# Datathon

## *10 Error 404 Team Not Found*

#### *2/22/2019*

# R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# Research question: Which states are the most vulnerable to droughts?

## Methodology

### We look at what are the states that severe droughts happen, then explore on what are the industries that earnings decrease is mostly related to droughts. Additionally we also look at the water quality on average in different states in order to come up with the conclusion of what are the states that are most vulnerable.

### The methodologies we use are random sampling and linear models to analyse the relationship between industry specific earnings and droughts.

## Read the datasets

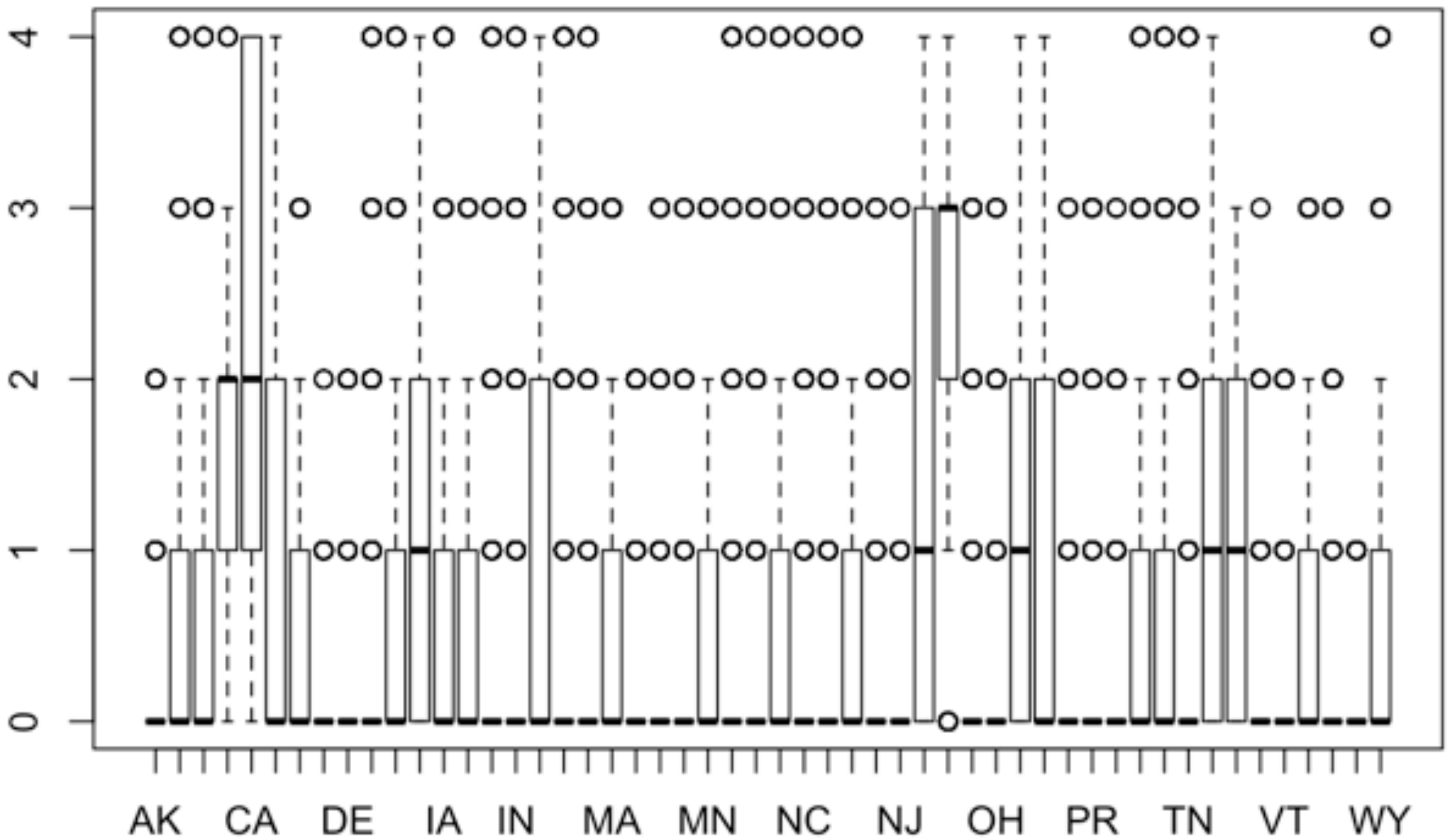
```
chemicals <- read.csv("chemicals.csv")
summary(chemicals)
```

```
##       fips                         county           state
##  Min.   : 6001   Polk County        : 10644   PA     : 86876
##  1st Qu.:19109   Washington County  : 10275   FL     : 74807
##  Median :33005   Marion County      :  9941   NY     : 71040
##  Mean   :30190   Hillsborough County:  9529   CA     : 65449
##  3rd Qu.:42027   Jefferson County   :  8915   WA     : 60681
##  Max.   :55141   Orange County      :  8738   MO     : 46051
##                  (Other)            :824277   (Other):477415
##       year                     cws_name              pws_id
##  Min.   :1999   WHISPERING PINES MHP:   261   FL6411132:    99
##  1st Qu.:2005   GREEN ACRES MHP     :   224   CO0118015:    98
##  Median :2009   MOUNTAIN VIEW MHP   :   209   FL6515234:    94
##  Mean   :2009   COUNTRY ESTATES     :   179   FL6521784:    94
##  3rd Qu.:2013   COUNTRYSIDE MHP     :   173   VT0005290:    94
##  Max.   :2016   COUNTRY ESTATES MHP :   171   VT0020455:    94
##                 (Other)             :881102   (Other)  :881746
##    pop_served              chemical_species
##  Min.   :      0   Arsenic          :142001
##  1st Qu.:    118   DEHP             : 72825
##  Median :    485   Halo-Acetic Acid:146132
##  Mean   :  10732   Nitrates         :329372
##  3rd Qu.:   3030   Trihalomethane  :154258
##  Max.   :8271000   Uranium          : 37731
##
##               contaminant_level      unit_measurement      value
##  Greater than MCL      : 13545   micrograms/L:882319   Min.   :     0.0
##  Less than or equal MCL:503954                         1st Qu.:     1.0
##  Non Detect            :364820                         Median :    11.3
##                                                        Mean   :   426.6
##                                                        3rd Qu.:   140.0
##                                                        Max.   :150000.0
##
```

```
droughts <- read.csv("droughts.csv")
```

