DATA OPEN REPORT - TEAM 12

**Non-Technical Executive Summary**

In the Data Open Datathon, we are offered with multiple datasets related to the human usage of water, industries information, and water qualities.

Our team focuses on

**1. Exploring the relationship between water problems (contamination, droughts, etc.) and human activities (industrial activities, population characteristics, etc.);**

**2. Identifying the effects of education level among citizens to water usage consumption.**

**Findings:** In the end, we found that the manufacturing and government administration have the largest impact on the level of water contamination, and such impact usually happens one year later. Moreover, there is a linear relationship between educational level and the amount of water usage: less educated citizens tend to use water less wisely

**Structure:** We divided our report into three main part: data preprocessing/cleaning, in which we conduct cleaning and processing of the dataset given to us, exploratory data analysis, in which we perform preliminary data analysis and visualization, and modeling, including the model for water quality and industrial component, and the model for water usage and educational level.

**Data Preprocessing/Cleaning**

For different dataset we adopted different methods for data cleaning:

We first prepared the Chemicals dataset three different ways for different purposes.

**Chemical(1, prepared for regression)**: This dataset contains information about chemicals found in water indexed by the community water system in the county. One of our goals is to use this set to generate a metric for measuring the water quality of a county. Because there are different chemicals, water system, and contaminant value for each county, we need to wrangle the data in such a way that we obtain one single value for each county. However, due to our lack of a standard for the “proper” amount of chemicals in water, we generated 6 different values -- corresponding to the 6 chemicals -- for each county (We later ran regression for each value.)

For the metric, we thought it made the most sense to aggregate the data for a particular county and chemical by the average of the contaminant value weighted by the population. For example, the metric for Alameda County’s Uranium in 2000 is the average of the contamination values across all water systems in the county that year weighted by population that each water system serves.

**Chemical(2) :** A new variable was created as well. The first one, “population\_affected”, is the sum of the population under all community water systems in each county, if there is at least one community water system that has a contaminant\_level that’s greater than mean concentration level.

**Chemical(3):** The chemical dataset is also transformed into two matrices of time series. The first one was transformed in the format such that each row represent a unique FIP code, and each column represents a year. The value in the time series matrix is the sum of the chemical value of all chemicals in the county and in that specific year. The second one was transformed in the same format except that the total population size under contaminated regions was used instead of the chemical value as in the first matrix.

**Droughts:** To extract and correlate useful and meaningful information from this data, we did following transformation to Droughts data:

1. To align it with 2010 (as Water\_Usage is for 2010), a subset with drought conditions occurred around 2010 was extracted.
2. Based on the assumption that only severe drought conditions are strong predictors of water usage patterns, time sliced data only from exceptional drought conditions (d4) was taken into consideration for various counties.

**Earnings(1, for regression model)**: This data contains the median earnings for various industries indexed by counties. We normalized each entry for each industry each year by dividing it by the total working population that year. We replaced “2,500-” and “250,000+” with “1,500” and “300,000” respectively to have numerical values to work with.

**Earnings(2):** A new variable, “industry\_size”, was created for each region by multiplying the number of median income in industry A, by the number of people working in industry A. In order to deal with industries without population information, we just use the average working population in the region. Furthermore, the matrix was transformed into a new time series data in the same format of the one in chemical data set. So the column becomes years from 1999 to 2016 and the row becomes FIP for each county.

**Education\_attainment:** This dataset was sliced to generate a subset about education level from the recent year (2000-2016) as this was further linked to water usage patterns (which belong to the year 2010). Therefore, time-based slicing was performed to extract recent data (1.5 decades). Also, two new derived columns after aggregating county were produced for future model building and prediction:

1. Total\_Population\_Per\_State
2. Total\_Less\_Educated\_Per\_State (Population belonging to Less\_than\_High\_School\_Diploma per state)

**Industry\_occupation(for regression model):** This data contains the estimated working population for various industries indexed by counties. We normalized each entry for each industry each year by dividing it with the total working population that year. For the missing values in this dataset, we replace them with the industry-based average percentage working population on that particular year.

