Team # 13: Knights of the Round Tableau

Water Contamination Violations With Respect to County Size and Industry Focus

Water quality has long been perceived as an issue that only developing or impoverished countries face with some rare exceptions in developed nations, such as Flint, MI. However, while Americans drink water that the majority believe is safe, potentially harmful contaminants, such as arsenic, copper, and lead, have been found in the tap water of every state in the nation. Even though we are a developed country with well-established water treatment systems, most water goes untreated as it travels from a point source, farms, and other sources into our lakes in rivers through means of travel outside of the sewage system. When considering sewage and wastewater (commercial, industrial, and agricultural activities), 850 billion gallons are released from the nation's aging and overwhelmed treatment systems, roughly 6.5% of the total amount of wastewater produced annually. This number is substantial. Although less than 10%, 850 billion gallons of water are introduced into the nations water systems with high levels of harmful chemicals.

After considering this fault in our wastewater treatment system, the question "Are America's treatment infrastructure overwhelmed?" arises. We hypothesized that the current treatment systems may be serving too many citizens that it can handle. However, upon further analysis, we found that the counties that reported a violation of clean water standards had a population that on average was 5x smaller than those who did not report a violation. This leads us to uncover water quality issues in rural areas

In rural areas, often the same industries that provide the community its "economic backbone" are the same industries that devastate the region's water supply the most. In regions near the Appalachian mountains, many individuals get their water from wells that are sunk or flooded mines that were once loaded with heavy metals. However, the industry that causes the most benefit to many communities but causes the most damage to water supply is agriculture. Nitrogen-based fertilizers and other chemicals used to grow crops, especially industrial agriculture, is often washed away into rivers and streams when it rains, arriving at water systems without being treated first. Fixing rural water problems can be costly, and finding the money to fix a problem for small populations is a massive issue debilitating the rural wastewater infrastructure. The agriculture industry is one of the heaviest users of water, but there is improper infrastructure in place that could be causing a large number of pollutants to enter the nations water systems. Could improving water treatment in rural areas solve most of America's water quality issues?

The Big Question

The question we wanted to answer to provide more insight to some of the issues regarding issues in rural areas is:

How does country size and industry focus affect water quality?

To answer this question we looked into 4 different areas:

1. Do water quality violations occur more in counties with low populations or high populations?
2. Do counties who have more violations than average use more water on average than those who have fewer violations? If so, what industries use more water?
3. Using total violations per county over a 6 year period (2010 - 2016), how can we predict the number of violations a county will having based on % of the workforce in each industry?
4. What chemicals should we expect to find in counties with high agriculture use to guide potential future treatment investments?

Initial Data Cleaning and Merging

We used R and Tableau to conduct our analyses. R was used to clean and organize the data, while Tableau was used to provide graphics.

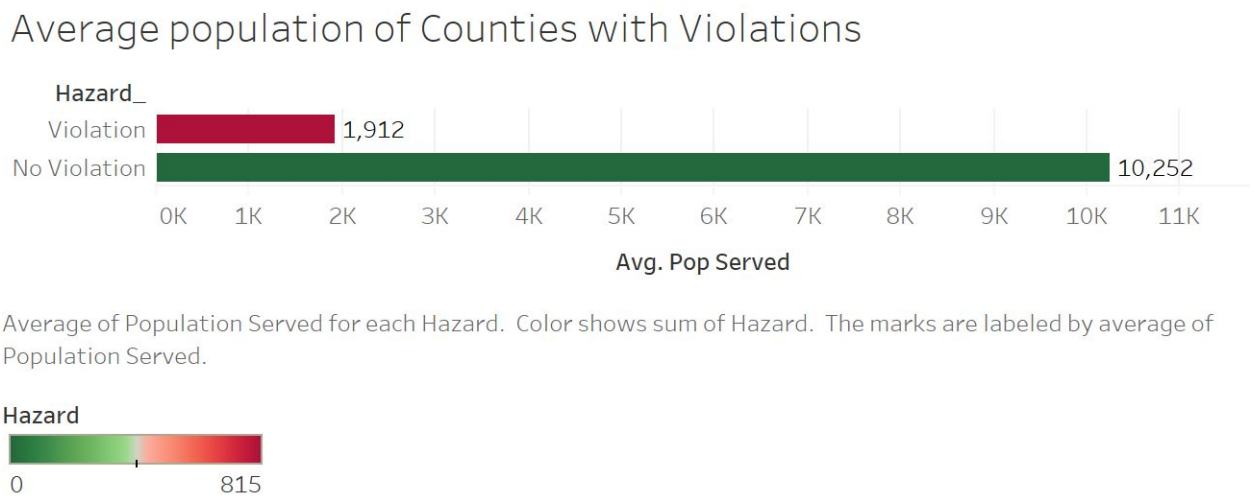
Upon first receiving the data, we conducted some quick cleaning and organization of the data. Fips, pws_idm, chemical species, and a few other variables were converted to factors in R. Additionally, a new variable that joined county and state was created in situations where FIPS numbers were not available and needed to be applied to certain datasets.