### First, we want to get a sense of what are the states that suffer from severe droughts.

```
droughts$vul <- ifelse(droughts$d4>0, 4, ifelse(droughts$d3>0, 3, ifelse(droughts$
d2>0, 2, ifelse(droughts$d1>0,1,0))))
plot(droughts$state,droughts$vul)
```




```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.1.0     ✔ purrr   0.3.0
## ✔ tibble  2.0.1     ✔ dplyr   0.7.8
## ✔ tidyr   0.8.2     ✔ stringr 1.3.1
## ✔ readr   1.3.1     ✔ forcats 0.3.0
```

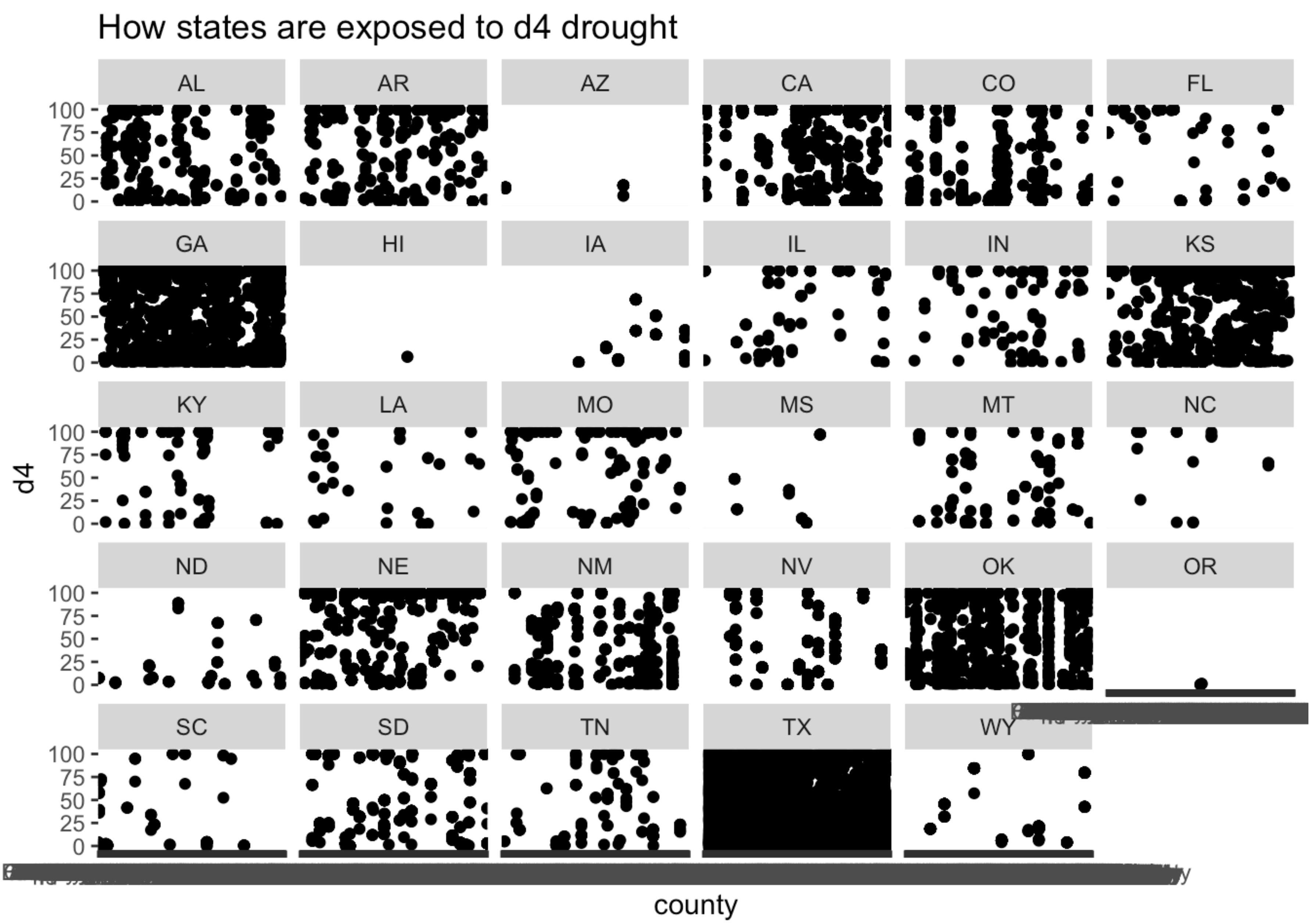
```
## ── Conflicts ────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
droughts %>% select(county, state, d4) %>% arrange(desc(d4)) %>% top_n(10)
```

```
## Selecting by d4
```

```
##                county state     d4
## 1   McPherson County    NE 100.01
## 2   McPherson County    NE 100.01
## 3   McPherson County    NE 100.01
## 4   McPherson County    NE 100.01
## 5   McPherson County    NE 100.01
## 6   McPherson County    NE 100.01
## 7   McPherson County    NE 100.01
## 8   McPherson County    NE 100.01
## 9   McPherson County    NE 100.01
## 10  McPherson County    NE 100.01
## 11  McPherson County    NE 100.01
## 12  McPherson County    NE 100.01
## 13  McPherson County    NE 100.01
## 14  McPherson County    NE 100.01
## 15  McPherson County    NE 100.01
## 16  McPherson County    NE 100.01
## 17  McPherson County    NE 100.01
## 18  McPherson