Capital Punishment without Conviction

Jail Deaths in America

Jonathan Brinkerhoff

Mathematics Department

Abstract

Reuters collected data on jail deaths for the largest jails throughout the United States. I analyzed this data under the assumption that liberal states perform better than conservative and poor states in guaranteeing the health and safety of their inmate populations. To this end, I selected California as a representative liberal state, Texas as a conservative state, Mississippi as a poor state, and Pennsylvania as a swing state. I used four metrics to compare the performance of these states: total deaths per year, the proportion of deaths relative to the average daily population per year, the proportion of suicides per year, and the proportion of illness-related deaths per year. Through exploratory data analysis and modeling, I found that it cannot be determined based on my selected states that liberal states guarantee better outcomes for their inmate populations than conservative states. In fact, California performed consistently worse than the other states with respect to my chosen metrics. Given that California has by far the largest inmate population, this suggests that jail or jail system size may be a better indicator of inmate well being than the nature of state government based on my subjective understanding. A more enlightening analysis might examine jails grouped by size.

Introduction

Jails are generally small facilities intended to hold persons accused of committing crimes but not released on bail, or persons serving short sentences for conviction of misdemeanor offenses. In their 2023 report, the Prison Policy Institute, a liberal think-tank, report that 1 in 3 incarcerated Americans are held in jails, and 80% of these inmates are not convicted of a crime [1]. Given that jails do not serve to punish or rehabilitate, they should be relatively benign institutions. It is concerning, therefore, when deaths occur in jails. Death in jail deprives the accused their right to a speedy trial as constitutionally guaranteed, and raises questions of human rights violations. At worst, in cases of abuse or neglect, deaths in jail amount to capital punishment without conviction.

Jail death data is compiled by the Justice Department, but not readily released to the public. through FOIA request, Reuters compiled jail death data for the largest jails in every state [2]. It is my opinion that conditions in jails are a reflection of the state's regard for constitutional and human rights, therefore I am interested in exploring via the Reuters data any relationship between the nature of a state's government and numbers of jail deaths in that state. I have selected several states as representative of certain manners of government. California is generally regarded as a fairly liberal state, and might be expected to treat it's inmate population with care, minimizing deaths. Texas is a conservative state, with an administration that is quite openly "tough on crime." Such a government might tolerate more deaths in its jails. Mississippi, and other deep southern states, is very poor, and may see more jail deaths simply due to a lack of resources to allocate to the establishment of humane conditions in it's jails. Finally, I have chosen Pennsylvania, a populous swing state that has experienced significant economic turbulence in recent decades, as a potentially interesting case. I intend to examine trends within and between the states I have selected with respect to death by type and deaths overall.

Data Overview

Reuters compiled data on 523 of the largest jails or jail systems throughout the United States, including 10 of the largest jails in each state; most of these are county jails. Between 2008 and 2019, 7,571 inmates died in these jails.

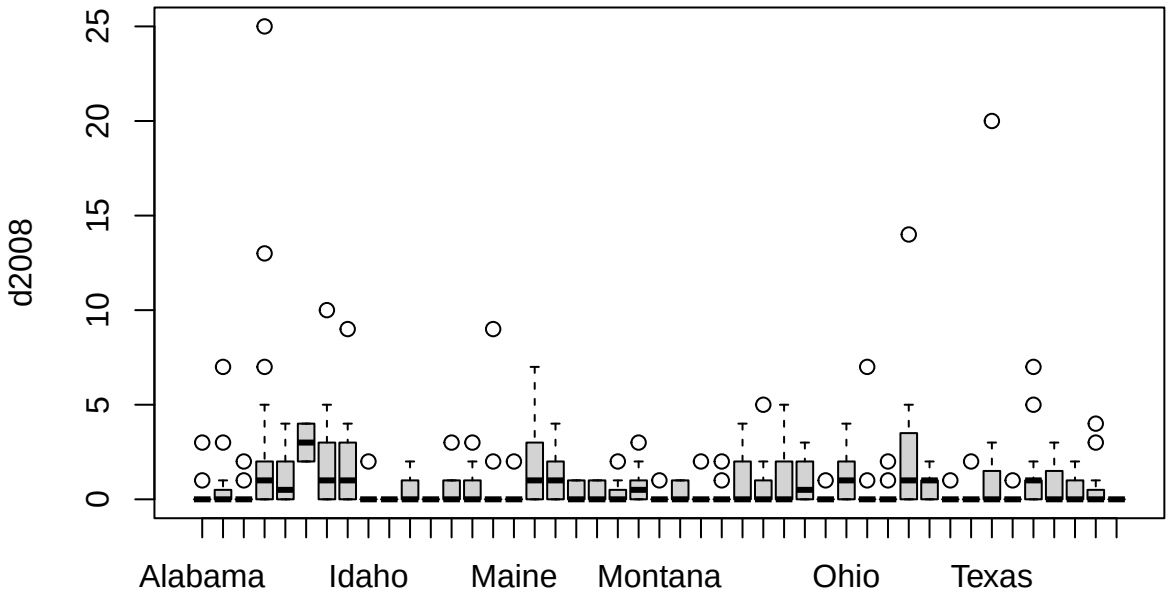
Reuters provides three csv files, a file called all_deaths which includes counts of deaths by jail for the years 2008 through 2019, all_jails which provides additional details on the individuals who died, and a codesheets file which provides details on the codes used in the other files. My analysis will make use of the all_deaths file exclusively. Information of particular interest to me are counts of death by type (illness, suicide, drug/alcohol, homicide, accident, other), total deaths, and the average jail populations.

There are missing values in the dataset. From a cursory examination, these values are missing due mainly to periodic reporting of deaths by particular jails. About four percent of the data is NA. Given that a small proportion of the data is missing, I am choosing to fill the NA values with zeroes to facilitate computation in R.

Of the states I selected, California has 38 jails and an average daily population of 38,368. Texas has 23 jails and an average daily population of 38,368. Mississippi has 11 jails and an average daily population of 3,996. Pennsylvania has 12 jails and an average daily population of 18,351. To facilitate data analysis, I have chosen to produce contingency tables for each state, with years as the row labels, and types of death as the column labels. For modeling, I produced a training set for the years 2008 through 2018 with columns for year, state, total deaths, suicide deaths, and illness-related deaths. My testing set included the last year, 2019.

