Team 30 - DS4A Colombia 4.0

Relationship between water quality and money in the US: A data analytical approach

Margarita Palacios Vargas, Julián Gómez García, Leonardo Briceño Montaña, Cristhian Córdoba, and Joham Alvarez Montoya

OVERVIEW

In this submission, we present the ***Exploratory Data Analysis*** of our final project of the DS4A Colombia 4.0 course.

INTRODUCTION

Water is the key to living. Most of our problems with politics and economics are small compared with a possible future without access to clean water. The issue of not being able to get a high-paying job can be negligible compared to the problem of not having clean water. However, some people think that water and environmental care produce a decrease in productivity, therefore, earnings and profitability of industries are decreased and people become consequently poorer (Clevenger, T., & Herbert, M., 2020).

The aim of this project is to discover hidden patterns between water quality and money in order to get insights into whether environmental carelessness produces more money as many non-environmentalists argue. There are also people arguing that better environmental indices is a function of countries' earnings and development.

To solve this, the following research question is proposed: **"Is there a relationship between profitability/income and water quality in United States Counties?"**

To accomplish this, three datasets were selected: chemicals provided by the Center of Disease Control and Prevention (information about water quality from 2000 to 2016 in the US), earnings from the US Census (information about profitability from 2010 to 2016 in the US) and industry occupation also from US Census (information about the location of industrial sectors from 2010 to 2016). In addition, as a reference dataset, water usage provided by the US Department of the Interior (information about the share of water usage in the US from 2010) was also used. External datasets such as region classification and county population were aggregated to the study.

METHODOLOGY

The Exploratory Data Analysis (EDA) can be divided into 5 steps:

I. **Data Wrangling**

- **Categorical and Numerical Variables Exploration:**
    - Describing numerical variables is the perfect way to understand in what terms the data is presented and if there could be interesting transformations to be held. However, everything seemed fine in this aspect.
    - Categorical variables can be more troublesome when their volume of possibilities is high. It was discovered that there are Counties with the same name in several States, so their FIPS codes had to be used in the assessment, which is unique.

- **Null values:** We looked up for missing values and the way this data needed to be handled. Most of the data sets were complete but, at the moment of merging, some missing values were found in the process.

II. **Exploratory Data Analysis (EDA)**

- **Merging:** After understanding the data, there was a need to merge the most important variables to evaluate. Luckily, our datasets had keys to merge among them (FIPS code and year).
- **Deeper exploration:** Interesting plots building, to visualize relationships of interest to construct models later on.

III. **Linear Model Building:** With the ingredients defined in the EDA, linear explanations of the variable of interest were assessed.

1. DATA WRANGLING

a. Water quality / Chemicals

Water quality data was retrieved from a table containing data from several water systems/sources per US county, their respective measure of each traced contaminant chemical present per reported year and the population each source served. The measure unit for each chemical was micrograms per liter, except for nitrates which used micrograms per liter.

Reports of six deleterious chemicals were included, including nitrates, trihalomethane, haloacetic acids, arsenic, DEHP and Uranium. The most reported chemical in the dataset was nitrates, less than 5% of the measures reported levels above the maximum contaminant level (MCL) of the total data span.

Because the assessment focused on conclusions upon Counties, having the information per water system was inconvenient. The value of contaminants was weighted per population served by each water system.

b. Earnings

This data contained, per year (2010-2016) and per County, the median of income adjusted by inflation. It also contained respective FIPS codes to ease the merging and had almost all the counties that the previous dataset reported.

c. Industry

The same keys the previous datasets had could also be found here: year (2010-2016) and FIPS code. This data contained the estimated working population (16 years and over) for various industries as well.