County    NE 100.01
## 19  McPherson County    NE 100.01
## 20  McPherson County    NE 100.01
## 21  McPherson County    NE 100.01
## 22  McPherson County    NE 100.01
## 23  McPherson County    NE 100.01
## 24  McPherson County    NE 100.01
## 25  McPherson County    NE 100.01
## 26  McPherson County    NE 100.01
## 27  McPherson County    NE 100.01
## 28  McPherson County    NE 100.01
## 29  McPherson County    NE 100.01
## 30  McPherson County    NE 100.01
## 31  McPherson County    NE 100.01
## 32  McPherson County    NE 100.01
## 33  McPherson County    NE 100.01
## 34  McPherson County    NE 100.01
## 35  McPherson County    NE 100.01
## 36  McPherson County    NE 100.01
## 37  McPherson County    NE 100.01
## 38  McPherson County    NE 100.01
## 39  McPherson County    NE 100.01
## 40  McPherson County    NE 100.01
## 41    Harding County    NM 100.01
## 42    Harding County    NM 100.01
## 43 Jeff Davis County    TX 100.01
## 44 Jeff Davis County    TX 100.01
## 45 Jeff Davis County    TX 100.01
## 46 Jeff Davis County    TX 100.01
## 47 Jeff Davis County    TX 100.01
## 48 Jeff Davis County    TX 100.01
## 49 Jeff Davis County    TX 100.01
```

```
droughts %>% filter(d4 != 0) %>% select(county, state, d4) %>% ggplot() + geom_poi
nt(mapping=aes(x=county, y= d4)) + facet_wrap(~state) + ggtitle("How states are ex
posed to d4 drought")
```