Exploratory Data Analysis

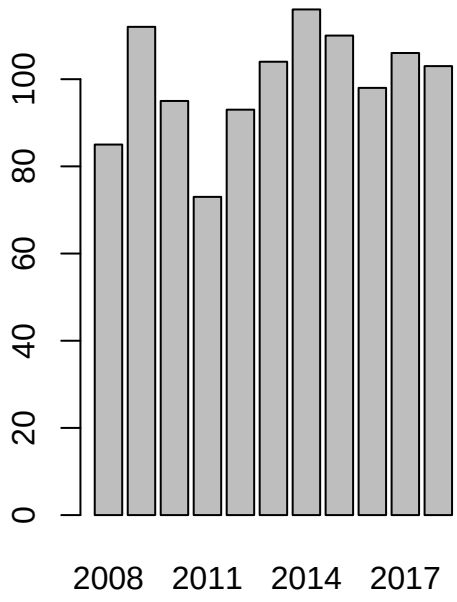
I could first examine the distribution of deaths by state to look for interesting patterns. Here is a series of boxplots that show the distribution of deaths by state in the year 2008.



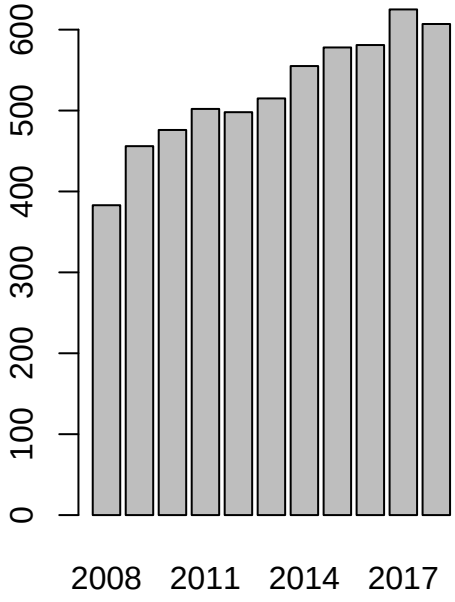
Distributions of deaths by state in 2008.

This plot appears fairly busy, and labels don't appear properly if the image is scaled down, but after examining these boxplots for a few years, I decided to include Pennsylvania in my comparative analysis; there appears to be a relatively high number of deaths in Pennsylvania jails. Pennsylvania is a swing state in general elections, and the state government is fairly equally divided between the two parties.

Next, I can examine in relation to other states. Following are two plots that compare yearly deaths in California to yearly deaths in all other states.

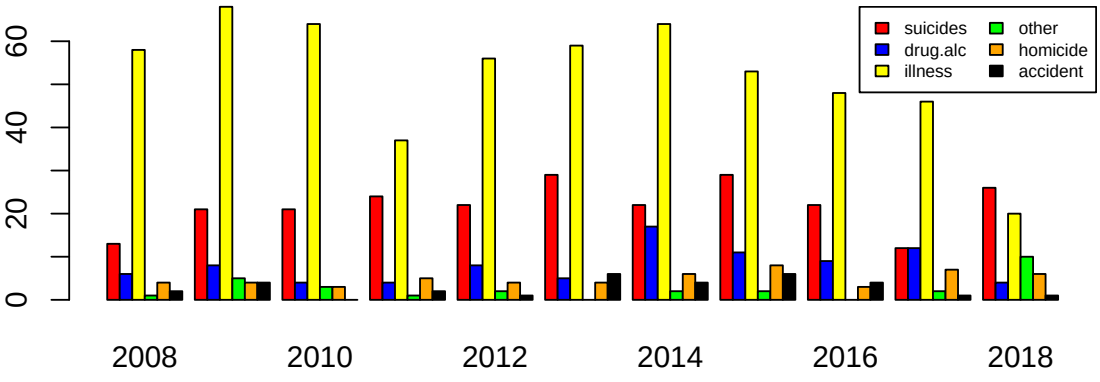


Total deaths is CA by year.



Total deaths in other states by year.

These plots indicate that there is upward trend in jail deaths nationally, but there is no such trend in California. I would want to see whether there is a trend in overall deaths in the states that I selected. Next I can examine my selected states in greater detail individually. Following is a plot demonstrating the distributions of deaths by type in California jails.

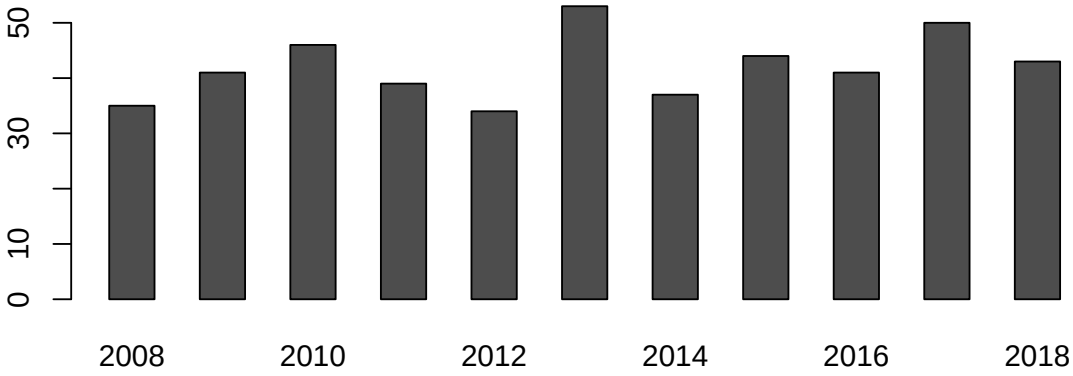


Deaths in CA by type.