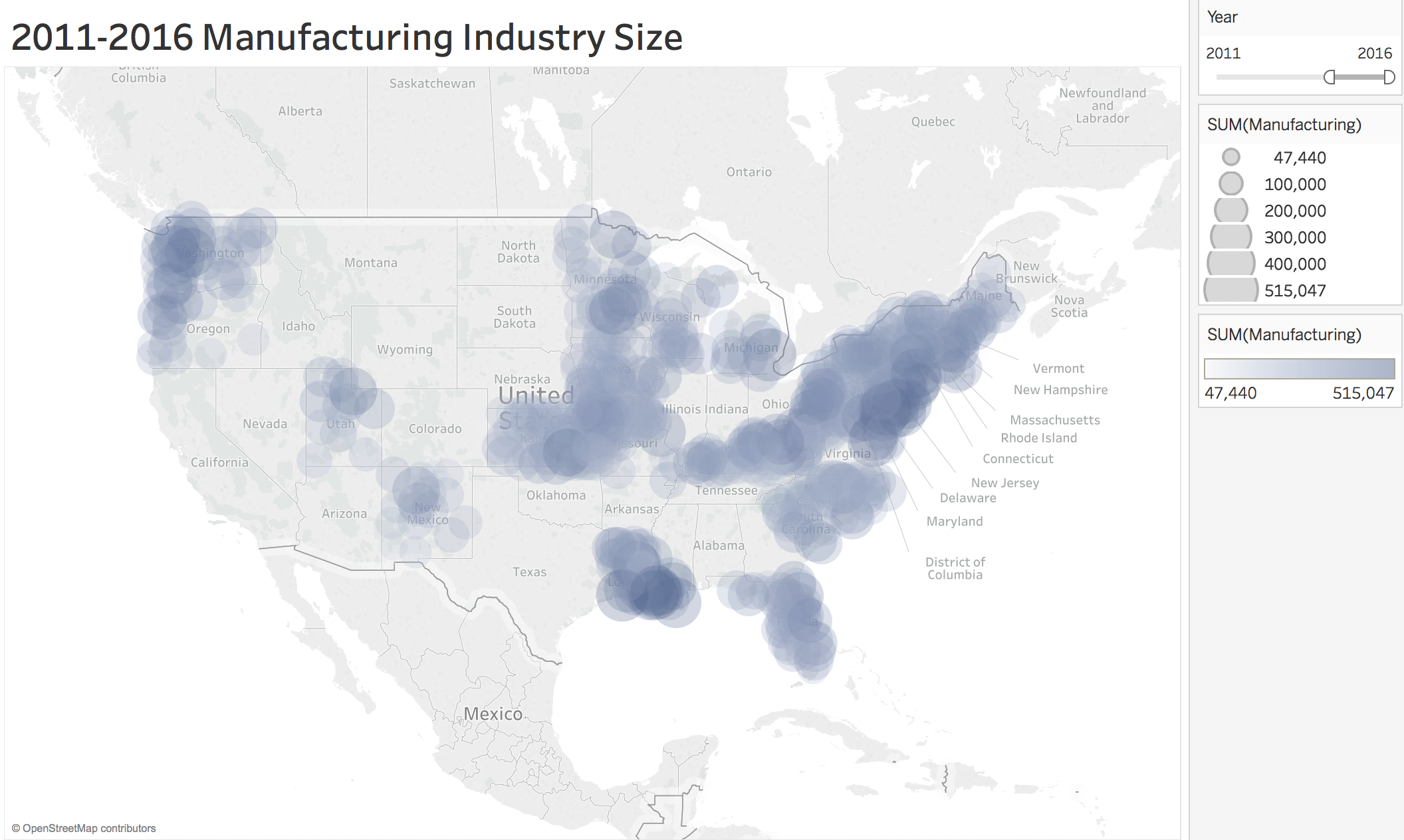
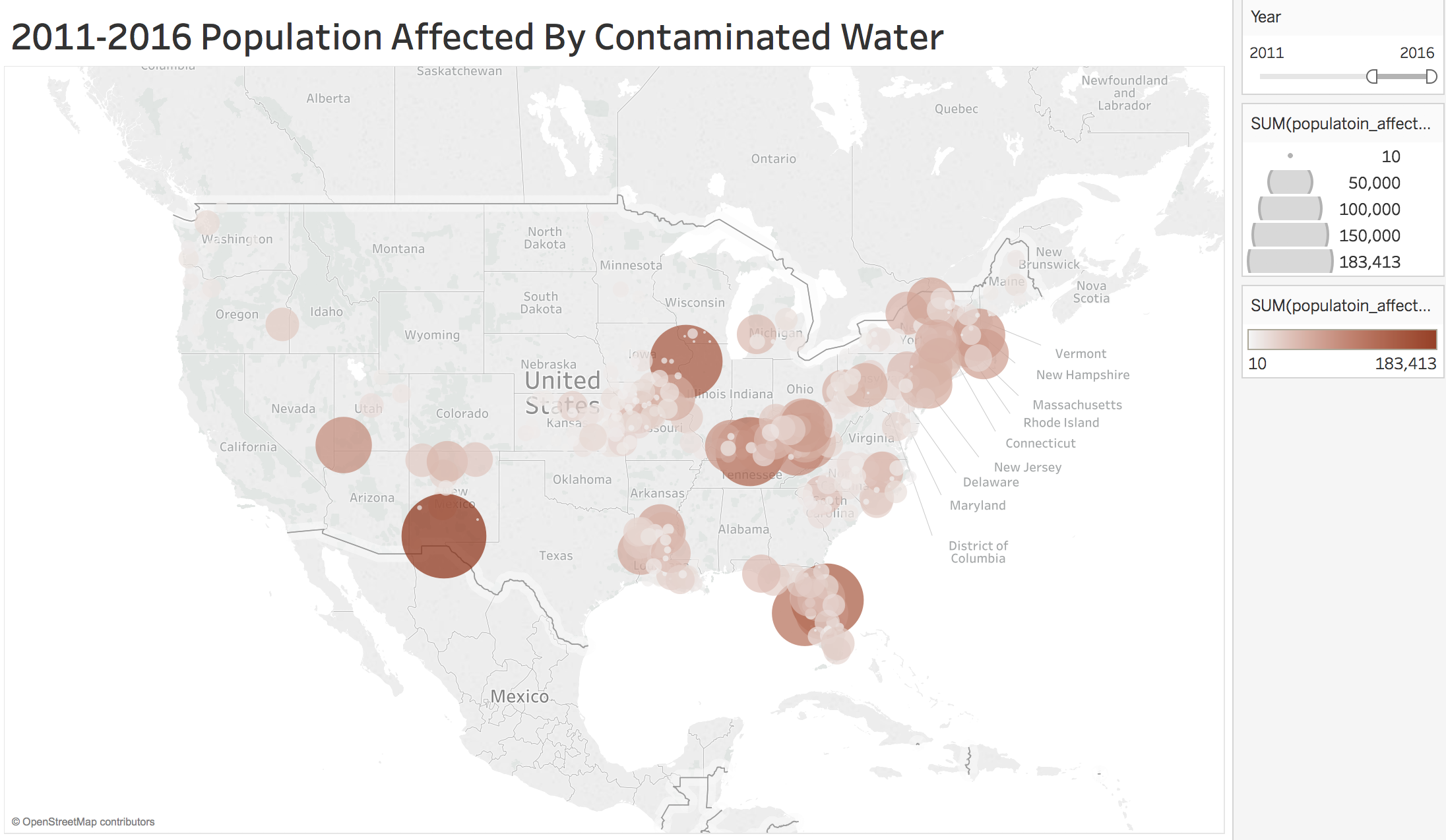
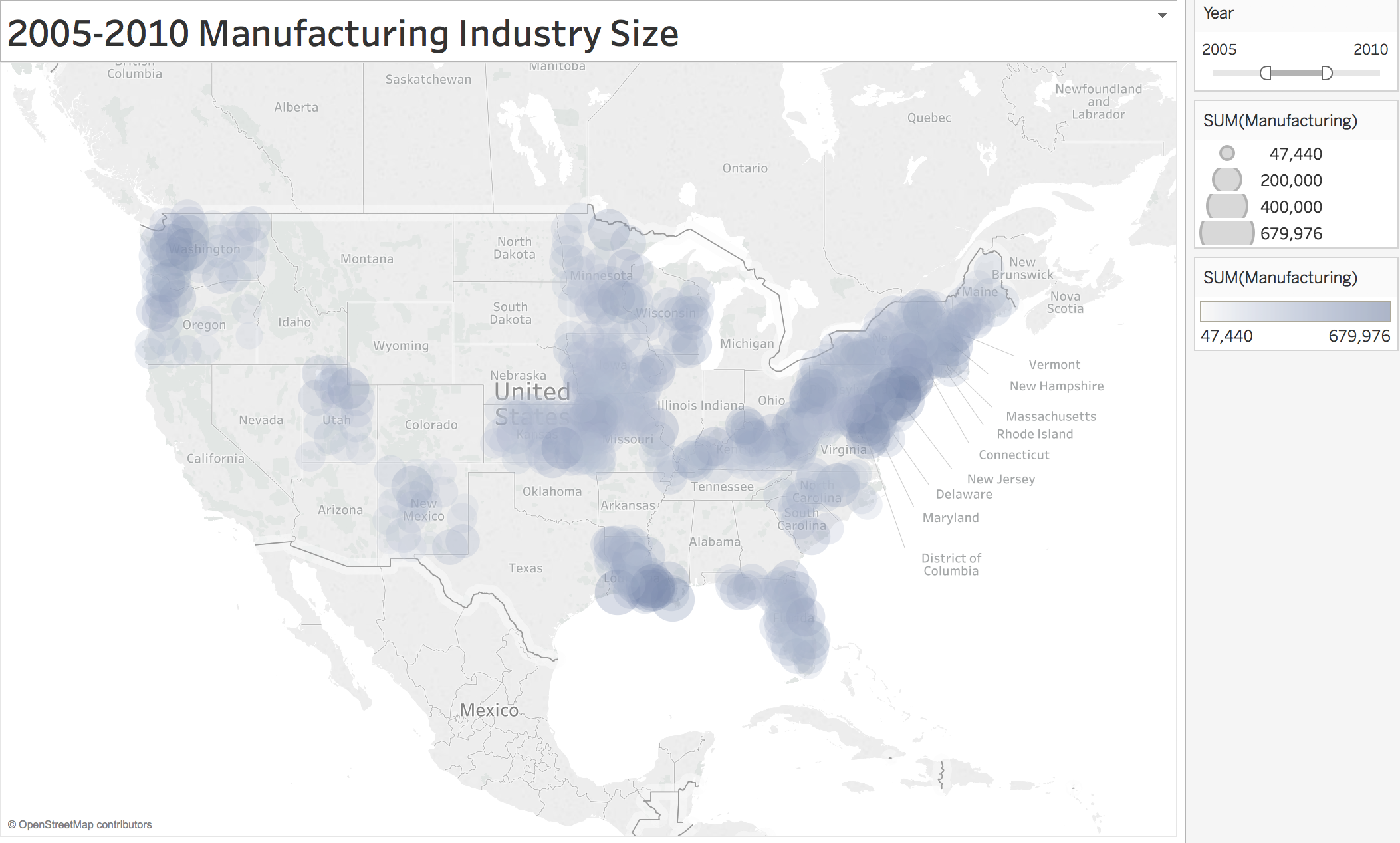
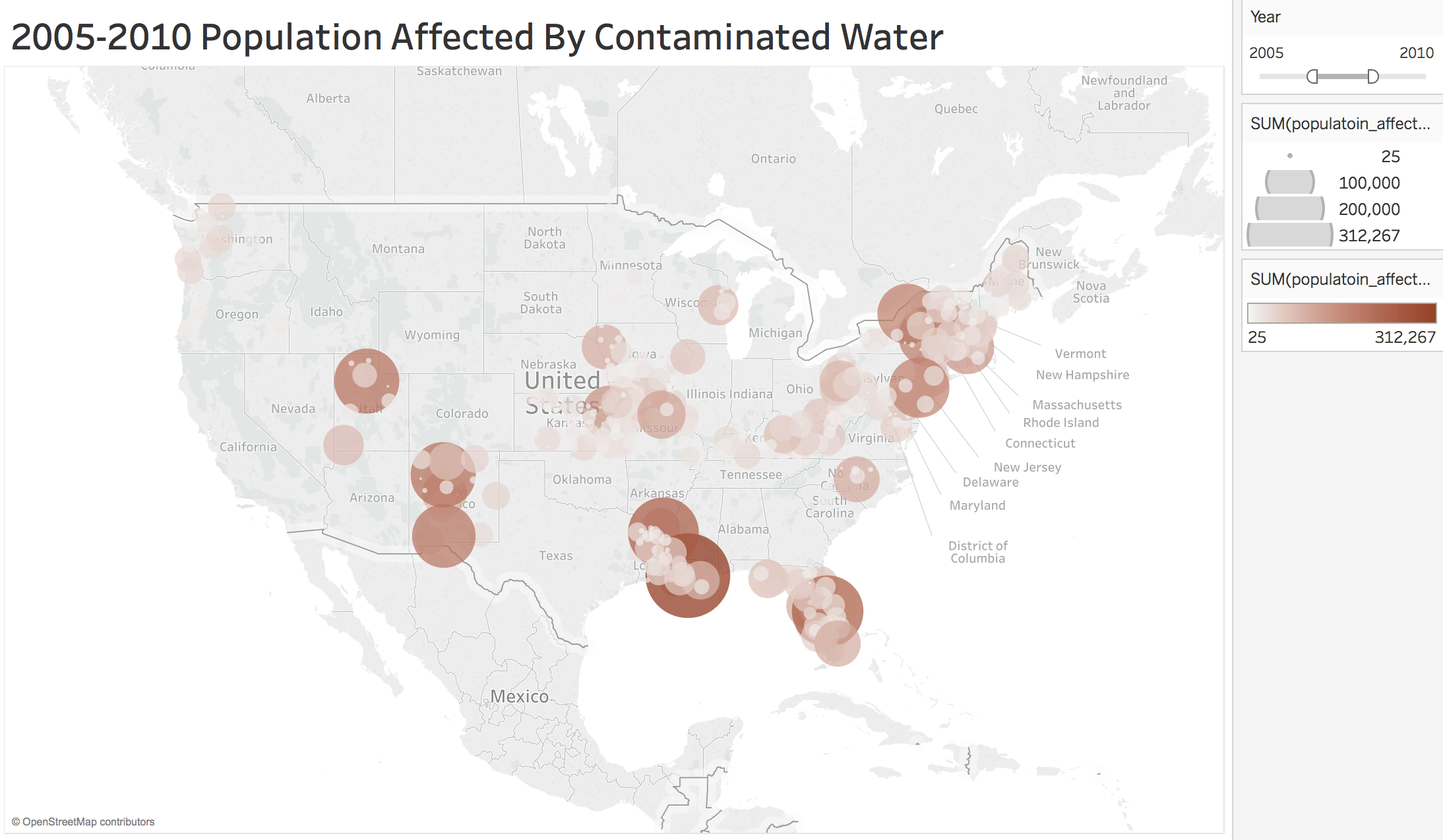
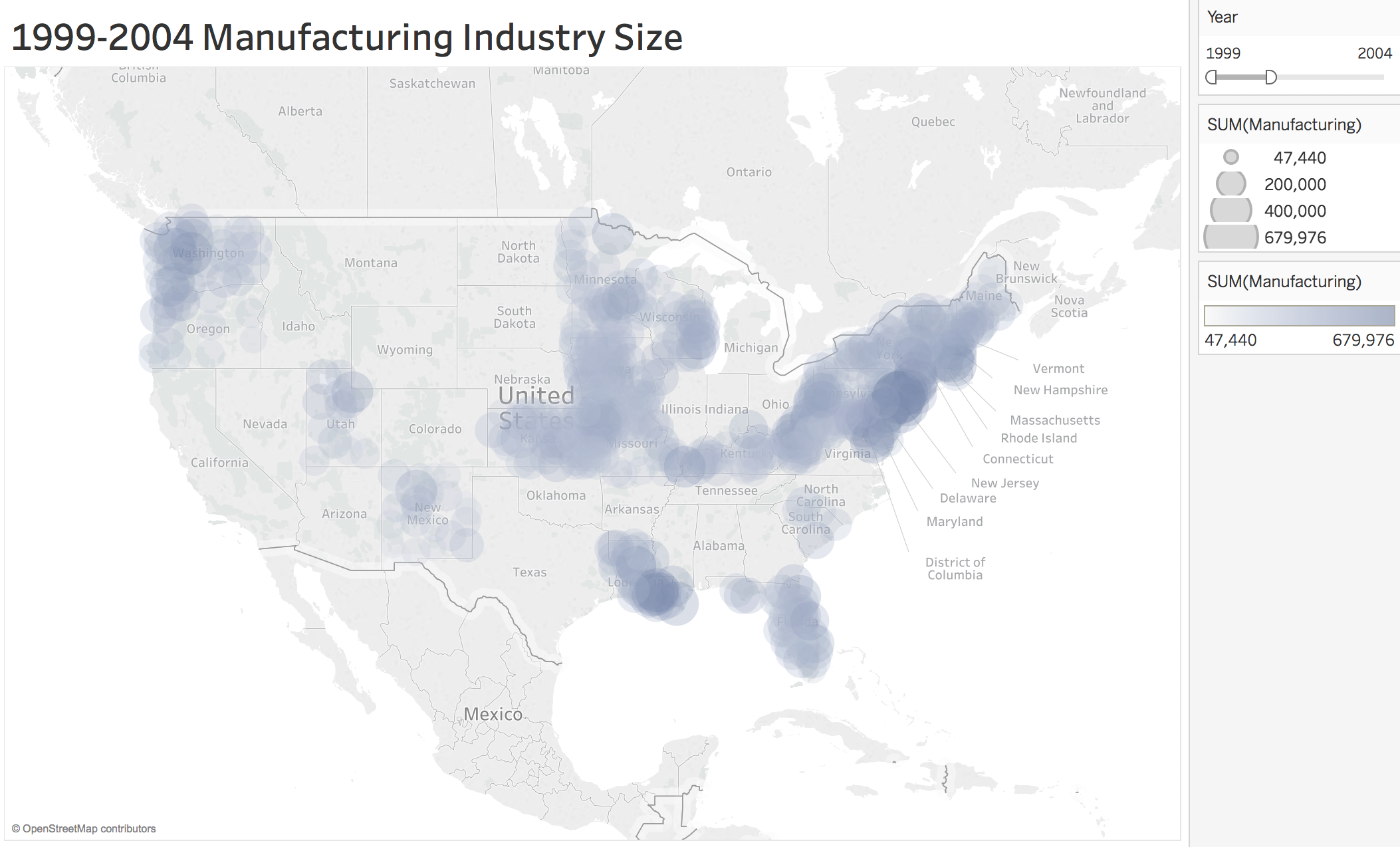
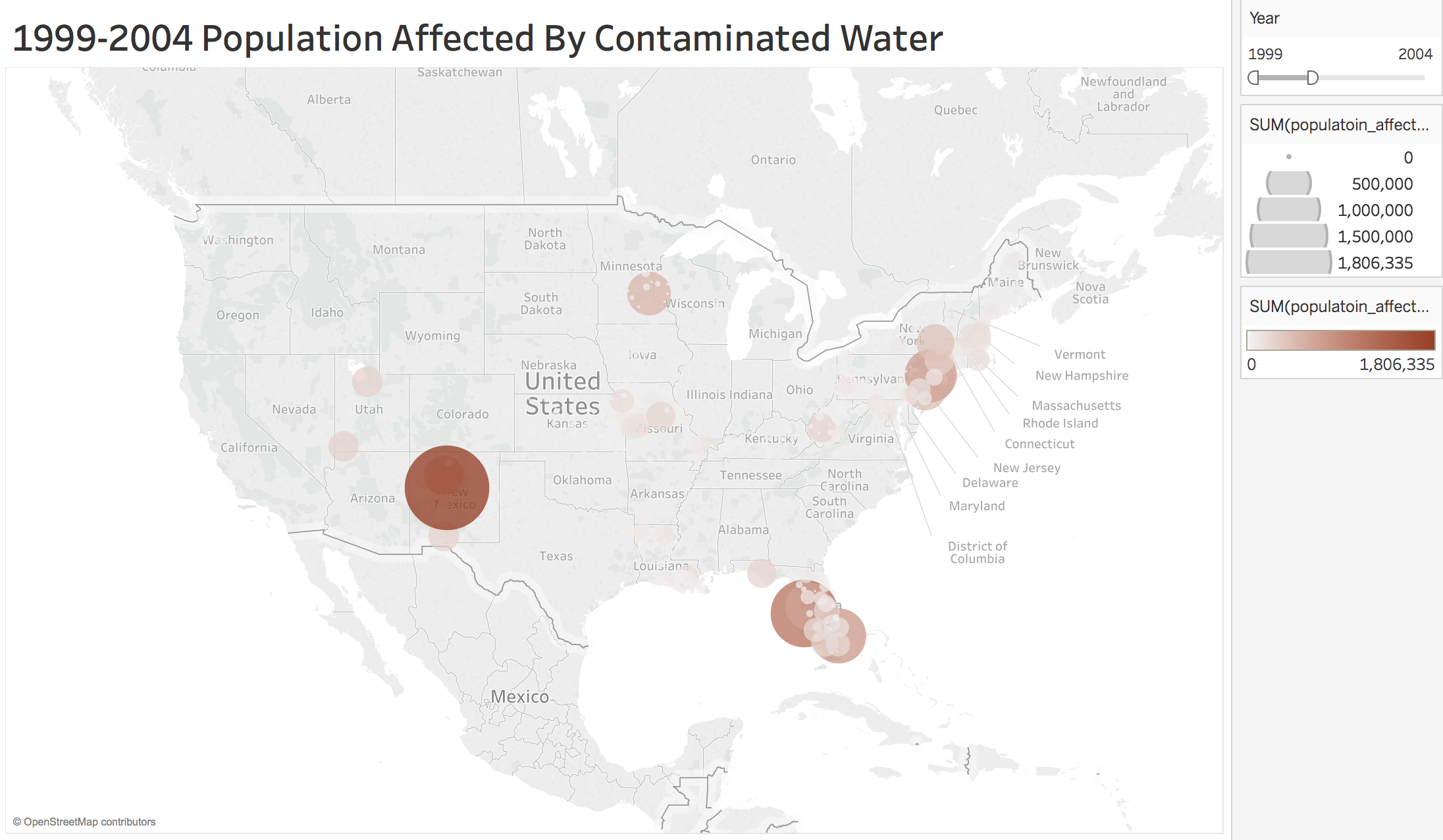
**Water\_usage:** The fact that this data has water usage information from the year 2010 only decides how this dataset will interact with other datasets. Feature Engineering using aggregated methods were performed on this dataset to extract only relevant features (States, County, Population, Total Usage, Irriagation\_Total\_Usage, Mining\_Total\_Usage, Thermoelectric\_Total\_Usage, Livestock\_Total\_Usage and Overall\_Total\_Usage) used to draw insights in subsequent steps.

