 1982 to 1988, we create dummy variable for years 1983 to 1988 in order to avoid exact collinearity.
Null Hypothesis: Ho : $\beta_{1982} = \beta_{1983} = \beta_{1984} = \beta_{1985} = \beta_{1986} = \beta_{1987} = \beta_{1988} = 0$
Alternative Hypothesis: $H_1$: Either one or all betas are non-zero

Since, our p value is 2.452e-05 which is <<< 0.05 (α), we reject the null hypothesis. Hence, year has significant effect on mortality.

```
        F test for individual effects

data:  log(mrall) ~ spircons + unrate + perinc + dry + mraidall + vmiles +  ...
F = 5.4419, df1 = 6, df2 = 275, p-value = 2.452e-05
alternative hypothesis: significant effects
```

Now that we have finalized on Individual and Time fixed model, we will get the significant columns.

1. While regressing against mlda, jaild , comserd , beertax, spircons, perinc, miles, gspch, unrate, we found that variables such as jaild, comserd,mlda, perincs, miles, gspch are not significant variables. Hence we removed them.

```
Twoways effects within Model

Call:
plm(formula = log(mrall) ~ mlda + jaild + comserd + beertax +
    spircons + perinc + miles + gspch + unrate, data = data2,
    effect = "twoways", model = "within")

Balanced Panel: n = 48, T = 7, N = 336

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-0.27371385 -0.03680932  0.00091437  0.03812998  0.19700784

Coefficients:
            Estimate  Std. Error t-value  Pr(>|t|)
mlda     -3.6188e-04  7.9068e-03 -0.0458 0.9635281
jaild     1.6160e-02  5.3217e-02  0.3037 0.7616200
comserd  -1.6495e-02  6.0284e-02 -0.2736 0.7845762
beertax  -1.2304e-01  7.3737e-02 -1.6686 0.0963334 .
spircons  3.6064e-01  5.3767e-02  6.7074 1.137e-10 ***
perinc    3.3025e-05  9.5762e-06  3.4487 0.0006524 ***
miles     7.9185e-07  9.4759e-07  0.8356 0.4040845
gspch     1.0887e-01  2.0106e-01  0.5415 0.5885985
unrate   -2.4535e-02  5.2947e-03 -4.6340 5.562e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2.1077
Residual Sum of Squares: 1.2868
R-Squared:      0.38951
Adj. R-Squared: 0.25086
F-statistic: 19.3533 on 9 and 273 DF, p-value: < 2.22e-16
```

Final ?????

Also, we see that laws such as mandatory jail service, mandatory community service do not have an impact on the vehicle mortality rate.