The states AL, AR, CA, GA, KS, OK, NE, SD, NM and TX are mostly affected by the most severe droughts. In TX, NE and NM, severe droughts affect to the extent of almost 100% of their population. These counties are all in the midwestern US.

One of the shortcomings of these facet wrap is that it takes d4 only as the indicator of severe droughts. Also, part of the reason for the density of points in TX is because it has more data points than the others. Despite these, however, this graph still shows us what are the states with more d4 happening than others.

## What industries are mostly affected by droughts?

Dummy model looks at how fish mining is affected.

```r
earnings <- read.csv("earnings.csv")
#head(earnings)
library(tidyverse)

set.seed(123)
sample_size <- floor(nrow(droughts) * 0.01)
sample_id <- sample(1:nrow(droughts), sample_size)
sample <- droughts[sample_id,]
# A random sample of 1% of data is chosen from the droughts, such that the followi
ng models will run quicklier.

prbl2 <- merge(earnings, sample, by="state")
prbl2$agri <- prbl2$total_agri_fish_mine
dummymodel <- lm(agri ~ none + d1 + d2 + d3 + d4, data = prbl2)
summary(dummymodel)
```

```
##
## Call:
## lm(formula = agri ~ none + d1 + d2 + d3 + d4, data = prbl2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32481  -8288  -1870   6484 150230
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 32442.1034    13.0881 2478.751  < 2e-16 ***
## none           -0.5164     0.1498   -3.448 0.000565 ***
## d1             16.8449     0.2335   72.143  < 2e-16 ***
## d2             25.3855     0.2427  104.577  < 2e-16 ***
## d3             16.0475     0.2908   55.186  < 2e-16 ***
## d4             10.3210     0.4013   25.716  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12340 on 6640468 degrees of freedom
##   (43261 observations deleted due to missingness)
## Multiple R-squared:  0.004367,   Adjusted R-squared:  0.004366
## F-statistic:  5825 on 5 and 6640468 DF,  p-value: < 2.2e-16
```

Contrary to what we thought, fish mining workers love drought. So fish mining industry is not negatively affected even during periods of water resource shortage.