Definitions

Contamination Violation = Yes (1) if Contamination level > MCL.
Agriculture Focused Counties = Top 80 counties by percentage of employed who work in the Agriculture industry
Manufacturing Focused Counties = Top 80 counties by percentage of employed who work in the Manufacturing industry

Detailed Analysis

1. Do water quality violations occur more in counties with low populations or high populations?

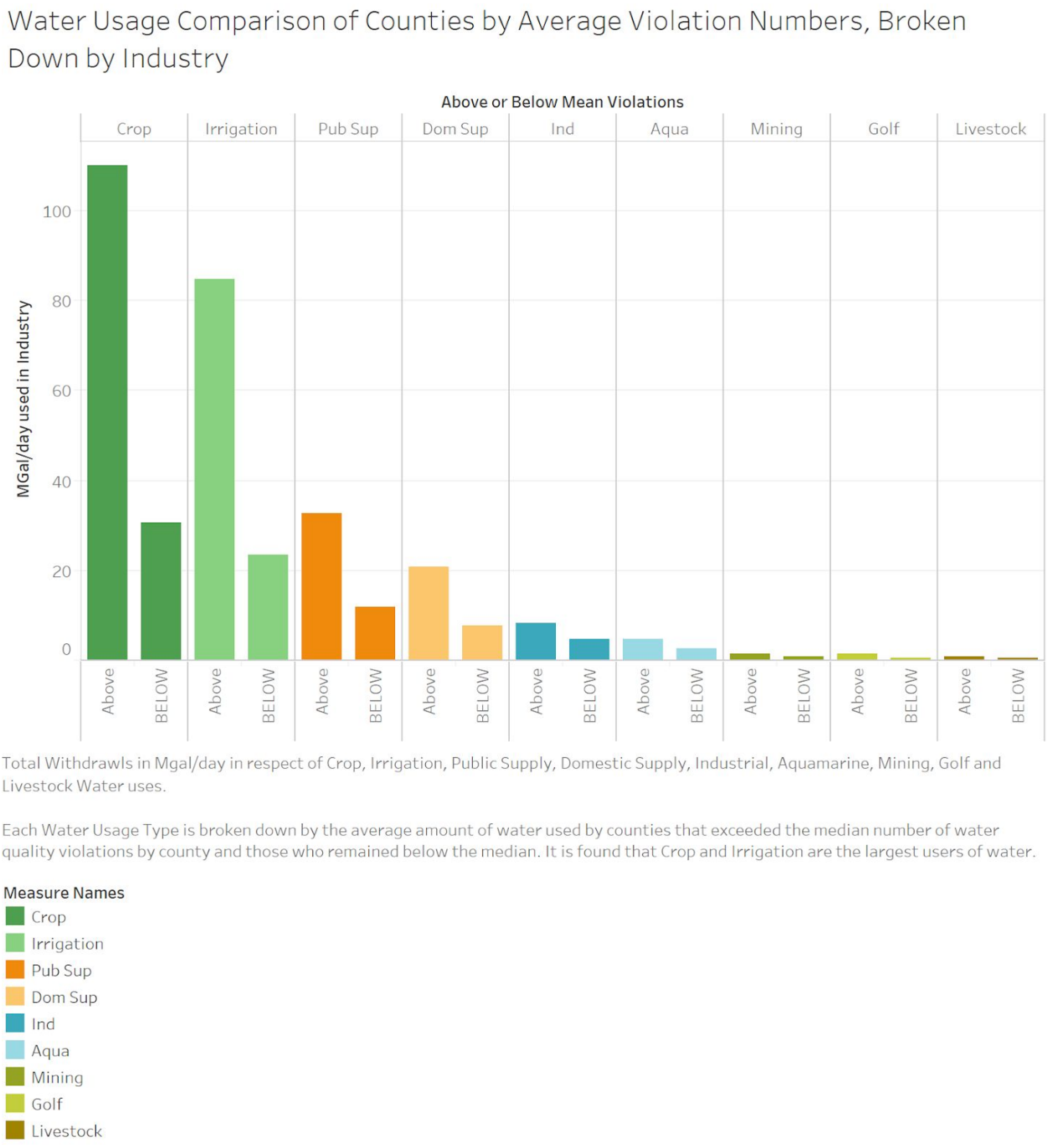


Comparing the populations served by the water systems that had contaminations vs the water systems that did not, we found the opposite of what we were expecting. We found that the water systems that had contamination, had around 80% fewer people served than those that did not.

Takeaway

Violations of water quality standards were more commonly found in smaller communities than large ones.

2. Do counties who have more violations than average use more water on average than those who have fewer violations? If so, what industries use more water?



First, to get a good baseline of contamination rates, we looked at every counties total contamination violations over a 6 year period, from 2010-2016. We took the average number of violations for all of the counties and created a baseline number of violations over that time period of **3.4** violations. Using this baseline we categorized the counties into two segments:

Above the Mean Violations and Below the Mean Violations. Dividing the counties based on this categorization, we plotted to see where and how the water was being used in counties that were Above the Mean number of violations and how they compared to their counterparts.

Takeaway: Not only did the counties that experience above the mean number of violations use much more water than those that did not, but they used significantly more water in two particular industries: Crops and Irrigation.

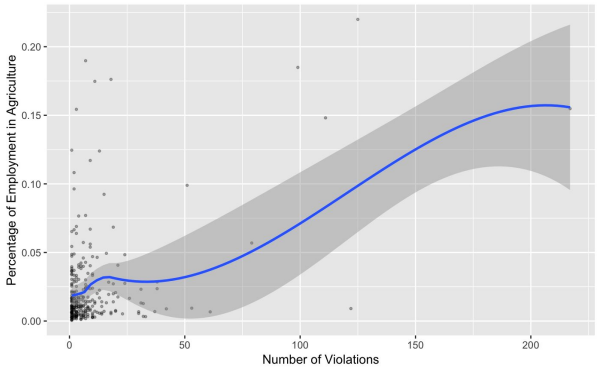
3. Using total violations per county over a 6 year period (2010 - 2016), how can we predict the number of violations a county will having based on % of workforce in each industry?