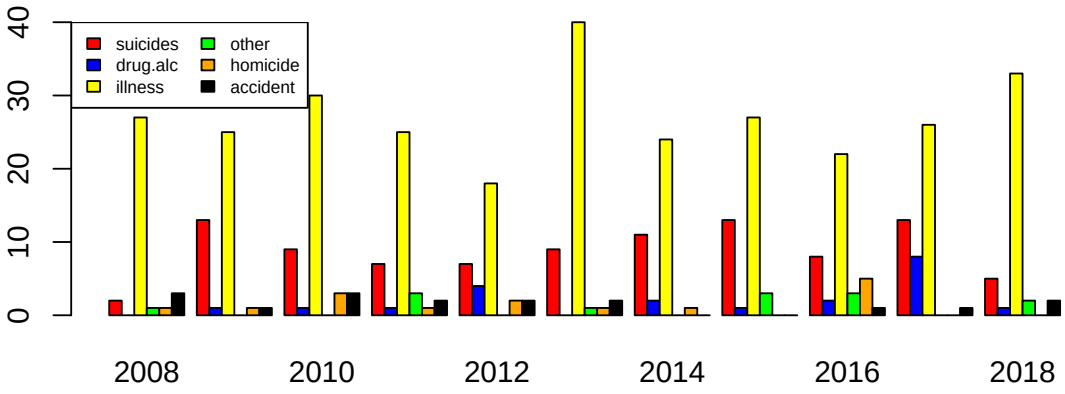
By far most of the deaths in California jails are due to illness. there is not an apparent trend in illness related deaths over time, however. Suicide is the second leading cause of death, and there may be an upward

trend in suicides. There are far fewer deaths related to the other causes represented in the dataset; of these, there may be an upward trend in drug/alcohol related deaths and homicides.

Next I will examine Texas, a fairly conservative state.



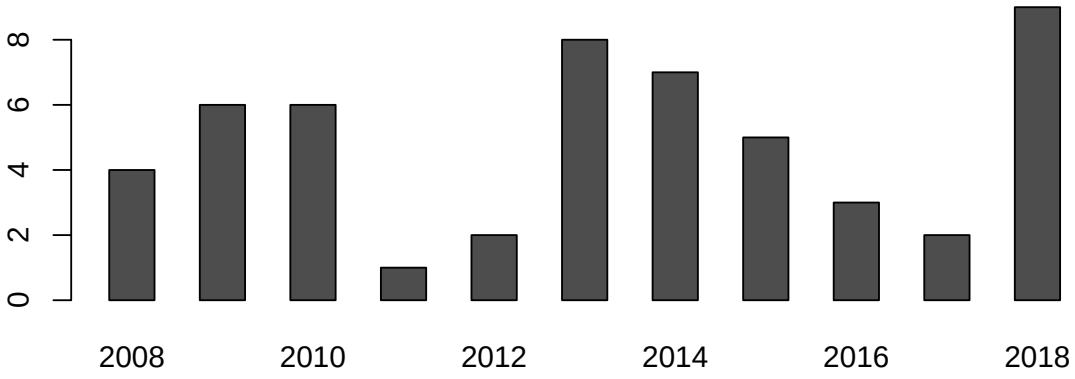
Total deaths in TX by year.



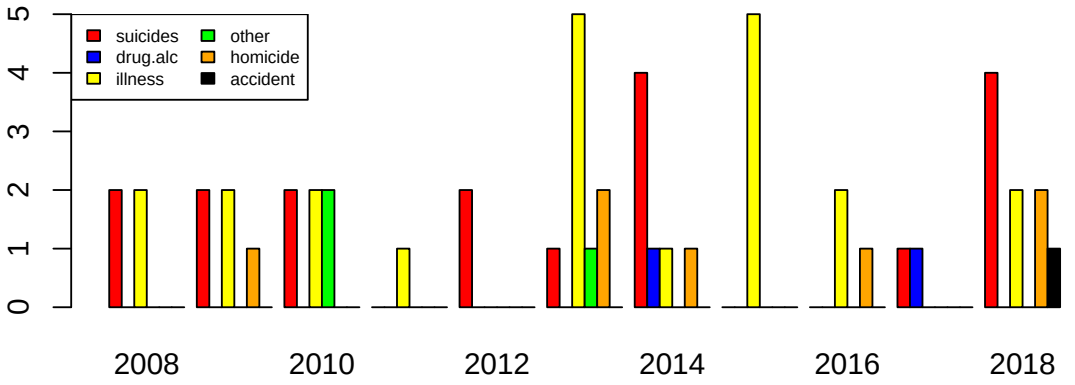
Deaths in TX by type.

As in California, there is not a clear trend in total deaths in Texas jails over time. Most of the deaths in Texas jails are attributed to illness, followed by suicides. Suicides appear to be increasing somewhat, though this is difficult to ascertain precisely from the barplot. Drug and alcohol related deaths appear to be increasing, and approach the number of suicides. Deaths due to other causes Appear negligible.

Now I will examine Mississippi, a poor southern state.



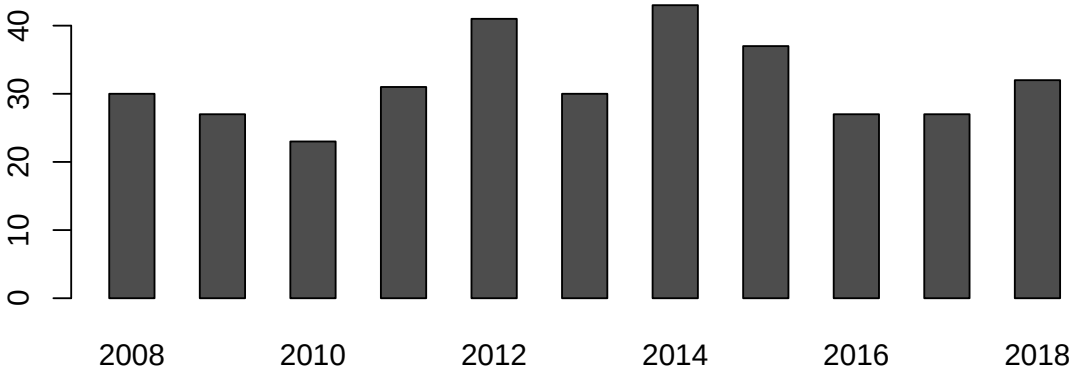
Total deaths in MS by year.