**Exploratory Data Analysis**

In order to answer the relationship between industries and the quality of water, we explore the role that the size of an industry has on the population under contaminated regions.

Details:

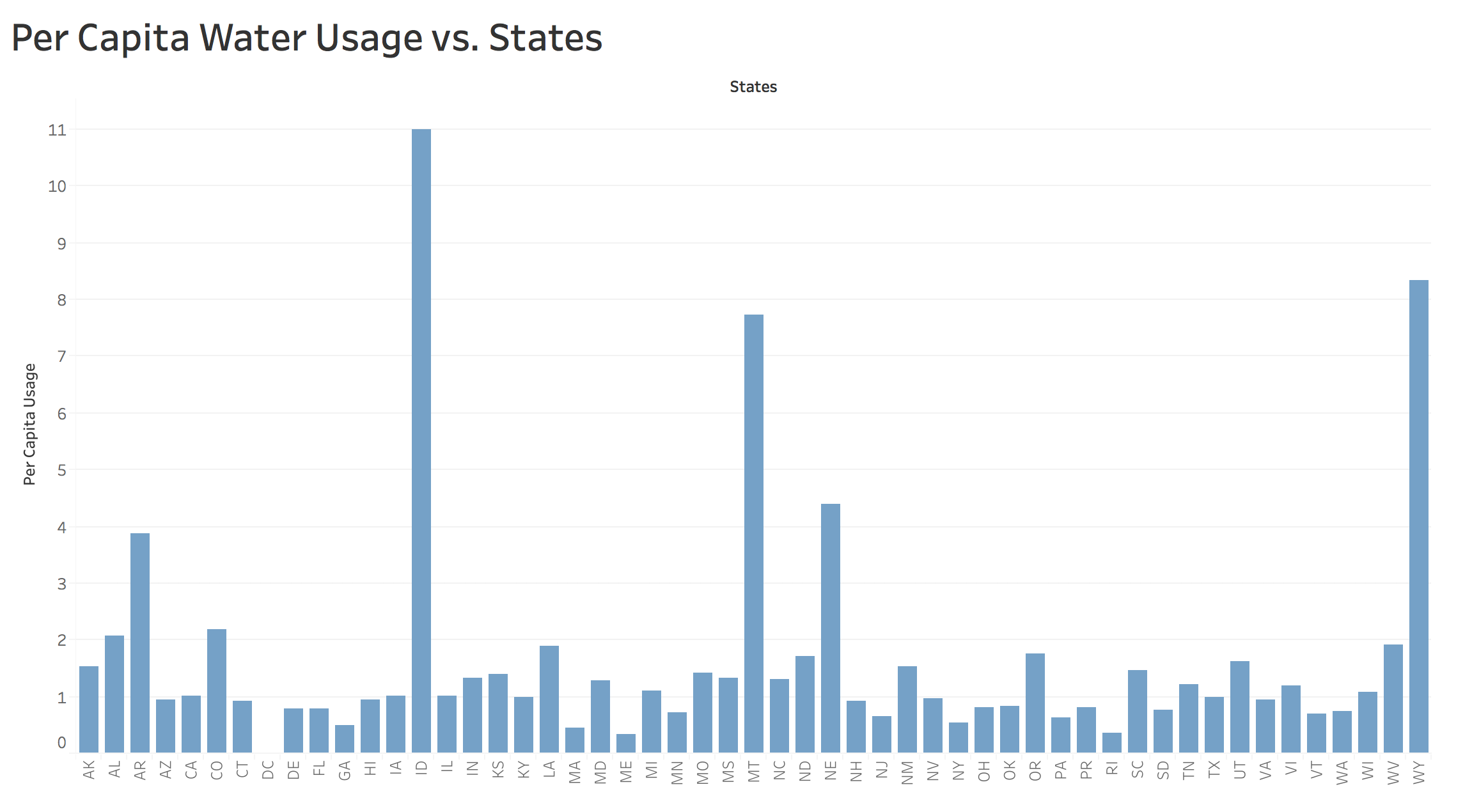
Among all the regions, the manufacturing industry has the largest correlation with the population in contaminated regions. As we can see from the three charts from three time period, the contaminated population have the tendency to spread from initially only a handful of regions, gradually moving to the geographical east and south side of the country.