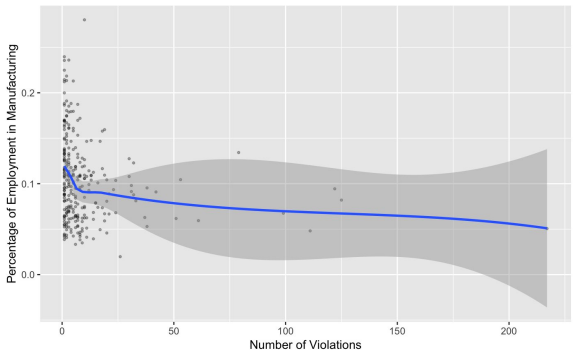
We built a linear regression model to determine what industries can be considered to predict number of violations based on the percentage of employees in each industry by the total employees in that county. After deleting all of the unnecessary variables, we found that in this regression model, “agriculture” and “manufacturing” are the two most significant factors. Therefore, we can conclude that number of violations can be estimated by:
Number of violations = 218 * agriculture -96 * manufacturing + 33.

```
lm(formula = X0 ~ +agriculture + construction + manufacturing +
    wholesale_trade + retail_trade + transport_utilities + information +
    finance_insurance_realestate + prof_scientific_waste + edu_health +
    arts_recreation + other + public_admin, data = iomerge)

Residuals:
    Min      1Q  Median      3Q     Max
-56.821  -6.454  -2.796   2.395 168.801

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                     33.67      48.32   0.697 0.486664
agriculture                    217.79      65.11   3.345 0.000951 ***
construction                   -61.05      93.53  -0.653 0.514550
manufacturing                  -96.08      50.69  -1.895 0.059219 .
wholesale_trade                177.22     160.56   1.104 0.270760
retail_trade                    33.79      82.68   0.409 0.683115
transport_utilities           -124.71     100.55  -1.240 0.216044
information                   -154.68     187.45  -0.825 0.410090
finance_insurance_realestate   -60.47      74.94  -0.807 0.420490
prof_scientific_waste           18.26      72.23   0.253 0.800627
edu_health                     -56.60      58.50  -0.968 0.334224
arts_recreation                -58.94      67.34  -0.875 0.382297
other                          138.61     149.63   0.926 0.355171
public_admin                       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.18 on 247 degrees of freedom
Multiple R-squared:  0.2133,    Adjusted R-squared:  0.1751
F-statistic: 5.582 on 12 and 247 DF,  p-value: 1.88e-08
```

We decided to remove duplicated fips and keep only one record for a unique fips because we made the assumption that the complexion of a county’s industrial focus would not change

significantly from year to year, which means after removing the duplicate records, the remaining record can be one of these six years, and we don't quite care which year it is.
Later we merged the six-year violations dataset with the industry occupation dataset to see whether there is a trend in both agriculture and manufacturing. Using the data from in Agriculture Focused Counties we noticed that as the number of violations increases, the percentage of employment in Agriculture also increases, and at the same time, the percentage of employment in Manufacturing decreases, which is consistent with the regression model to some extent.





Furthermore, we calculated the mean employment numbers within the Agriculture Focused Counties for both agriculture and manufacturing, which results in a number of 84,329 in agriculture and 135,622 in manufacturing.

Takeaway: The number of violations a county experienced between 2010-2016 is significantly, positively correlated with the percentage of the county's workforce devoted to Agriculture. The number of violations a county experience between 2010-2016 is significantly, negatively correlated with the county's workforce devoted to Manufacture.

4. What chemicals should we expect to find in counties with high agriculture use to guide potential future treatment investments? And how can we explain contaminants based on industry and geography?

After some online research, we tracked down the most common origins of the six contaminants that our data set tested. Below is a table describing the common sources:

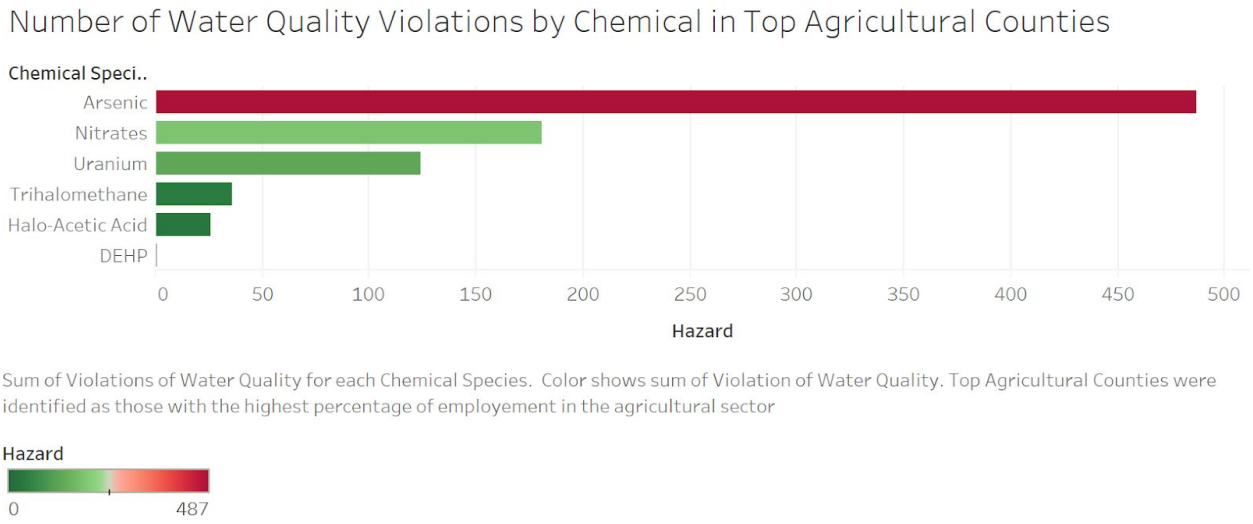
Arsenic	Erosion of natural deposits; runoff from orchards, runoff from glass and electronics production wastes
DEHP	Discharge from rubber and chemical factories