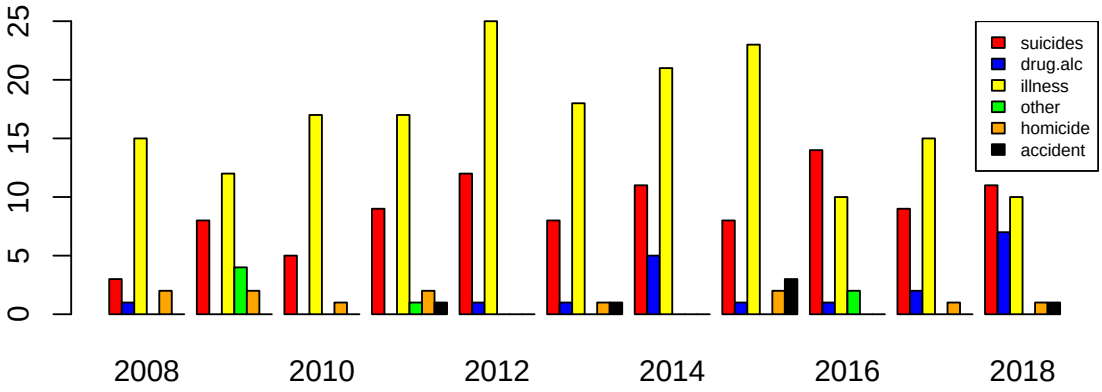
Deaths by type in MS.

Death trends in Mississippi jails are not very apparent from the bar plots. Mississippi has relatively few jails in the survey, and a relatively small inmate population.

Finally I will examine Pennsylvania, a fairly populous swing state.



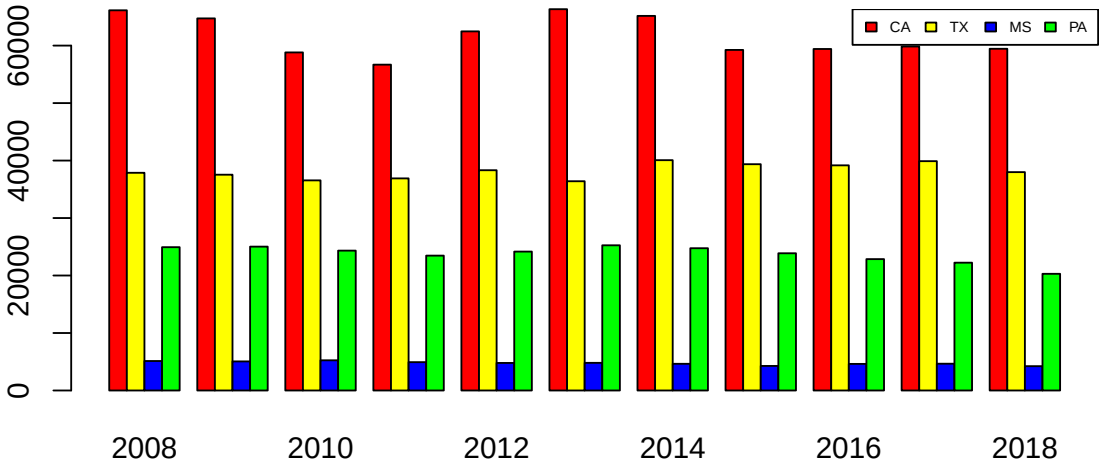
Total deaths in PA by year.



Deaths in PA by type.

Total deaths in Pennsylvania jails appear to spike in the middle of the data set, around year 2013. As in California and Texas, most of the deaths are attributed to illness, followed by suicide. Illness related deaths appear to spike around 2013 then decline, but suicides appear to continue to increase. Drug and alcohol related deaths appear to be increasing as well. Deaths attributed to other causes appear negligible.

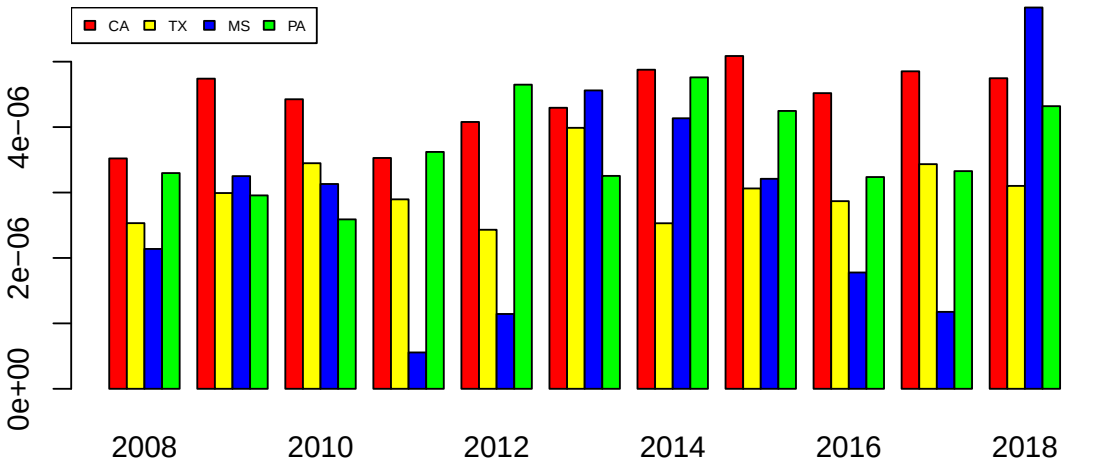
Next I would like to examine trends in population numbers over time. The following plot displays the average daily jail population per state for each year.



Average daily inmate population by state.

Inmate populations appear fairly constant by state, though Pennsylvania's inmate population appears to be decreasing somewhat.

Next, I want to examine death rates by state, given that the states I have selected have significantly different inmate populations. Following is a plot of the daily death rates by state for each year.



Daily death rates by state.

Death rates fluctuate a bit, but California generally has the highest, followed by Texas and Pennsylvania. These are the states with the largest inmate populations.

Modeling Methods

I intend here to examine trends within and between states with respect to types of death. It is apparent from my EDA that illness and suicide account for most of the deaths in any state, so I will focus on these. First, I want to examine trends in total deaths across states. I will treat total deaths as the response, and year and state as the predictors. I will start with a linear regression model, since the results are fairly easily interpretable, then fit a Poisson regression model, ridge, lasso, elasticnet, and a simple hidden-layer neural network, to compare predictive accuracy. I will use the mean squared error as my measure of predictive accuracy. I have split my data into training and testing sets. Data for years 2008 through 2018 is training data, and data for the year 2019 is testing data, and I will use my models to predict the number of total deaths by state in the year 2019.