Per Capita Water Usage across US States :

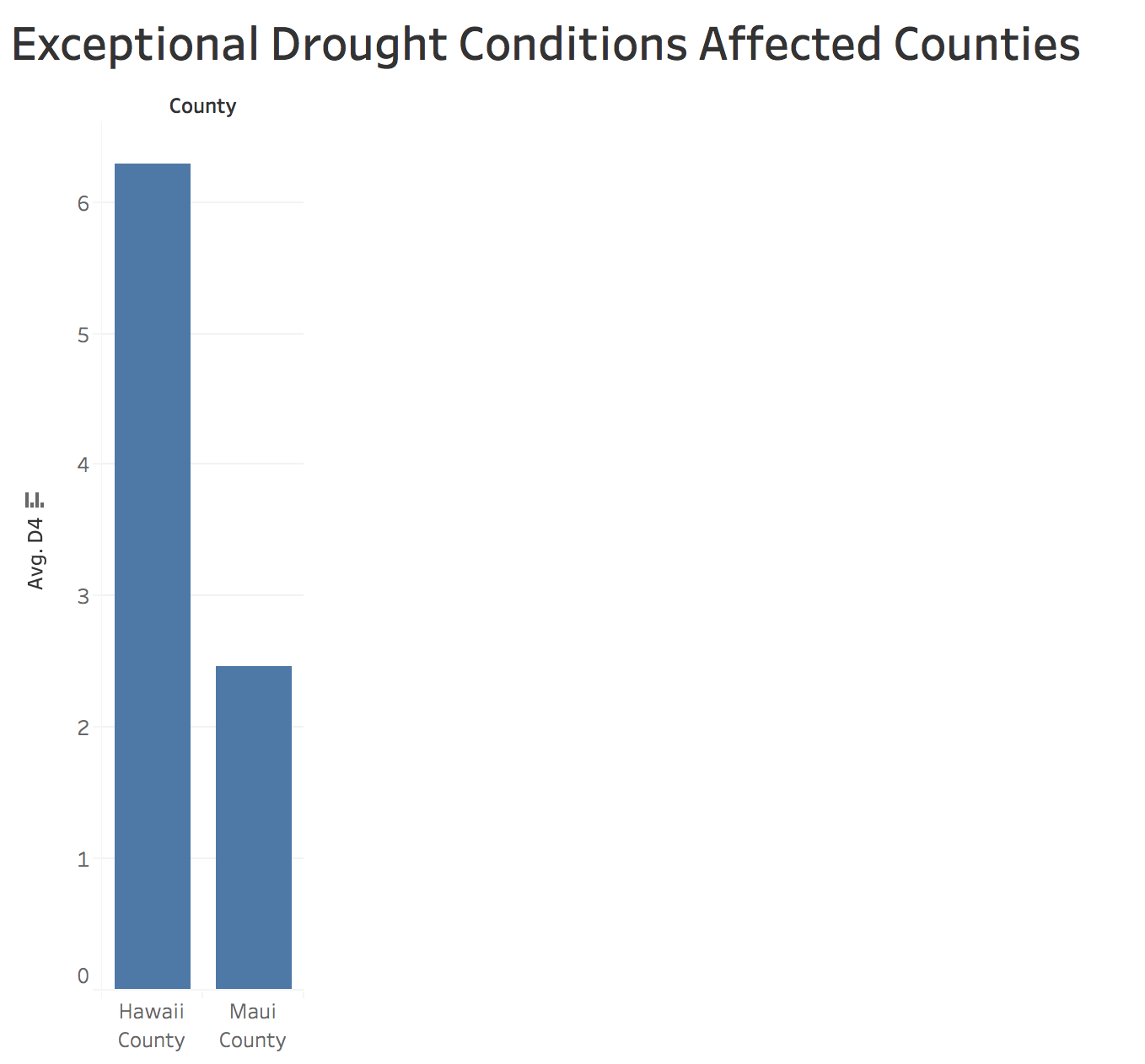
To estimate the water usage patterns across different states in the US, we came up with new measure - **Per Capita Water Usage**