Halo-Acetic-Acid	Byproduct of drinking water disinfection
Nitrates	Runoff from fertilizer use; leaking from septic tanks, sewage; erosion of natural deposits
Trihalomethane	Byproduct of disinfecting drinking water
Uranium	Erosion of natural deposits

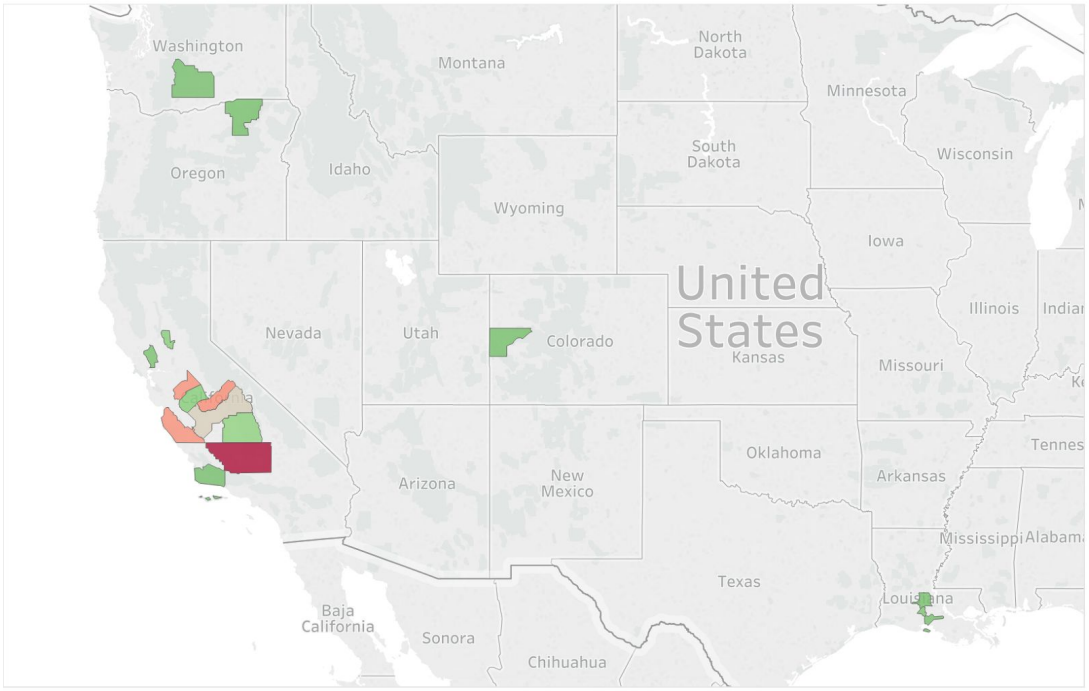
We then grouped the 4 of the main contaminants into two groups, based on their most common sources of contamination.

Nitrates & Arsenic: These contaminants mostly come from runoff related to fertilized used in agriculture as well as erosion of natural deposits. We associated these with Agriculture Focused Counties.

Trihalomethane & Halo-Acetic-Acid: When county water commissions are aware of their water -- regardless of source -- needs treatment that will introduce chemicals to disinfect the water. However, they simultaneously can create contaminants as the chemicals design to disinfect will create byproducts when the interact with the composition of the water that are harmful to humans.



Map of Arsenic Water Quality Violations in top Agricultural Counties

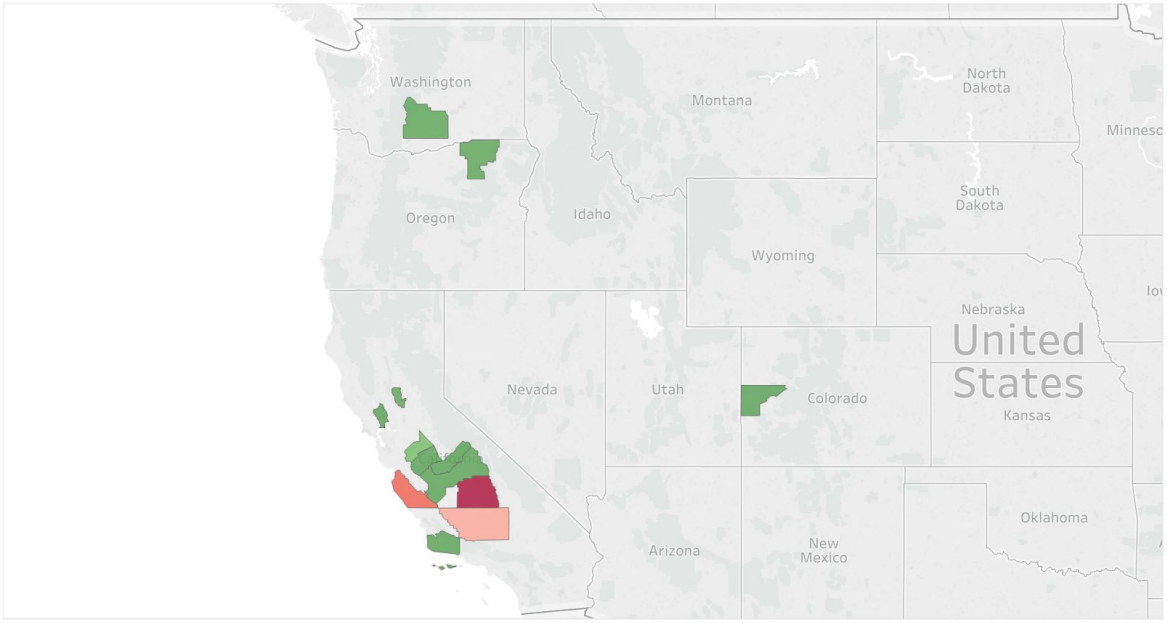


Color shows sum of Violation by Arsenic in each county. The data is filtered on Chemical Species, which keeps Arsenic. The view is filtered on County