After examining trends in total deaths, I want to examine trends in total death rates, suicide rates, and illness-related death rates. Given that the states I selected have dramatically different jail population sizes, I would expect proportions to yield more meaningful results than counts. I will obtain proportions by dividing deaths (total, suicides, and illness-related) by the average daily population, then run the same models as before. Once again, I will assess the predictive accuracy of each model by the value of the mean squared error.

Modeling Results

Following are the coefficients produced by a linear model, treating state and year as the explanatory variables, and total deaths as the response.

```
  (Intercept)       statems       statepa       statetx          year
-1250.0795455   -94.7272727   -67.9090909   -57.4545455     0.6704545
```

The negative coefficients suggest every state sees fewer jails deaths than California, the baseline state; Mississippi sees about 95 fewer deaths, Pennsylvania 68 fewer deaths, and Texas 57 fewer deaths per year. This is not particularly interesting, as California has by far the largest jail population, followed by Texas, then Pennsylvania, then Mississippi. The positive coefficient for the year suggests that total deaths are increasing somewhat over time. Based on the p-values, all of the coefficients are statistically significant. The Poisson regression model produces coefficients that tell the same story. It would be more interesting to look as death rates, given the significant difference in jail populations between the selected states. That said, I still want to examine the predictive accuracy of my selected models, so I will predict the total deaths for the year 2019 with each model, and calculate the mean squared error (MSE) from each prediction. Following is a table of MSE's for each model.

```
       model       mse
1     linear 144.13068
2    poisson  90.67883
3 elasticnet 150.36379
4      ridge 195.64349
5      lasso 150.58270
6  neuralnet  63.31187
```

The Poisson model seems to excel at modeling this data, which makes sense as this is count data. Linear regression, elasticnet, lasso, and ridge exhibit similar performance; perhaps ridge, lasso, or elasticnet would have performed better if the model included more predictors. The neural network predicts very well with minimal tuning (I specified 2 hidden layers with 5 neurons respectively), which is quite impressive.

Next, I will perform a comparative analysis of the death ratios, suicide ratios, and illness-related death ratios of these states. First, I will fit linear models and interpret the coefficients. Following are the coefficients produced by a linear model treating state and year as predictors, and total death ratio as the response.

```
  (Intercept)       statems       statepa       statetx          year
-6.247571e-02 -5.895214e-04 -2.792117e-04 -5.106279e-04  3.183841e-05
```

Given that the coefficients for all the states are negative, it appears as though the baseline state, California, exhibits the greatest ratio of deaths to it's average inmate population. Texas and Mississippi exhibit similar death ratios, and Pennsylvania exhibits the lowest, though the coefficient associated with pennsylvania may not be statistically significant. The positive coefficient associated with the year indicates that death rates overall are increasing with time.

Next, I will examine suicide ratios across states. As before, I will fit a linear model with state and year as predictors, and suicide ratio as the response. Following are the coefficients generated from the model.

```
  (Intercept)       statems       statepa       statetx          year
-2.847321e-02 -1.013351e-05  2.323125e-05 -1.267770e-04  1.432197e-05
```

In this case, a positive coefficient for Pennsylvania indicates that Pennsylvania exhibits a higher suicide rate than the baseline state, California. Negative coefficients for Mississippi and Texas indicate lower suicide rates than California, and Mississippi has the lowest. The positive coefficient year indicates that suicide rates are increasing overall with time. Based on corresponding p-values, however, none of these coefficients are statistically significant, so I wouldn't draw any conclusions from this output.

Next, I will examine the ratio of illness-related deaths to average population across states. Following are the coefficients produced by a linear model, treating year and state as predictors, and illness-related death ratio as the response.

```
  (Intercept)       statems       statepa       statetx          year
 1.514656e-02 -4.128897e-04 -1.431791e-04 -1.301073e-04 -7.106909e-06
```

Once again, all the states exhibit lower illness-related death ratios than the baseline state, California. Pennsylvania and Texas are about the same, and Mississippi is about four times higher than Texas and Pennsylvania. Based on the negative coefficient for the year, illness related deaths appear to be decreasing over time. I should note that, based on the p-values, only the coefficient associated with Mississippi is statistically significant.

It seems, based on the results of the linear models, that there isn't a clear relationship between states and types of death. California, the "liberal" state, consistently demonstrated the worst outcomes for inmates, whether I looked at total deaths or death rates. Texas, a "conservative" state, fared relatively well. It might be reasonable to assume that the size of a jail system is a better indication of a jail's condition rather than location.

Now I will examine the predictive accuracy of each of my models on the response variables previously explored. Following is a table of MSE's for each models prediction of total death rates in the year 2019.

```
       model          mse
1     linear 1.200646e-07
2    poisson 1.052841e-07
3 elasticnet 1.348586e-07
4      ridge 1.522412e-07
5      lasso 1.214737e-07
6  neuralnet 3.482134e-07
```

All of these MSE's are tiny, as are the death ratios. Poisson regression exhibits the best performance here, and neural net the worst. Perhaps with some tuning, neural net might perform better. The linear model predicts fairly well, and elasticnet, ridge, and lasso still struggle somewhat; lasso performs the best of the three. Next I will examine a table of MSE's for prediction of suicide ratios in the year 2019.

```
       model          mse
1     linear 1.052667e-06
2    poisson 2.303592e-08
3 elasticnet 3.744672e-08
4      ridge 3.457531e-08
5      lasso 5.467072e-08
6  neuralnet 5.673184e-08
```

The linear model produces the worst predictions here, which was to be expected given thta the coefficients produced by the model lacked statistical significance. The neural net is on par with elasticnet and ridge, and lasso performs the worst. Poisson regression demonstrates the best predictive accuracy again. Next I will examine a table of MSE's for prediction of illness-related death ratios in the year 2019.

```
       model          mse
1     linear 5.693303e-07
2    poisson 7.727481e-08
3 elasticnet 6.478137e-08
4      ridge 6.695817e-08
5      lasso 5.533268e-08
6  neuralnet 7.475030e-08
```

The linear model produces the worst predictions again, but ridge surpasses Poisson; there must be some colinearity that ridge accounts for. Elasticnet and lasso perform about equally well, better than Poisson and neural net.