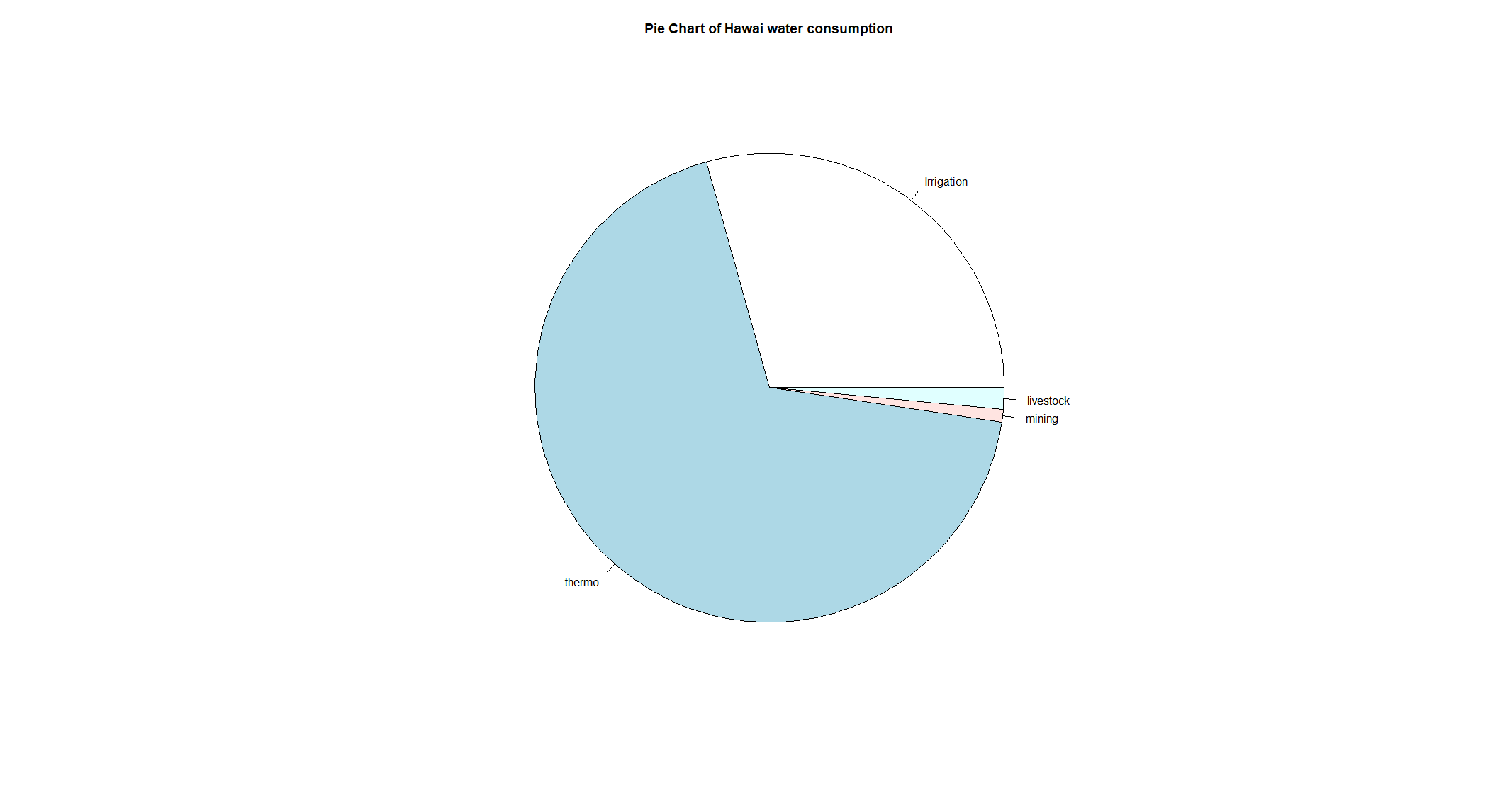
**Per Capita Water Usage = Water Consumption by all the counties in a state / Total Population of a State**

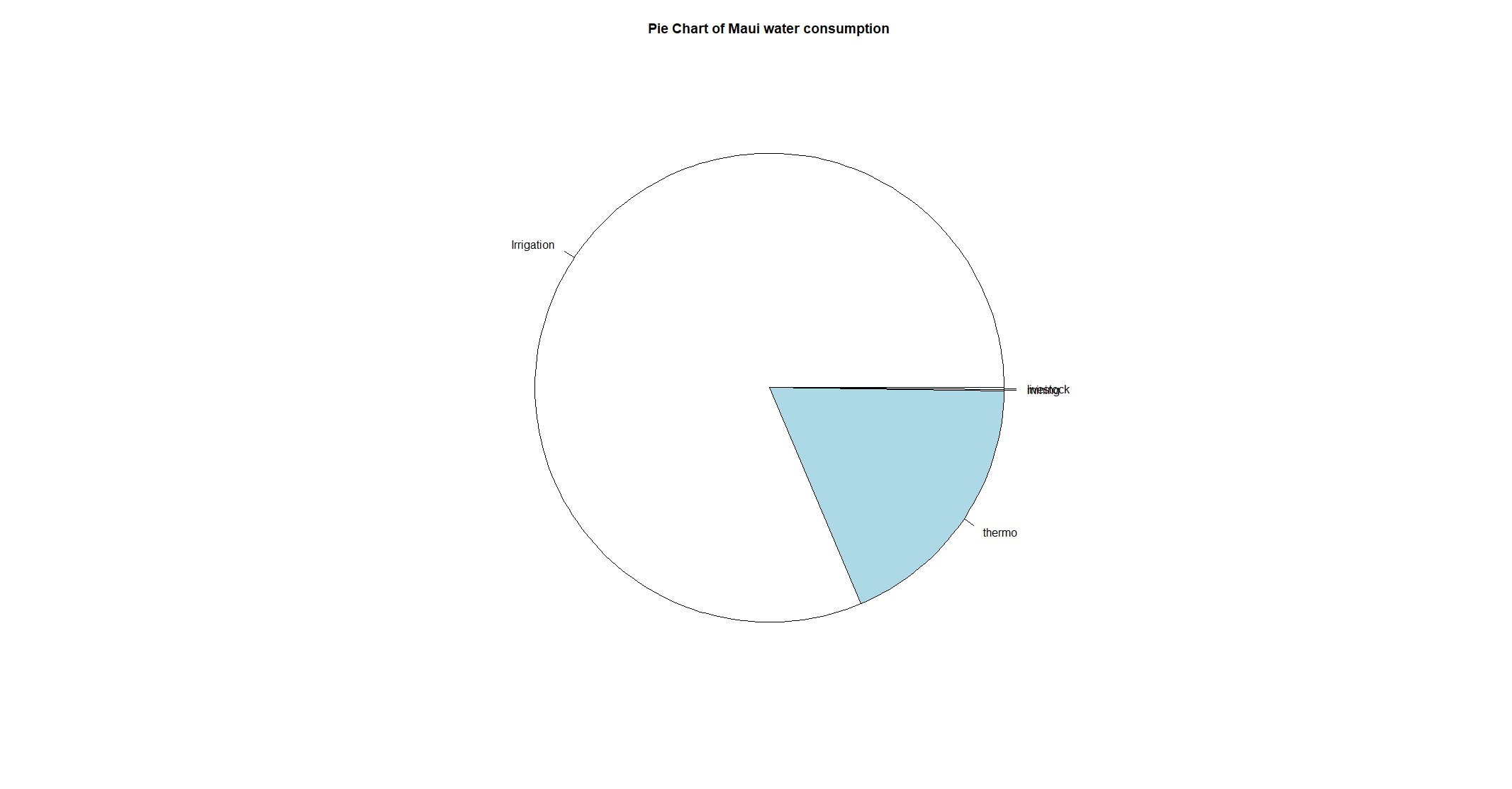


As we can see from the above plot, there are few states (ID, MT, NE and WY) with exceptionally high consumption of water per person. In subsequent steps, we go in-depth to reason this pattern.

After exploring Drought dataset, we found that following county (Hawaii and Maui) were hit severely by d4 drought conditions. The subsequent pie charts show the water consumption pattern in these counties:







**Thermo Projects and Irrigation are the potential factors to create drought-like conditions.**

**Modeling/Prediction**

**Prediction of Water Quality Based on Industry**

Besides the exploratory data analysis, we also conduct regression analysis on the relationship between the industrial component and water quality. Specifically, we tried to fit a regression model to predict the water quality of a particular county given the industrial information of that county.

Feature engineering

The features we used to perform our analysis includes the data from Industrial\_occupation and earnings dataset, after data cleaning and preprocessing. Hypothetically, since different counties with different industrial components would have different impact on their environments. For instance, a county with large manufacturing/oil extraction plants would be more likely to have poor water quality than a county focusing on education/entertainment industry. Example features are percentage of working population in the manufacture industry (industry\_occupation$manufacture), the median earnings for employees in mining, quarrying, oil/gas extraction(earnings$mining\_quarrying\_oilgas\_extract), etc.

In addition, we also include the lag 1 variables for each of the features described above. The main reason is we think it will takes a little bit time for the human activities to have the true effects on the environment. For example, when a factory disposes contaminated water to the ecosystem, it usually takes half a year to detect such contamination. In this case the lag 1 variables will be very informative features.