# Model 1 looks at fish hunt.

```r
model1 <- lm(agri_fish_hunt ~ none + d1 + d2 + d3 + d4, data = prbl2)
summary(model1)
```

```
##
## Call:
## lm(formula = agri_fish_hunt ~ none + d1 + d2 + d3 + d4, data = prbl2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24021  -5837   -823   4912 223658
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 26467.3183    10.7321 2466.192  < 2e-16 ***
## none           -1.2572     0.1229  -10.234  < 2e-16 ***
## d1              0.5340     0.1913    2.791  0.00525 **
## d2              0.2094     0.1989    1.053  0.29231
## d3              0.2353     0.2381    0.988  0.32304
## d4             -1.0636     0.3286   -3.237  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10070 on 6581077 degrees of freedom
##   (102652 observations deleted due to missingness)
## Multiple R-squared:  4.705e-05,	Adjusted R-squared:  4.629e-05
## F-statistic: 61.93 on 5 and 6581077 DF,  p-value: < 2.2e-16
```

#### While fish hunt is not significantly affected by the droughts.

# Model 2 looks at construction.

```
model2 <- lm(construction ~ none + d1 + d2 + d3 + d4, data = prbl2)
summary(model2)
```

```
##
## Call:
## lm(formula = construction ~ none + d1 + d2 + d3 + d4, data = prbl2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29919  -4950   -440   4273  62028
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32027.9113     9.1971 3482.40   <2e-16 ***
## none            3.9145     0.1053   37.19   <2e-16 ***
## d1             -1.9047     0.1648  -11.56   <2e-16 ***
## d2             -5.2017     0.1711  -30.40   <2e-16 ***
## d3             -7.6110     0.2035  -37.41   <2e-16 ***
## d4             -4.9787     0.2798  -17.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7501 on 4950337 degrees of freedom
##   (1733392 observations deleted due to missingness)
## Multiple R-squared:  0.002203,   Adjusted R-squared:  0.002202
## F-statistic:  2186 on 5 and 4950337 DF,  p-value: < 2.2e-16
```

Construction is severely impacted by the drought. As we know, construction generally consumes much water resources. That may be why when there is a drought these industries suffer.

# Model 3 below looks at financial services industry.

```
model3 <- lm(fin_ins_realest ~ none + d1 + d2 + d3 + d4, data = prbl2)
summary(model3)
```

```
##
## Call:
## lm(formula = fin_ins_realest ~ none + d1 + d2 + d3 + d4, data = prbl2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32068  -6344  -1705   4296 216503
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 33895.0438    11.3445 2987.793  < 2e-16 ***
## none            0.1696     0.1298    1.307    0.191
## d1             -0.9030     0.2030   -4.448 8.66e-06 ***
## d2             -3.9788     0.2116  -18.806  < 2e-16 ***
## d3             -1.7055     0.2534   -6.730 1.69e-11 ***
## d4              6.7260     0.3504   19.194  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10670 on 6594740 degrees of freedom
##   (88989 observations deleted due to missingness)
## Multiple R-squared:  0.0001575,  Adjusted R-squared:  0.0001568
## F-statistic: 207.8 on 5 and 6594740 DF,  p-value: < 2.2e-16
```

We can see that finance, insurance and real estate industries are greatly impacted by d1, d2 and d3 level of droughts. For example, if there is 1% more population affected by d2, there will be 3.97 decrease in earnings of financial industries. To the contrary, however, a most severe d4 drought has a positive impact on these industries. What's more, if there is no drought, financial services are not going to be significantly impacted.

Since construction is clearly strongly affected by the drought, we want to see which states are primary focused on this industry.

```
industry_occupation <- read.csv("industry_occupation.csv")

construct_pop = industry_occupation %>% group_by(state) %>% summarize(num = sum(co
nstruction), perct=num/sum(total_employed)) %>% arrange(desc(perct)) %>% top_n(10)
```

```
## Selecting by perct
```

```
construct_pop
```

```
## # A tibble: 10 x 3
##    state     num  perct
##    <fct>   <dbl>  <dbl>
##  1 MT     156325 0.0829
##  2 LA     810856 0.0793
##  3 TX    5741994 0.0790
##  4 AK     122846 0.0777
##  5 ND     104697 0.0737
##  6 WY      44507 0.0731
##  7 HI     328532 0.0727
##  8 CO    1133205 0.0724
##  9 OK     530526 0.0705
## 10 ID     217200 0.0684
```

```
#edu_pop = industry_occupation %>% group_by(state) %>% summarize(num = sum(edu_hea
lth), perct=num/sum(total_employed)) %>% arrange(desc(perct)) %>% top_n(10)
```

# We can see that the top three states where construction is their major industry are MT, LA and TX. From graph “How states are exposed to d4 drought”, we can see that TX is subject to severe droughts. Considering the fact that the construction industry is negatively impacted by droughts from linear model 3, TX is very vulnerable to droughts. It’s the same case with MT, AL and OK.