Summary

Based on my EDA and the results of modeling, I cannot conclude that more liberal states produce better outcomes for their inmate populations than conservative and poor states. In fact California, my baseline "liberal" state, generally performs worse than the other selected states with respect to total deaths and death rates. The other states exhibit inconsistent results, but fairly consistently perform better than California; only Pennsylvania demonstrated a higher suicide rate, but this was not a statistically significant result. California has a significantly higher average inmate population than the other states, around 1.6 times larger than Texas, the state with the next largest average population, and despite having a large inmate population Texas, my model "conservative" state, generally demonstrates good outcomes for inmates.

Future work might entail grouping states by region; patterns may be more evident if I grouped states together based on my subjective criteria of similarity, rather than using single states as representative of certain political orientations. Some such groupings might be West Coast, Midwestern, Southern, and Northeastern states. However, it seems that based on the analysis I performed, the size of a jail or jail system would a better indicator of expected outcomes for inmates, given that the state with the largest inmate population consistently demonstrated the worst outcomes. If I were to start this project again, I would select jails from throughout the country, group by size (small, medium and large), and analyze outcomes for these groups.

References

[1] (12/11/2023). *Mass Incarceration: The Whole Pie 2023.* The Prison Policy Institute. https://www.prisonpolicy.org/reports/pie2023.html

[2] (12/11/2023). *Dying Inside: The Hidden Crisis in America's Jails.* Reuters. https://www.reuters.com/investigates/section/usa-jails/