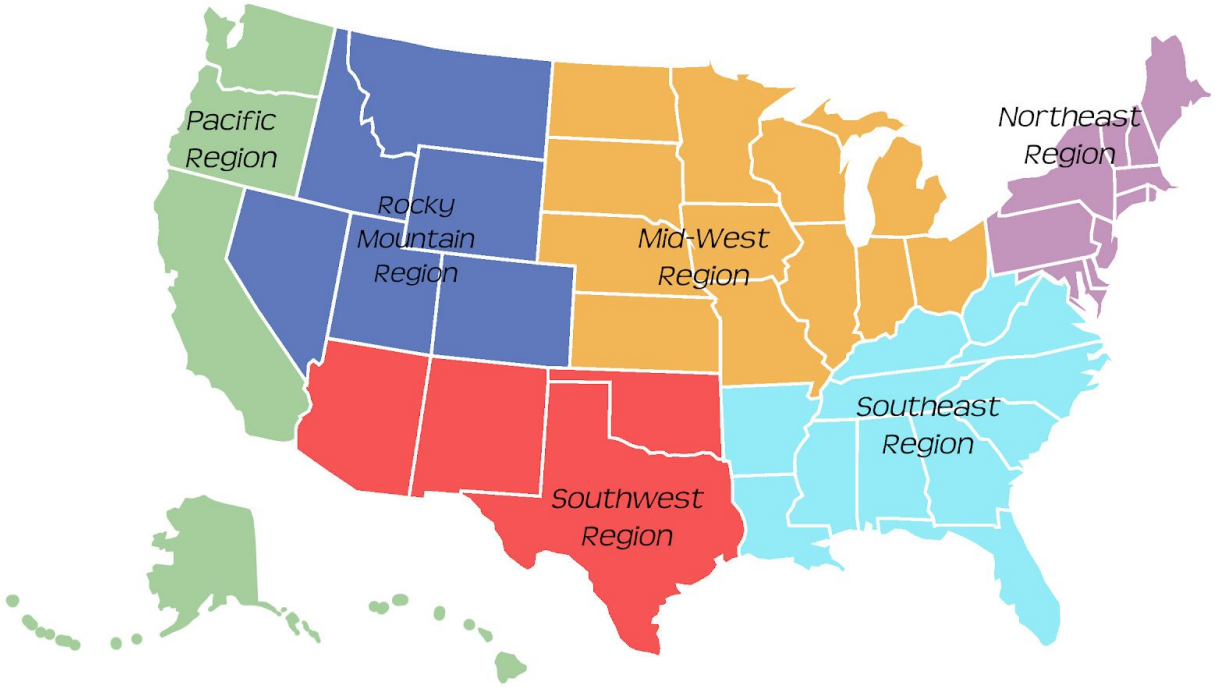
d. Water Usage

This dataset contained a lot of information about the origin of water supplies and their uses. However, for the purposes of this analysis, the population variable was the only one used.

e. Region

The purpose of this assessment is to see if there are influences among the income of different counties. Therefore, a geographical evaluation is inherent. To simplify meddling with coordinates we defined the following regions within the USA:

Graph 1. Regions of the USA



Source: https://www.blendspace.com/lessons/Z68uhrZRNQZxLA/regions-of-the-united-states

In this map it is shown that the northwest region is also called the Rocky Mountain region.

f. Political Party Affinity

Though it is more complex than that, politics in the US are commonly simplified to adherence to the two traditional american parties: Democratic and Republican. Using historical data of the Counties in 2016, the affinity of a County to a certain party was defined by the votes casted.

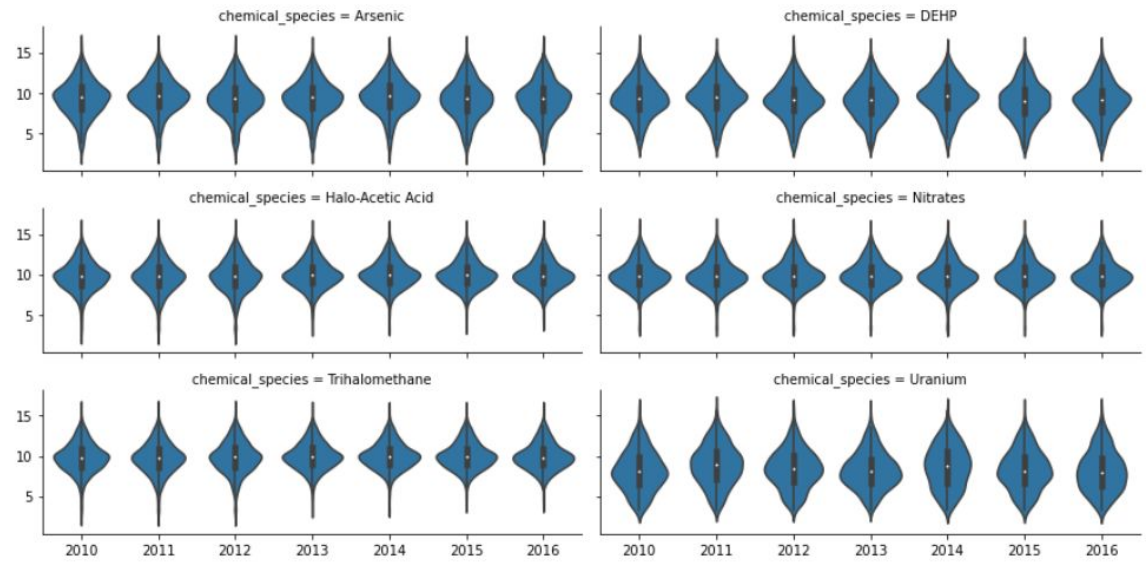
Time series Handling

The data has different time frames and is imperative to take acerted decisions, because the objective is not to predict values through time, and the way it is organized can lead to auto-correlation problems.

- **Population Served by Water Systems/Sources**

In the following graph it can be seen that, per chemical, the behavior is almost the same as its distribution:

Graph 2. Number of report of chemicals concentrations in community water systems in the US, from 2010 to 2016, per year



Source: own elaboration.

However, ANOVA Tests were conducted to statistically conclude.

Table 1. ANOVA Tests results - years and population.

	sum_sq	df	F	PR(>F)
year	1.179739e+11	1.0	0.835235	0.360768
Residual	6.240269e+15	44180.0	NaN	NaN

It can be concluded that there is no statistically significant difference between the population served through the years. However, there is neither a comparison with chemical species nor with its interaction with the year. Two-way ANOVA is conducted.