Hazard

0 152

Map of Nitrate Water Quality Violations in top Agricultural Counties



Color shows sum of Violation by Nitrate in each county. The data is filtered on Chemical Species, which keeps Nitrates. The view is filtered on County

Hazard

0 70

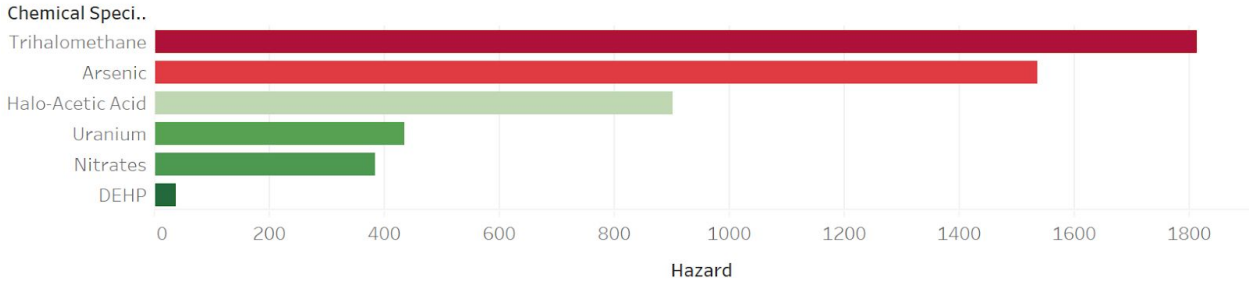
Country Breakdown of Violations of Water Quality by Chemical

		Chemical Species					
State	County	Arsenic	Nitrates	Uranium	DEHP	Halo-Ace..	Trihalom..
CA	Fresno County	37	1	13	0		
	Kern County	152	30	35	0		
	Kings County	17	0	1	0		
	Madera County	66	0	59	0		
	Merced County	13	0	0	0		
	Monterey County	63	47	1	0		
	Napa County	3	0	0	0		
	Santa Barbara County	1	1	0	0		
	Santa Cruz County	0	1	0	0		
	Stanislaus County	63	11	5	0		
	Sutter County	13	2	0	0		
	Tulare County	19	70	10	0		
CO	Mesa County	0	0	0	0	1	1
LA	Iberia Parish	1		0	0	0	0
	Lafayette Parish	6		0	0	0	1
	St. Landry Parish	0		0	0	1	3
MI	Van Buren County					0	1
NM	San Juan County	0	0	0	0	4	5
NY	Livingston County	0	0	0	0	0	9
OR	Douglas County	0	0	0	0	6	3
	Marion County	19	0	0	0	0	0
	Umatilla County	0	2	0	0	0	0
PA	Indiana County	0	0	0	0	11	5
	Somerset County	0	0	0	0	2	0
WA	Benton County	1	5	0	0	0	0
	Franklin County	3	8	0	0	0	0
	Grant County	6	1	0	0	0	0
	Lewis County	2	0		0	1	0
	Yakima County	1	2	0	0	0	0
WV	Harrison County	1	0		0	0	8

Sum of Violations broken down by Chemical Species vs. State and County. Color shows sum of Violations in each county. The marks are labeled by sum of Violations.

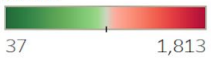
Hazard

Reports of Violations of Water Quality by Chemical

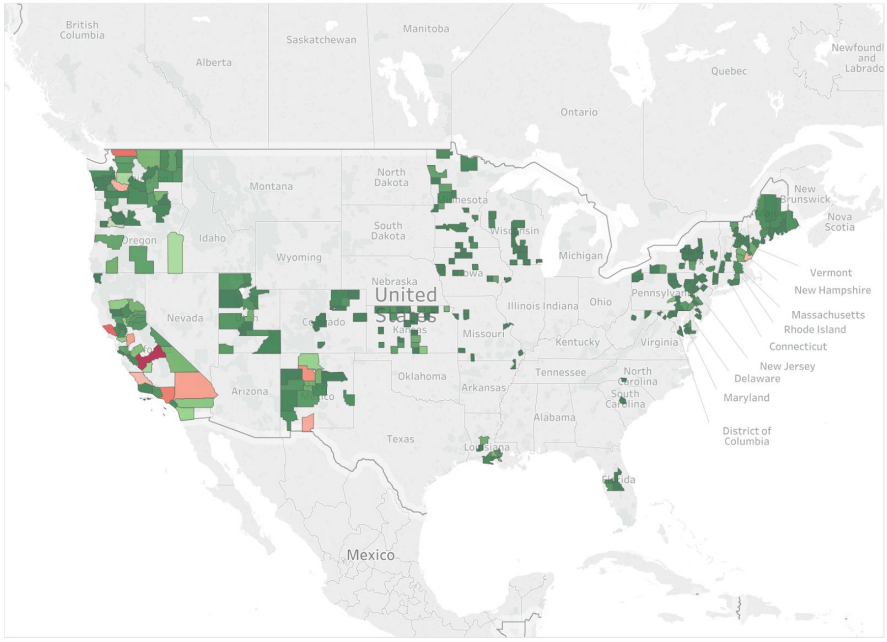


Sum of Hazard for each Chemical Species. Color shows sum of Hazard.