## What about the financial industries?

```
fin_pop = industry_occupation %>% group_by(state) %>% summarize(num = sum(finance_
insurance_realestate), perct=num/sum(total_employed)) %>% arrange(desc(perct)) %>%
top_n(10)
```

```
## Selecting by perct
```

```
fin_pop
```

```
## # A tibble: 10 x 3
##    state     num  perct
##    <fct>   <dbl>  <dbl>
##  1 SD     105510 0.100
##  2 DE     294351 0.0974
##  3 IA     549629 0.0968
##  4 NE     350974 0.0954
##  5 CT    1136314 0.0914
##  6 NJ    2573023 0.0860
##  7 MN    1152078 0.0828
##  8 NY    5042621 0.0827
##  9 AZ    1562032 0.0822
## 10 MO    1066830 0.0787
```

The top three states that has finance as its mojor indutry are SD, DE and IA. Thus these states are not quite vulnerable to droughts.

## Taking a look at the agriculture industry as a whole,

```
agri_pop = industry_occupation %>% group_by(state) %>% summarize(num = sum(agricul
ture), perct=num/sum(total_employed)) %>% arrange(desc(perct)) %>% top_n(10)
```

```
## Selecting by perct
```

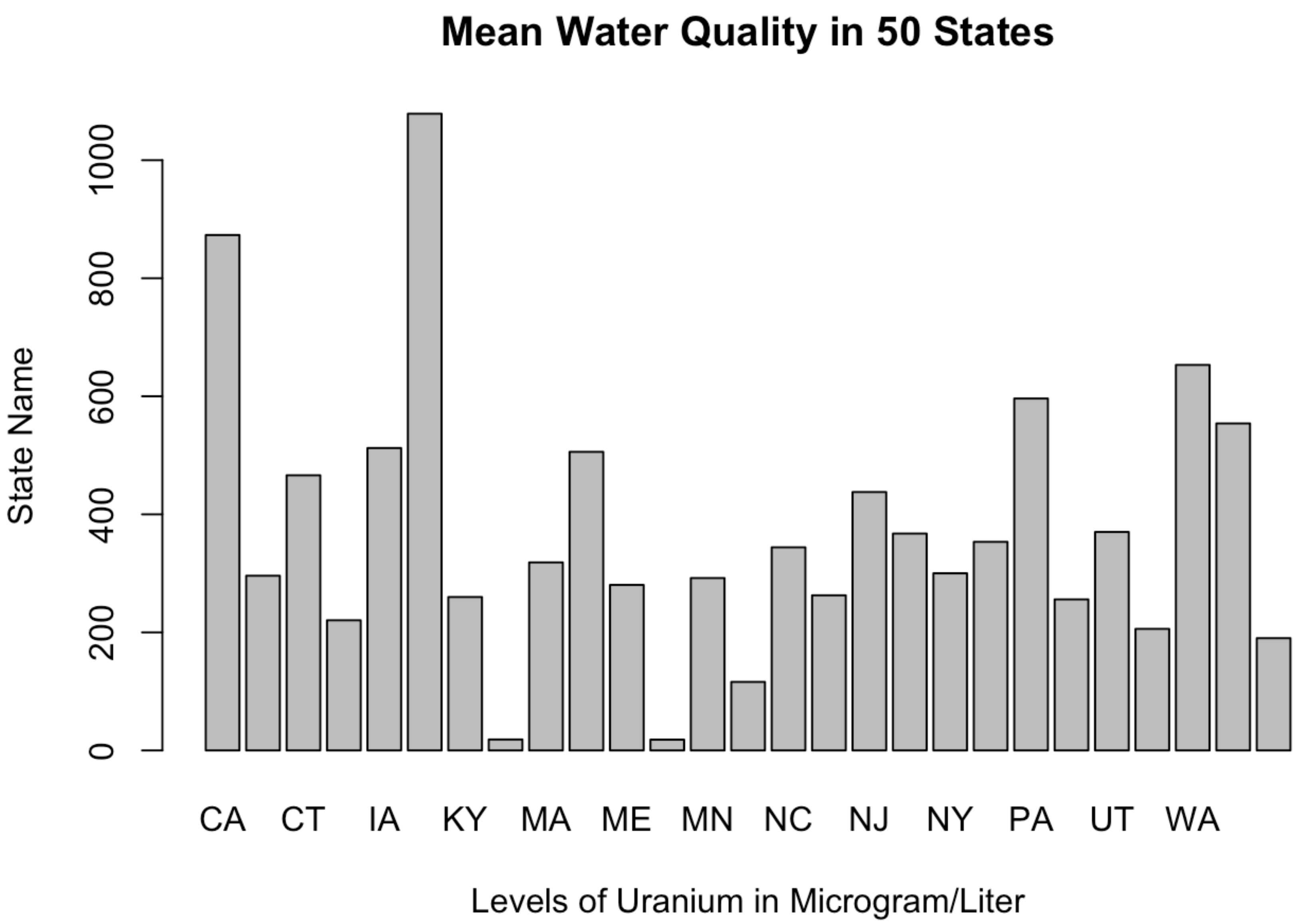
```
agri_pop
```

```
## # A tibble: 10 x 3
##    state     num  perct
##    <fct>   <dbl>  <dbl>
##  1 WY      39500 0.0648
##  2 AK      60600 0.0383
##  3 ND      49692 0.0350
##  4 MT      65886 0.0349
##  5 LA     345242 0.0338
##  6 ID      90729 0.0286
##  7 OK     210427 0.0280
##  8 NM     111303 0.0268
##  9 OR     285061 0.0268
## 10 TX    1694592 0.0233
```