Appendix

```
#################################################EDA#################################################
data = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/Allstatesinsurvey/all_jails
#replace NA values with zeroes
data = replace(data,is.na(data),0)
attach(data)

boxplot(d2008~state, sub = "Distributions of deaths by state in 2008.")#all states

deaths.ca = data[35:72,9:19]
deaths.other = data[-(35:72),9:19]

#vector of total yearly deaths in CA
yearly.deaths.ca = numeric()
for(i in 1:dim(deaths.ca)[2]){
  yearly.deaths.ca[i] = sum(deaths.ca[,i])
}

#vector of total yearly deaths in US except CA
yearly.deaths.other = numeric()
for(i in 1:dim(deaths.other)[2]){
  yearly.deaths.other[i] = sum(deaths.other[,i])
}

par(mfrow = c(1, 2))
barplot(yearly.deaths.ca,names.arg = c(2008:2018),sub = "Total deaths is CA by year.")
barplot(yearly.deaths.other,names.arg = c(2008:2018),sub = "Total deaths in other states by year.")

#contingency table for CA#############################################################################
data2 = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/CA.csv",header=T)

#california
train.ca = data2[-12,]
test.ca = data2[12,]

#deaths by type
colors = c("red","blue","yellow","green","orange","black")
barplot(t(train.ca[,-c(1,2,9)]),beside=T,col=colors,names.arg = c(2008:2018),sub = "Deaths in CA by typ
legend("topright",fill=colors,legend = colnames(train.ca[,-c(1,2,9)]),cex=0.6,ncol=2)

#contingency table for TX#############################################################################
data2 = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/TX.csv",header=T)
#texas
train.tx = data2[-12,]
test.tx = data2[12,]
par(mfrow = c(2, 1))
#All deaths by year in TX
barplot(t(train.tx[,2]),beside=T,names.arg = c(2008:2018),sub = "Total deaths in TX by year.",legend =

#deaths by type in TX
barplot(t(train.tx[,-c(1,2,9)]),beside=T,col=colors,names.arg = c(2008:2018),sub = "Deaths in TX by typ
```

```r
legend("topleft",fill=colors,legend = colnames(train.tx[,-c(1,2,9)]),cex=0.6,ncol = 2)

#contingency table for MS#############################################################################
data2 = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/MS.csv",header=T)
#mississippi
train.ms = data2[-12,]
test.ms = data2[12,]
par(mfrow = c(2, 1))
data2 = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/PA.csv",header=T)

#All deaths by year in MS
barplot(t(train.ms[,2]),beside=T,names.arg = c(2008:2018),sub = "Total deaths in MS by year.",legend = c

#deaths by type
barplot(t(train.ms[,-c(1,2,9)]),beside=T,col=colors,names.arg = c(2008:2018),sub = "Deaths by type in MS
legend("topleft",fill=colors,legend = colnames(train.tx[,-c(1,2,9)]),cex=0.6,ncol=2)

#contingency table for PA#############################################################################
data2 = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/PA.csv",header=T)
#pennsylvania
train.pa = data2[-12,]
test.pa = data2[12,]
par(mfrow = c(2, 1))
#All deaths by year in PA
barplot(t(train.pa[,2]),beside=T,names.arg = c(2008:2018),sub = "Total deaths in PA by year.")

#deaths by type
barplot(t(train.pa[,-c(1,2,9)]),beside=T,col=colors,names.arg = c(2008:2018),sub = "Deaths in PA by type
legend("topright",fill=colors,legend = colnames(train.tx[,-c(1,2,9)]),cex=0.6,ncol=1)

#Average daily population comparison
adp = cbind(train.ca[,9],train.tx[,9],train.ms[,9],train.pa[,9])
colors2 = c("red","yellow","blue","green")
barplot(t(adp), beside=T, col=colors2, names.arg = c(2008:2018),sub = "Average daily inmate population
legend("topright",fill=colors2,legend = c("CA","TX","MS","PA"),cex=0.5,ncol=4)

#Death rates comparison
rates = cbind((train.ca[,2]/365)/train.ca[,9], (train.tx[,2]/365)/train.tx[,9], (train.ms[,2]/365)/train
barplot(t(rates),beside=T,col=colors2, names.arg = c(2008:2018),sub = "Daily death rates by state.")
legend("topleft",fill=colors2,legend = c("CA","TX","MS","PA"),cex=0.5,ncol=4)

#################################################Modeling#############################################
deaths = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/totaldeaths.csv",header=T
#############################Models for total deaths#################################################
model.lin = lm(deaths~state+year, data=deaths)
model.lin$coefficients

library(glmnet)
test.deaths = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/totaldeaths.test.csv

#linear model
pred.lin = predict(model.lin,newdata=test.deaths)
mse.lin = mean((test.deaths$deaths - pred.lin)^2)
```

```r
#poisson
model.pois = glm(deaths~state+year, data=deaths, family = poisson(link="log"))
pred.pois = exp(predict(model.pois,newdata=test.deaths))
mse.pois = mean((test.deaths$deaths - pred.pois)^2)

x <- model.matrix(deaths~state+year - 1, data = deaths)
y = deaths$deaths

#elasticnet
cv3 = cv.glmnet(x, y, alpha = .5)
best.lam.elasticnet = cv3$lambda.min
model.elastic = glmnet(x,y,alpha = 0.5,lambda=best.lam.elasticnet)
pred.el = predict(model.elastic,newx = model.matrix(deaths~state+year - 1, data=test.deaths))
mse.elastic = mean((test.deaths$deaths - pred.el)^2)

#lasso
cv1 = cv.glmnet(x, y, alpha = 1)
best.lam.lasso = cv1$lambda.min
model.lasso = glmnet(x,y,alpha = 1,lambda=best.lam.lasso)
pred.la = predict(model.lasso,newx = model.matrix(deaths~state+year - 1, data=test.deaths))
mse.lasso = mean((test.deaths$deaths - pred.la)^2)

#ridge
cv2 = cv.glmnet(x, y, alpha = 0)
best.lam.ridge = cv2$lambda.min
model.ridge = glmnet(x,y,alpha = 0,lambda=best.lam.ridge)
pred.ri = predict(model.ridge,newx = model.matrix(deaths~state+year - 1, data=test.deaths))
mse.ridge = mean((test.deaths$deaths - pred.ri)^2)

#neuralnet
library(tidyverse)
library(neuralnet)
library(mltools)
library(data.table)

deaths$state = as.factor(deaths$state)
newdata <- one_hot(as.data.table(deaths)) #dummy code states
newdata$deaths = scale(newdata$deaths) #normalize deaths?
newdata$year = scale(newdata$year) #normalize year

model.neural = neuralnet(deaths~state_ca+state_ms+state_pa+state_tx+year,data=newdata,hidden=c(5,5))

test.deaths$state = as.factor(test.deaths$state)
new.test = one_hot(as.data.table(test.deaths))
new.test$deaths = scale(new.test$deaths)
new.test$year = 0

pred = predict(model.neural,new.test)
unscaled.pred = (pred*sd(deaths$deaths))+mean(deaths$deaths)

neural.mse = mean((unscaled.pred - test.deaths$deaths)^2)

table = data.frame(model = c("linear","poisson","elasticnet","ridge","lasso","neuralnet"),
```

```
                    mse = c(mse.lin,mse.pois,mse.elastic,mse.ridge,mse.lasso,neural.mse))
table

#############################################################Models for death rates, suicide rates, and illnes
deaths = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/totaldeaths.csv",header=T
test.deaths = read.csv("C:/Users/jbrin/OneDrive/Documents/Fall 2023/Math 533/final/totaldeaths.test.csv

deaths$death.ratio = deaths$deaths/deaths$avgpop
deaths$suicide.ratio = deaths$suicide/deaths$avgpop
deaths$illness.ratio = deaths$illness/deaths$avgpop

test.deaths$death.ratio = test.deaths$deaths/test.deaths$avgpop
test.deaths$suicide.ratio = test.deaths$suicide/test.deaths$avgpop