Hazard



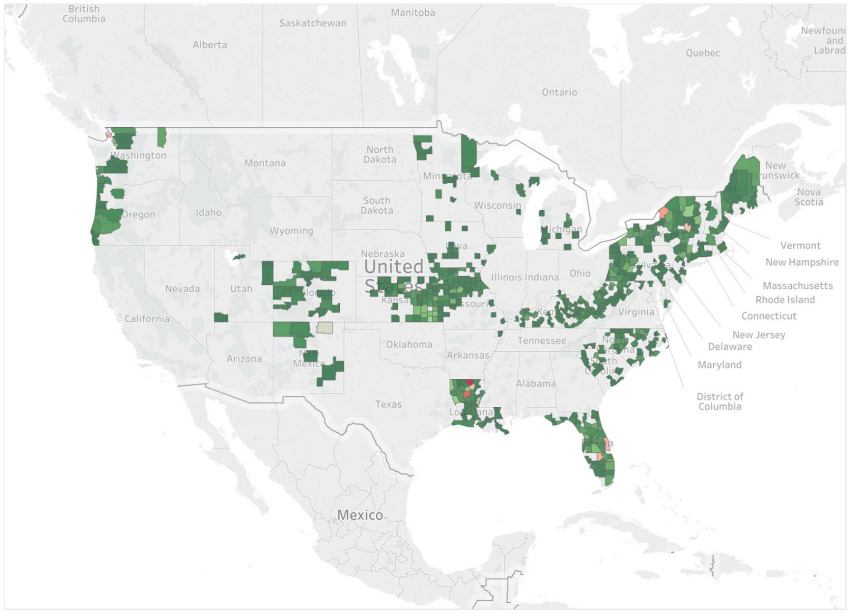
Counties with Reported Violations of Water Quality due to common Agricultural pollutants (Nitrate and Arsenic)



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Hazard. Details are shown for State and County. The data is filtered on Chemical Species, which keeps Arsenic and Nitrates. The view is filtered on sum of Hazard, which ranges from 1 to 56.

Hazard

Counties with Reported Violations of Water Quality due to common Drinking Water Cleaning and Mining pollutants (Trihalomethane and Halo-Acetic-Acid)



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Hazard. Details are shown for State and County. The data is filtered on Chemical Species, which keeps Halo-Acetic Acid and Trihalomethane. The view is filtered on sum of Hazard, which ranges from 1 to 152.

Hazard

1 57

<u>Takeaway:</u> The graph above and the maps below illustrate the difference in contaminant composition when comparing Agriculture focused counties with the entire country. Agriculture Focused counties are disproportionately affected by Nitrates and Arsenic. Moreover, the map shows that areas that have mining industries (West Virginia) or poor groundwater geological make ups (Florida's limestone) have to try hard to disinfect their water which inadvertently creates harmful byproducts.

<u>Conclusion</u>

We have found that counties that are focused in agriculture are more likely to have violations of water quality. In such counties, the contaminants are far more concentrated on Nitrates and Arsenic when compared to the rest of the country. While in counties that are not classified as Agriculture Focused, a majority of their contaminants are related to the by-products of trying to disinfect untreated water. The lack of these contaminants in Agriculture Focused counties means that their current water cleaning treatment solution is not producing a significant amount of these by-product contaminants; which is good.

However, due to the existence of Nitrates and Arsenic, their current treatment solution and method cannot treat the high volume demand of water in their county in time. In order to reduce the contaminants level in water of these countries a few options can be considered:

1. Agriculture and Crop companies need to reduce water usage so that treatment companies can properly process the water before sending it out
2. Farmers need to reduce or change their fertilization methods or solutions
3. Treatment centers need to create formula that effectively cleans water faster to handle their demand

Recommendation for Agriculture Focused Counties: Farmers and Croppers should work in conjunction with County Treatment centers to create fertilizing solutions that simultaneously compliment Treatment Center solutions, so that Farmers and Croppers can maximize their harvest while using Fertilizer that is compatible with Treatment Centers solution for disinfection water.

In addition to these issues above, there is an indirect effect on the overall health of rural Americans that contaminated water can exacerbate. According to the CDC, rural Americans are already disproportionately more likely to die from cancer. While this is attributed to a variety of different factors, such as tobacco use and obesity, rural Americans also have to contend with inadequate public health infrastructure as well. They are more uninsured, and rural health care locations are declining steadily. This leaves rural Americans poorly equipped to cope with the consequences of drinking contaminated water that will build on their pre-existing issues. Perhaps by investing in better water treatment infrastructure, we can improve the chances for these Americans a fighting chance against many of the other health issues they face. This could decrease the cost of insuring rural Americans, benefitting health insurance companies and allow more health centers in these areas to remain open or to grow now that costs to insure have gone down.