Model fitting

To start with, we state the main assumptions, which we sort of implied above:

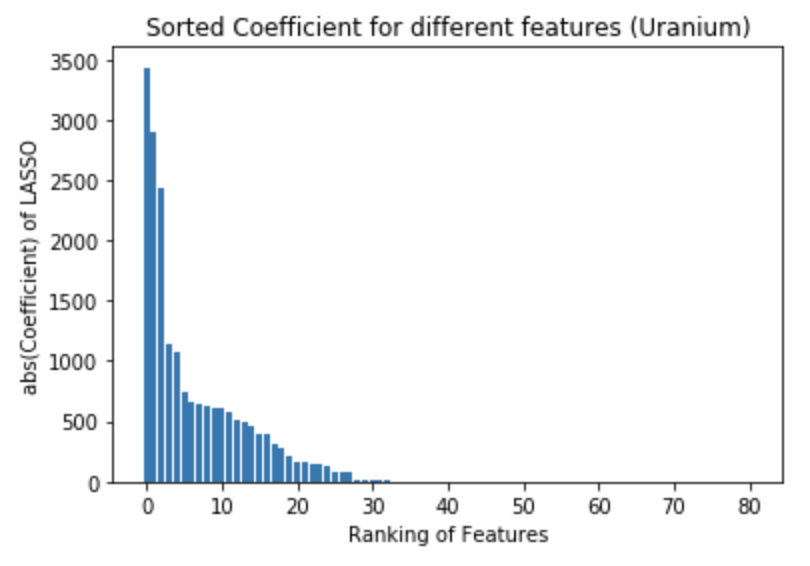
1. There exists a linear relationship between industrial components of a county, and the water contamination level of a county.
2. The water contamination and the industrial development are correlated with time, and there is a linear relationship between the corresponding lag variables

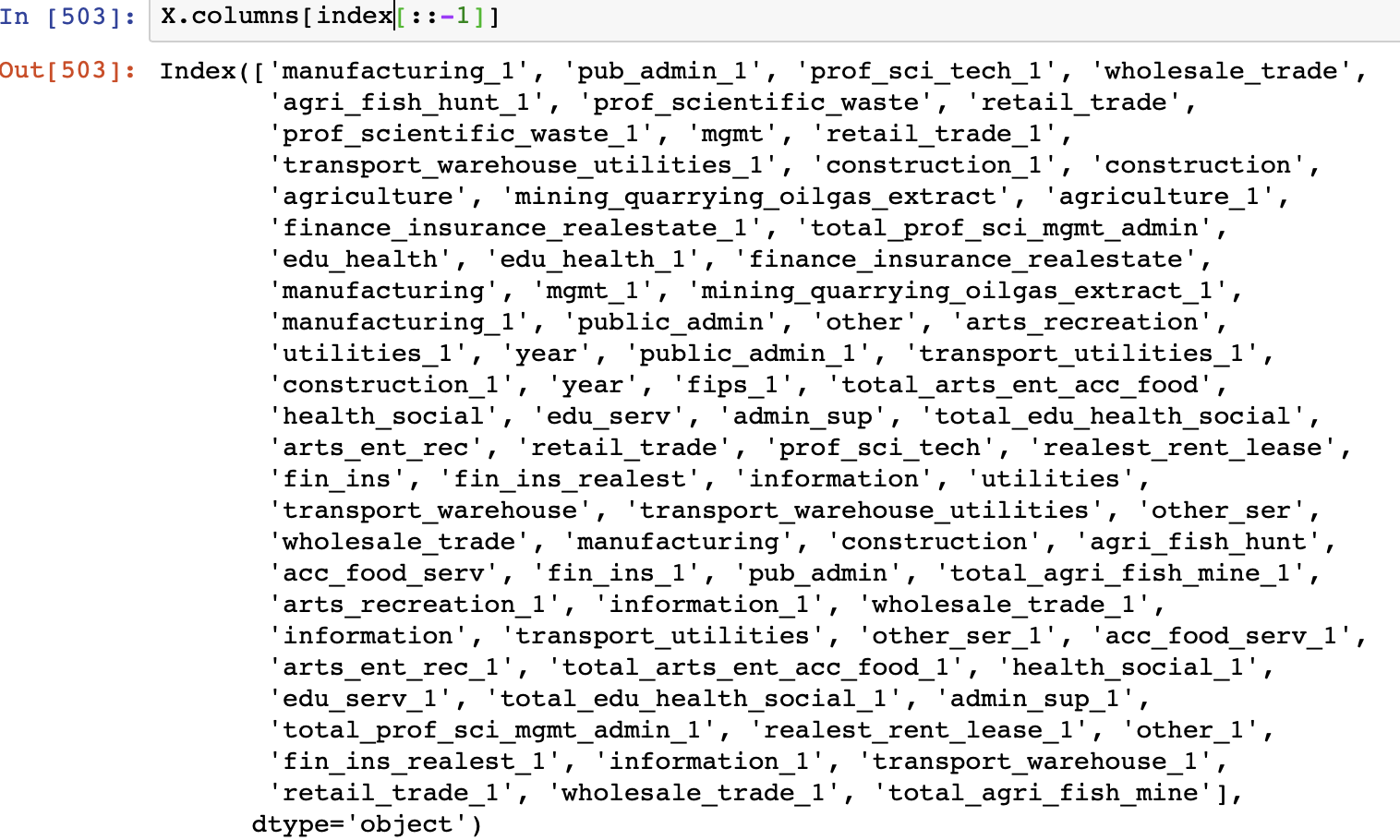
These two assumptions are well-supported by our exploratory data analysis section above:

1. The high positive correlation between industrial development and water contamination implies a possible linear relationship
2. Both the water contamination and industrial development are highly correlated across time. For example, we can see that the location of the water contamination stays the same for the course of ten years.

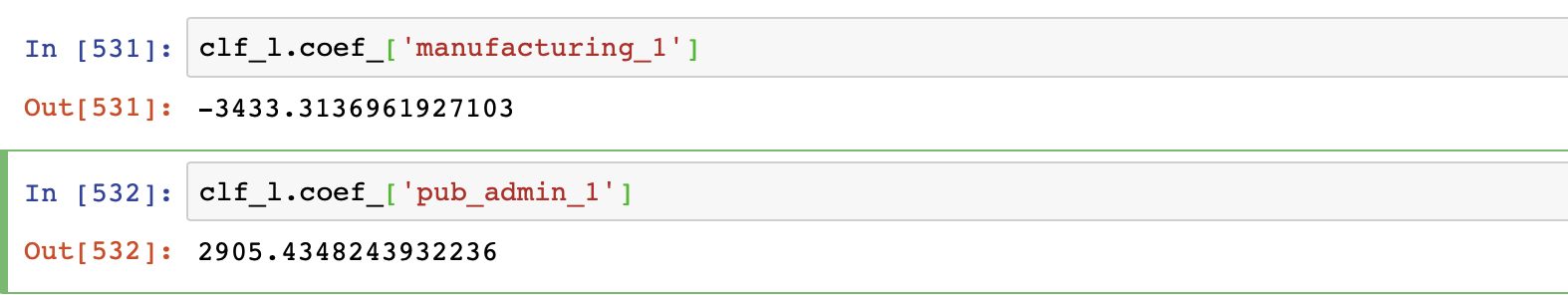
We decided to only show our model against the Uranium contamination data because we assumed that Uranium is the most harmful out of all of the substances and should hold priority for analysis. In particular, we get to see which industry potentially contributes to uranium contamination. Interestingly, we saw similar results when applying our model to predict the contamination level of Arsenic, DEHP, and Halo-Acetic Acid.