test.deaths$illness.ratio = test.deaths$illness/test.deaths$avgpop

lin.model.dratio = lm(death.ratio~state+year,data=deaths)
lin.model.sratio = lm(suicide.ratio~state+year,data=deaths)
lin.model.illratio = lm(illness.ratio~state+year,data=deaths)

lin.pred1 = predict(lin.model.dratio,newdata = test.deaths)
lin.pred2 = predict(lin.model.sratio,newdata = test.deaths)
lin.pred3 = predict(lin.model.illratio,newdata = test.deaths)

lin.mse1 = mean((lin.pred1 - test.deaths$death.ratio)^2)
lin.mse2 = mean((lin.pred1 - test.deaths$suicide.ratio)^2)
lin.mse3 = mean((lin.pred1 - test.deaths$illness.ratio)^2)

lin.model.dratio$coefficients

#poisson again
pois.model.dratio = glm(death.ratio~state+year, data=deaths, family = poisson(link="log"))
pois.model.sratio = glm(suicide.ratio~state+year, data=deaths, family = poisson(link="log"))
pois.model.iratio = glm(illness.ratio~state+year, data=deaths, family = poisson(link="log"))

pois.pred1 = exp(predict(pois.model.dratio,newdata=test.deaths))
pois.pred2 = exp(predict(pois.model.sratio,newdata=test.deaths))
pois.pred3 = exp(predict(pois.model.iratio,newdata=test.deaths))

pois.mse1 = mean((test.deaths$death.ratio - pois.pred1)^2)
pois.mse2 = mean((test.deaths$suicide.ratio - pois.pred2)^2)
pois.mse3 = mean((test.deaths$illness.ratio - pois.pred3)^2)

#elasticnet, lasso, ridge
library(glmnet)
#elasticnet
#death ratio
x <- model.matrix(death.ratio~state+year - 1, data = deaths)
y = deaths$death.ratio
cv3 = cv.glmnet(x, y, alpha = .5)
best.lam.elasticnet = cv3$lambda.min
model.elastic = glmnet(x,y,alpha = 0.5,lambda=best.lam.elasticnet)
pred.el1 = predict(model.elastic,newx = model.matrix(death.ratio~state+year - 1, data=test.deaths))
mse.elastic1 = mean((test.deaths$death.ratio - pred.el1)^2)
```

```
#suicide ratio
x <- model.matrix(suicide.ratio~state+year - 1, data = deaths)
y = deaths$suicide.ratio
cv3 = cv.glmnet(x, y, alpha = .5)
best.lam.elasticnet = cv3$lambda.min
model.elastic = glmnet(x,y,alpha = 0.5,lambda=best.lam.elasticnet)
pred.el1 = predict(model.elastic,newx = model.matrix(suicide.ratio~state+year - 1, data=test.deaths))
mse.elastic2 = mean((test.deaths$suicide.ratio - pred.el1)^2)
#illness ratio
x <- model.matrix(illness.ratio~state+year - 1, data = deaths)
y = deaths$illness.ratio
cv3 = cv.glmnet(x, y, alpha = .5)
best.lam.elasticnet = cv3$lambda.min
model.elastic = glmnet(x,y,alpha = 0.5,lambda=best.lam.elasticnet)
pred.el1 = predict(model.elastic,newx = model.matrix(illness.ratio~state+year - 1, data=test.deaths))
mse.elastic3 = mean((test.deaths$illness.ratio - pred.el1)^2)

#lasso
#death ratio
x <- model.matrix(death.ratio~state+year - 1, data = deaths)
y = deaths$death.ratio
cv1 = cv.glmnet(x, y, alpha = 1)
best.lam.lasso = cv1$lambda.min
model.lasso = glmnet(x,y,alpha = 1,lambda=best.lam.lasso)
pred.la = predict(model.lasso,newx = model.matrix(death.ratio~state+year - 1, data=test.deaths))
mse.lasso1 = mean((test.deaths$death.ratio - pred.la)^2)
#suicide ratio
x <- model.matrix(suicide.ratio~state+year - 1, data = deaths)
y = deaths$suicide.ratio
cv1 = cv.glmnet(x, y, alpha = 1)
best.lam.lasso = cv1$lambda.min
model.lasso = glmnet(x,y,alpha = 1,lambda=best.lam.lasso)
pred.la = predict(model.lasso,newx = model.matrix(suicide.ratio~state+year - 1, data=test.deaths))
mse.lasso2 = mean((test.deaths$suicide.ratio - pred.la)^2)
#illness ratio
x <- model.matrix(illness.ratio~state+year - 1, data = deaths)
y = deaths$illness.ratio
cv1 = cv.glmnet(x, y, alpha = 1)
best.lam.lasso = cv1$lambda.min
model.lasso = glmnet(x,y,alpha = 1,lambda=best.lam.lasso)
pred.la = predict(model.lasso,newx = model.matrix(illness.ratio~state+year - 1, data=test.deaths))
mse.lasso3 = mean((test.deaths$illness.ratio - pred.la)^2)

#ridge
#death ratio
x <- model.matrix(death.ratio~state+year - 1, data = deaths)
y = deaths$death.ratio
cv2 = cv.glmnet(x, y, alpha = 0)
best.lam.ridge = cv2$lambda.min
model.ridge = glmnet(x,y,alpha = 0,lambda=best.lam.ridge)
pred.ri = predict(model.ridge,newx = model.matrix(death.ratio~state+year - 1, data=test.deaths))
mse.ridge1 = mean((test.deaths$death.ratio - pred.ri)^2)
#suicide ratio
```

```
x <- model.matrix(suicide.ratio~state+year - 1, data = deaths)
y = deaths$suicide.ratio
cv2 = cv.glmnet(x, y, alpha = 0)
best.lam.ridge = cv2$lambda.min
model.ridge = glmnet(x,y,alpha = 0,lambda=best.lam.ridge)
pred.ri = predict(model.ridge,newx = model.matrix(suicide.ratio~state+year - 1, data=test.deaths))
mse.ridge2 = mean((test.deaths$suicide.ratio - pred.ri)^2)
#illness ratio
x <- model.matrix(illness.ratio~state+year - 1, data = deaths)
y = deaths$illness.ratio
cv2 = cv.glmnet(x, y, alpha = 0)
best.lam.ridge = cv2$lambda.min
model.ridge = glmnet(x,y,alpha = 0,lambda=best.lam.ridge)
pred.ri = predict(model.ridge,newx = model.matrix(illness.ratio~state+year - 1, data=test.deaths))
mse.ridge3 = mean((test.deaths$illness.ratio - pred.ri)^2)

#neural network
#death ratio
deaths$state = as.factor(deaths$state)
newdata <- one_hot(as.data.table(deaths)) #dummy code states
newdata$death.ratio = scale(newdata$death.ratio) #normalize deaths
newdata$year = scale(newdata$year) #normalize year
model.neural = neuralnet(death.ratio~state_ca+state_ms+state_pa+state_tx+year,data=newdata,hidden=c(5,5)
test.deaths$state = as.factor(test.deaths$state)
new.test = one_hot(as.data.table(test.deaths))
new.test$death.ratio = scale(new.test$death.ratio)
new.test$year = 0
pred = predict(model.neural,new.test)
unscaled.pred = (pred*sd(deaths$death.ratio))+mean(deaths$death.ratio)
neural.mse1 = mean((unscaled.pred - test.deaths$death.ratio)^2)
#suicide ratio
newdata$suicide.ratio = scale(newdata$suicide.ratio) #normalize
model.neural = neuralnet(suicide.ratio~state_ca+state_ms+state_pa+state_tx+year,data=newdata,hidden=c(5
new.test$suicide.ratio = scale(new.test$suicide.ratio)
pred = predict(model.neural,new.test)
unscaled.pred = (pred*sd(deaths$suicide.ratio))+mean(deaths$suicide.ratio)
neural.mse2 = mean((unscaled.pred - test.deaths$suicide.ratio)^2)
#illness ratio
newdata$illness.ratio = scale(newdata$illness.ratio) #normalize
model.neural = neuralnet(illness.ratio~state_ca+state_ms+state_pa+state_tx+year,data=newdata,hidden=c(5
new.test$illness.ratio = scale(new.test$illness.ratio)
pred = predict(model.neural,new.test)
unscaled.pred = (pred*sd(deaths$illness.ratio))+mean(deaths$illness.ratio)
neural.mse3 = mean((unscaled.pred - test.deaths$illness.ratio)^2)

#predictive accuracy for death rates
table1 = data.frame(model = c("linear","poisson","elasticnet","ridge","lasso","neuralnet"),
                    mse = c(lin.mse1,pois.mse1,mse.elastic1,mse.ridge1,mse.lasso1,neural.mse1))

#predictive accuracy for suicide rates
table2 = data.frame(model = c("linear","poisson","elasticnet","ridge","lasso","neuralnet"),
                    mse = c(lin.mse2,pois.mse2,mse.elastic2,mse.ridge2,mse.lasso2,neural.mse2))
```

```r
#predictive accuracy for illness-related death rates
table3 = data.frame(model = c("linear","poisson","elasticnet","ridge","lasso","neuralnet"),
                    mse = c(lin.mse3,pois.mse3,mse.elastic3,mse.ridge3,mse.lasso3,neural.mse3))
```