External Sources:

https://newrepublic.com/article/147011/rural-americas-drinking-water-crisis

https://www.nrdc.org/stories/water-pollution-everything-you-need-know#common

http://www.everythingconnects.org/water-pollution.html

Appendix (Unsuccessful Analysis)

- Regression model combining chemicals dataset and census dataset - https://www.kaggle.com/muonneutrino/us-census-demographic-data

```{r}
census <- read.csv("Census.csv")
census$fips <- census$FIPS
census$FIPS <- NULL

census$CensusId<- NULL
census$TotalPop<- NULL
census$result <- NULL
census$CountyState2 <- NULL
census$Illegal <- NULL
ccmerge <- merge(x=six_year_violations,y=census, by.x = 'X6001',by.y='fips' )
ccmerge$fips <- NULL
ccmerge$X6001 <- NULL

ccmodel <- lm(X0~ . ,data = ccmerge)
summary(ccmodel)

ccmodel2 <- lm(X0~Men+Women+White+Black+Native+Asian+Pacific+Income +
IncomePerCap
+IncomePerCapErr+Drive+Carpool+Transit+Walk+OtherTransp+WorkAtHome+MeanCommute
+Employed+Unemployment, data = ccmerge)
summary(ccmodel2)

ccmodel3 <- lm(X0~Men+Women+White+Black+Native+Pacific+Income + IncomePerCap
+IncomePerCapErr+Walk+OtherTransp+MeanCommute+Employed+Unemployment, data =
ccmerge)
summary(ccmodel3)

ccmodel4 <- lm(X0~Men+Women+White+Black+Native+Pacific+Income + IncomePerCap
+IncomePerCapErr+MeanCommute+Employed+Unemployment, data = ccmerge)
summary(ccmodel4)

cor(ccmerge$X0, ccmerge$ChildPoverty)
```

```
Call:
lm(formula = X0 ~ Men + Women + White + Black + Native + Pacific +
    Income + IncomePerCap + IncomePerCapErr + MeanCommute + Employed +
    Unemployment, data = ccmerge)

Residuals:
    Min      1Q  Median      3Q     Max
-37.540  -2.819  -1.308   1.035 174.551

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.760e+01  2.878e+00   6.117 1.22e-09 ***
Men              4.836e-04  5.050e-05   9.577  < 2e-16 ***
Women           -2.682e-04  5.004e-05  -5.360 9.66e-08 ***
White           -1.950e-01  2.185e-02  -8.926  < 2e-16 ***
Black           -1.700e-01  3.014e-02  -5.642 2.01e-08 ***
Native          -2.003e-01  7.349e-02  -2.726 0.006488 **
Pacific          3.347e+00  1.428e+00   2.344 0.019210 *
Income          -1.891e-04  4.923e-05  -3.842 0.000127 ***
IncomePerCap     4.464e-04  1.001e-04   4.459 8.85e-06 ***
IncomePerCapErr -1.581e-03  3.491e-04  -4.529 6.39e-06 ***
MeanCommute      1.653e-01  5.495e-02   3.008 0.002678 **
Employed        -2.142e-04  2.056e-05 -10.418  < 2e-16 ***
Unemployment    -2.168e-01  1.081e-01  -2.005 0.045137 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.056 on 1481 degrees of freedom
Multiple R-squared:  0.2106,    Adjusted R-squared:  0.2042
F-statistic: 32.93 on 12 and 1481 DF,  p-value: < 2.2e-16
```

<u>The result is pretty significant, but we don’t think the result can be easily explained by reasoning</u>

- Urban & Rural - https://seer.cancer.gov/seerstat/variables/countyattribs/ruralurban.html

```{r}
urbanrural <- read.csv("County_Urban_Metro_Rural.csv")
combineurbanrural <- merge(urbanrural, contam_by_agriandmanu, by.y="fips", by.x="FIPS")
combineurbanrural$newclass <-
ifelse(combineurbanrural$Rural.Urban.Continuum.Code.2013<3,"Urban", "Rural")
combineurbanrural$newclass <- as.factor(combineurbanrural$newclass)
combineurbanrural$X<-NULL
combineurbanrural$X.1<-NULL
combineurbanrural$X.2<-NULL
combineurbanrural$X.3<-NULL
combineurbanrural$County<-NULL
combineurbanrural$State<-NULL


aa<-tapply(combineurbanrural$agriculture,combineurbanrural$newclass,mean,na.rm=T)
barplot(aa)
aaa<- subset(combineurbanrural, combineurbanrural$newclass =="Rural")
nrow(aaa)

baseplot3 <- ggplot(data = combineurbanrural, aes(x=agriculture, y=X0,colour=newclass))
```

```
baseplot3 + geom_point(size = 0.7, alpha = 0.3)  +geom_smooth()
baseplot4 <- ggplot(data = combineurbanrural, aes(x=manufacturing, y=X0, colour=newclass))
baseplot4 + geom_point(size = 0.7, alpha = 0.3)+geom_smooth()
```

We found that the data after we combined number of violations, urban&rural classifications, industry occupation, the rural percentage for existing dataset is less than 10%, which may not be convincing to adress conclusions on urban/rural classifications

Appendix (Code)

- Creating Subsets and Writing Out CSV files

```{r}
Water_System_By_Year <- read.csv("chemicals.csv")
Earnings_By_Year <- read.csv("earnings.csv")
Industry_Pop_By_Year <- read.csv("industry_occupation.csv")
Water_Metrics_2010 <- read.csv("water_usage.csv")
Water_Metrics_Dictionary <- read.csv("water_usage_dictionary.csv")
#Census_2010 <- read.csv("CountyData.csv")
```