We utilized LASSO regression and Cross-validation to further validate our assumptions and investigate which features would have the biggest impact on water quality. For each different chemicals, we performed a grid search on the LASSO L1-penalty parameter alpha and decided to fit a LASSO model with alpha = 0.04. Then we sorted the absolute value of the coefficients, and plot the following graph (in this case, Uranium). The three spikes at the left side of the graph indicate the huge impacts on water quality. Interestingly, our LASSO model indicates that these features are: Manufacturing lag 1, Public administration lag 1, and professional science technology lag 1. We interpret that these are the most important factors that influence water quality. And such impact would usually come at the next year.





Furthermore, the coefficient value for manufacturing lag 1 and public administration lag 1 has opposite coefficient, indicates that these two factors have opposite impact on water quality.



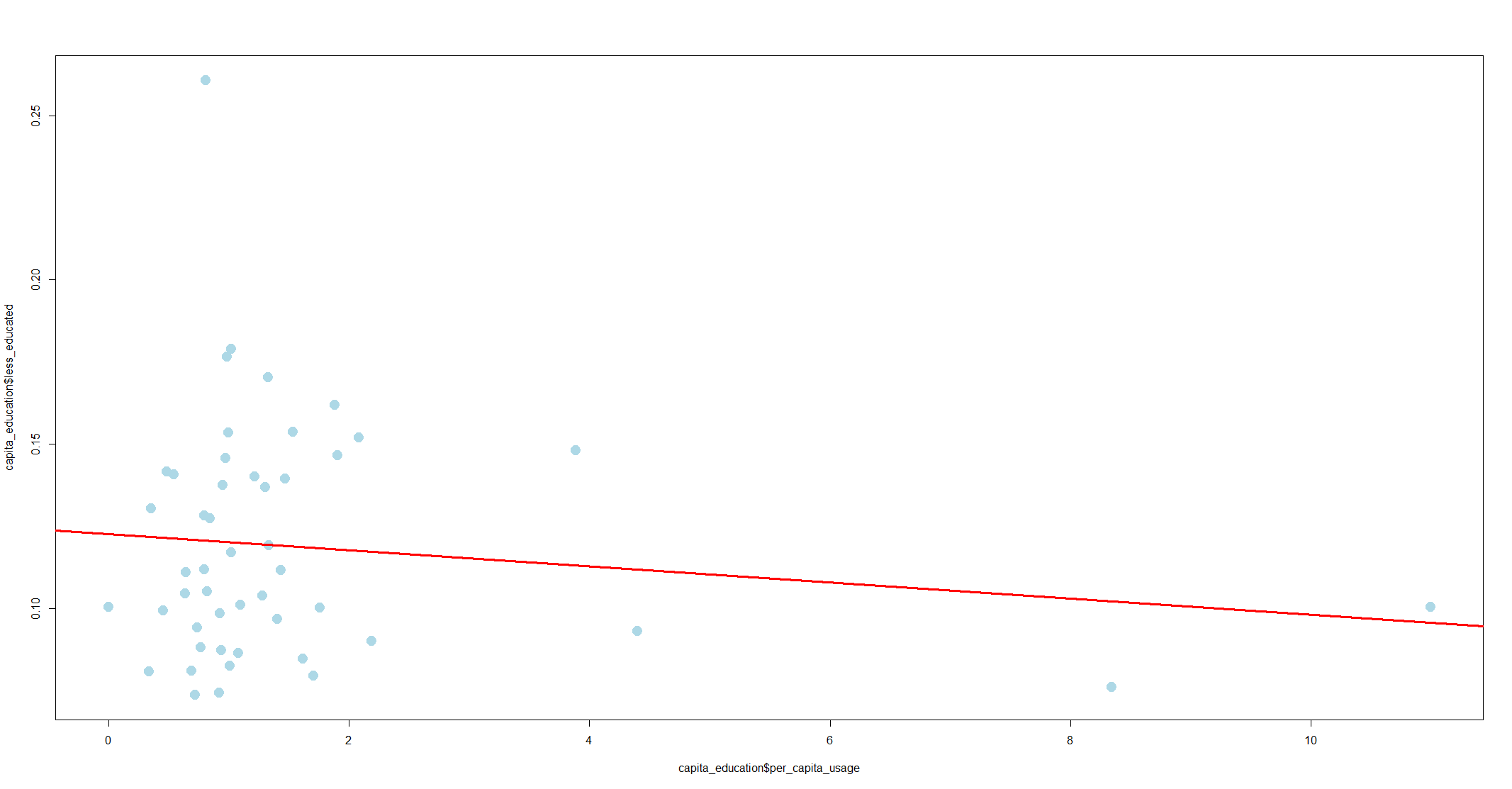
**Water Usage Prediction based on Education Level of Citizens**

Surprisingly, the education data of citizens from recent census delivers strong correlation and connection with the water usage across the states.

**Assumption:**

We assumed the fact that literate or well-educated people are much aware of the water crisis and tend to save the environment by using water wisely.

This can be seen from the following regression plot over **Per Capita Water Usage** and **Percentage of Less-Educated Citizens**.



**Model:**

**The regressor predicts that with the increased percentage of less-educated citizens, the per capita usage of water increases.**

This is evident through the regression line as well.

**Outliers :**  Outliers here are the states where per capita usage is still high even with a lesser percentage of less educated people. This can be justified by the fact that these states have either huge industrial base ( like Arkansas) or huge thermal projects (like Alabama) or may be vast irrigated lands.

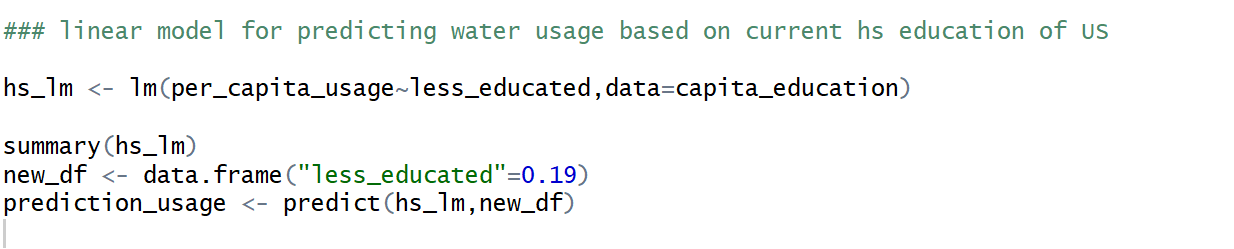
**Prediction:**

**Test data -**

The nation's high school graduation rate hit **81 percent** in 2012-13, the highest level since states adopted a new uniform way of calculating graduation rates five years ago. "America's students have achieved another record-setting milestone," U.S. Secretary of Education Arne Duncan said.

*Source :* [*https://www.ed.gov/news/press-releases/us-high-school-graduation-rate-hits-new-record-high*](https://www.ed.gov/news/press-releases/us-high-school-graduation-rate-hits-new-record-high)

This gives us the test data as 0.19 (Percentage of people without high school graduation = 19%) for our regression model.

Code snippet of a linear model predicting on current education rate

**We found the per capita water usage for United States to be around 1.06 which completely aligns with the values we found for each state across the US.**

**Conclusion**

In conclusion, we found that the manufacturing and government administration have the largest impact on the level of water contamination, and such impact usually happens one year later. Moreover, there is a linear relationship between educational level and the amount of water usage: less educated citizens tend to use water less wisely.