Fish hunt is severely impacted by drought. Fishing is a part of agriculture industry. So the states rely most on agriculture are also vulnerable to droughts. The top three are WY, AK and ND.

## Water quality in different states

```
meanLevels <- tapply(chemicals$value, chemicals$state, mean)
barplot(meanLevels)
title (main = "Mean Water Quality in 50 States", xlab= "Levels of Uranium in Micro
gram/Liter", ylab= "State Name")
```




```
names((sort(meanLevels,decreasing=T))[1:5])
```

```
## [1] "KS" "CA" "WA" "PA" "WI"
```

By observing the water quality in different states, determined by the mean value of chemicals in all waters in each state. The top five states with worst water quality are KS, CA, WA, PA and WI. Therefore, these states are the most vulnerable to drought, in the sense that if there is a water shortage, these states will most probably have less clean water to use than others.

## Analytical and Modeling rigor

By concluding that certain states are more vulnerable to droughts based on their top industries we did have to make assumptions. We assumed low construction had a negative impact as well as low agriculture because those were the areas with the highest correlation to high drought levels. We also assumed that the mean reflected all counties in each state as an accurate representation of the uranium levels (in Micrograms/Liter) in the water.

# Conclusion

From the problem statement, we were most intrigued by the sample question of "What counties are most vulnerable in the event of a drought? Do droughts have an effect on industry specific earnings? Through our analysis we were able to discover the counties most susceptible to droughts which allowed to to develop an understanding of which states in the country were more prone to droughts. This was a great starting point that allowed us to expand our knowledge into greater researches.

By finding the greatest industry in specific states, ones that had large industries more vulnerable to droughts, such as agriculture and construction, we were able to conclude which countries were more noticeably vulnerable. By comparing the median and mean chemical levels in the counties of each of the states, we were able to detect there are multiple varying causes for different levels of droughts in each of these states and that overall, TX, MT, LA and KS are noticeably vulnerable to droughts.