#Cleaning
```{r}
Water_System_By_Year$fips <- as.factor(Water_System_By_Year$fips)
Earnings_By_Year$fips <- as.factor(Earnings_By_Year$fips)
Industry_Pop_By_Year$fips <- as.factor(Industry_Pop_By_Year$fips)
Water_Metrics_2010$fips <- as.factor(Water_Metrics_2010$fips)

Water_System_By_Year$pws_id <- as.factor(Water_System_By_Year$pws_id)
Water_System_By_Year$chemical_species <-
as.factor(Water_System_By_Year$chemical_species)
```
```{r}
Water_System_By_Year
Water_Metrics_2010
```
```{r}
Water_System <- subset(Water_System_By_Year, year=="2010")
Water_System$Hazard <- ifelse(Water_System$contaminant_level=="Greater than MCL", 1, 0)
Water_System
write.csv(Water_System, file="Water_System.csv")
```
```{r}
County <- read.csv("County_Classification.csv")
County$Classification <- ifelse(County$Rural.Urban.Continuum.Code.2013<4, "Urban",
ifelse(County$Rural.Urban.Continuum.Code.2013>3 &
County$Rural.Urban.Continuum.Code.2013<7, "Metro", "Rural"))
```

```
County$Real_Classification  <- ifelse(County$Rural.Urban.Continuum.Code.2013<4, "Urban", "Rural")
write.csv(County, "Urban_Rural.csv")
```
```{r}
Water_Metrics <- merge(Water_Metrics_2010, County, by.x="fips", by.y="FIPS")
Water_Metrics
```
```{r}
#years <- c("2010" , "2011" , "2012" , "2013" , "2014" , "2015" , "2016" )
Water_System_By_Year
Water_System_5 <- subset(Water_System_By_Year, year== "2010" | year=="2011" | year=="2012" | year=="2013"| year=="2014" | year=="2015" | year=="2016")
Water_System_5$Hazard <- ifelse(Water_System_5$contaminant_level=="Greater than MCL", 1, 0)
Water_System_5
write.csv(Water_System_5, file="Water_System5years.csv")
```
```{r}
Ag <- read.csv("Agriculture.csv")
SubWater2 <- subset(Water_System_By_Year, year=="2010")
SubWater$Hazard <- ifelse(SubWater$contaminant_level=="Greater than MCL", 1, 0)
Agchemicals2 <- merge(Ag, Water_System_5, by.x="fips", by.y = "fips", all.x = TRUE )
write.csv(Agchemicals2, "Ag_Chemicals2.csv")
```

- Regression Model for Section 3

```{r}
industry_occupation <- read.csv("industry_occupation.csv")
industry_occupation$fips <- as.factor(industry_occupation$fips)
industry_occupation<- industry_occupation[!duplicated(industry_occupation$fips),]
industry_occupation$agriculture <-  industry_occupation$agriculture/
industry_occupation$total_employed
industry_occupation$construction <-  industry_occupation$construction/
industry_occupation$total_employed
industry_occupation$manufacturing <-  industry_occupation$manufacturing/
industry_occupation$total_employed
industry_occupation$wholesale_trade <-  industry_occupation$wholesale_trade/
industry_occupation$total_employed
industry_occupation$retail_trade <-  industry_occupation$retail_trade/
industry_occupation$total_employed
industry_occupation$transport_utilities <-  industry_occupation$transport_utilities/
industry_occupation$total_employed
```

```
industry_occupation$information  <-  industry_occupation$information /
industry_occupation$total_employed
industry_occupation$finance_insurance_realestate <-
industry_occupation$finance_insurance_realestate/ industry_occupation$total_employed
industry_occupation$prof_scientific_waste <-  industry_occupation$prof_scientific_waste/
industry_occupation$total_employed
industry_occupation$edu_health <-  industry_occupation$edu_health/
industry_occupation$total_employed
industry_occupation$arts_recreation <-  industry_occupation$arts_recreation/
industry_occupation$total_employed
industry_occupation$other <-  industry_occupation$other/ industry_occupation$total_employed
industry_occupation$public_admin <-  industry_occupation$public_admin/
industry_occupation$total_employed
```
```{r}
iomerge<- merge(industry_occupation, six_year_violations, by.x="fips", by.y="X6001")
iomerge$fips <- NULL
iomerge$year <- NULL
iomerge$state <- NULL
iomerge$geo_id <- NULL
iomerge$county <- NULL
iomodel<- lm(X0~ ., data=iomerge)
summary(iomodel)
```

- Graphs for section 3

```{r}
contam_by_agriandmanu <- merge(industry_occupation, six_year_violations,by.x="fips",
by.y="X6001")
contam_by_agriculture<-
contam_by_agriandmanu[order(contam_by_agriandmanu$agriculture,decreasing = T),]
top10agri <- contam_by_agriculture[1:30,]
write.csv(top10agri,"Top10 Agriculture.csv")
contam_by_manufact<-
contam_by_agriandmanu[order(contam_by_agriandmanu$manufacturing,decreasing = T),]
top10manu <- contam_by_manufact[1:30,]
write.csv(top10manu,"Top10 Manufacturing.csv")
```
```{r}
#contam_by_manufact <- subset(contam_by_manufact, contam_by_manufact$X0<=125)
baseplot1 <- ggplot(data = contam_by_agriculture, aes(y=X0, x=agriculture))
baseplot1 + geom_point(size = 0.7, alpha = 0.3)  +geom_smooth()+ylab("Number of
Violations")+ xlab("Percentage of Employment in Agriculture")
```

```
baseplot2 <- ggplot(data = contam_by_manufact, aes(y=X0, x=manufacturing))
baseplot2 + geom_point(size = 0.7, alpha = 0.3)  +geom_smooth()+ylab("Number of
Violations")+ xlab("Percentage of Employment in Manufacturing")
```
Calculate average employees in selected industry
```{r}
top10agriur <- combineurbanrural[order(combineurbanrural$agriculture, decreasing=T),]
top10agriur <- top10agriur[1:80,]
top10manuur <- combineurbanrural[order(combineurbanrural$manufacturing, decreasing=T),]
top10manuur <- top10manuur[1:80,]
totalemploymeanbyagri<-mean(top10agriur$total_employed)
totalemploymeanbymanu<-mean(top10manuur$total_employed)
```