Table 2. Anova Tests results - years, population and chemicals

	sum_sq	df	F	PR(>F)
chemical_species	3.608927e+12	5.0	5.112282	0.000109
year	1.306431e+11	1.0	0.925322	0.336087
year:chemical_species	4.518991e+11	5.0	0.640145	0.669073
Residual	6.236208e+15	44170.0	NaN	NaN

- In fact, there is no statistically significant difference between the population served through the years.
- The interaction between years and chemicals is not significant.
- However, there seems to be a difference between the population served among chemicals.
    - Say our significance is at a 1%.

Table 3. Anova Tests results - years and chemicals

Nitrates

	sum_sq	df	F	PR(>F)
year	7.745035e+10	1.0	0.417083	0.518412
Residual	1.688898e+15	9095.0	NaN	NaN

Arsenic

	sum_sq	df	F	PR(>F)
year	2.523391e+11	1.0	1.360332	0.243515
Residual	1.477308e+15	7964.0	NaN	NaN

Trihalomethane

	sum_sq	df	F	PR(>F)
year	2.733231e+10	1.0	0.268357	0.604449
Residual	9.246009e+14	9078.0	NaN	NaN

Halo-Acetic Acid

	sum_sq	df	F	PR(>F)
year	4.146399e+10	1.0	0.411491	0.52123
Residual	9.146456e+14	9077.0	NaN	NaN

DEHP

	sum_sq	df	F	PR(>F)
year	1.203680e+11	1.0	0.940079	0.332299
Residual	7.153610e+14	5587.0	NaN	NaN

Uranium

	sum_sq	df	F	PR(>F)
year	6.358842e+10	1.0	0.415661	0.519154
Residual	5.153948e+14	3369.0	NaN	NaN

Within each chemical the median of population is statistically equal. Hence, averaging the values, per county and chemical, is reasonable.

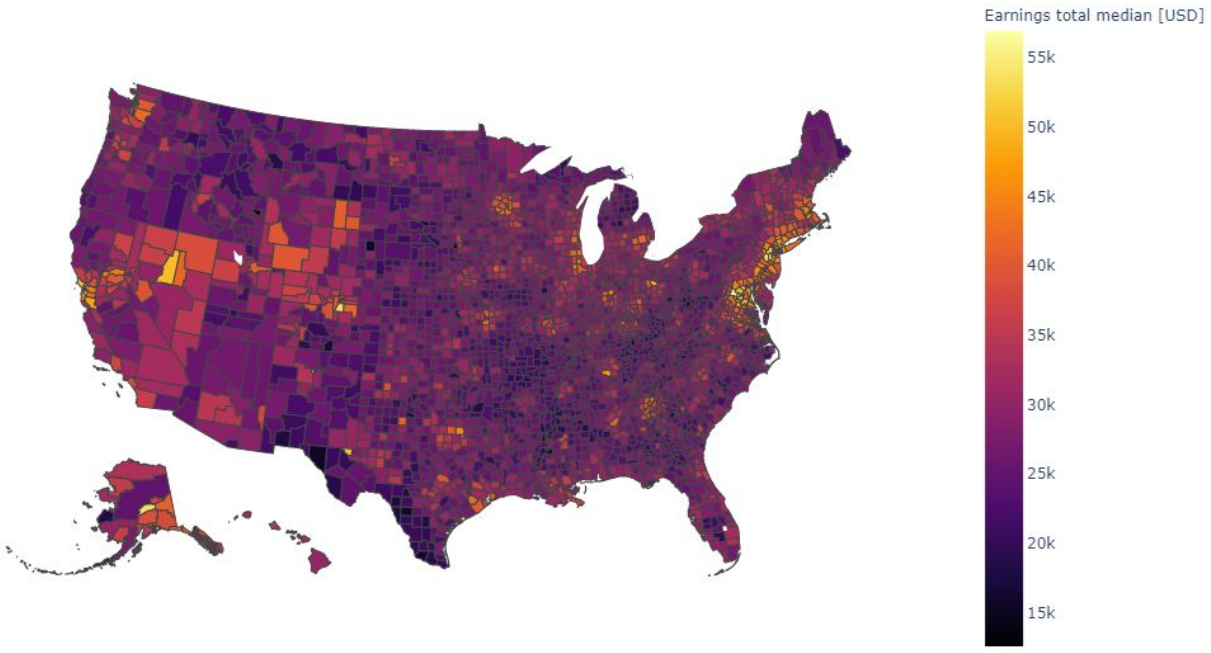
- **Median of Income**

The median of income conveniently is inflation adjusted, so averaging the values per county is reasonable.

2. EXPLORATORY DATA ANALYSIS

The first step in our Exploratory Data Analysis (EDA) was to check patterns according to location among key variables such as total median of earnings. From Figure 1, it is possible to see that the distribution of earnings in the country is very dependent on the location with two identified hotspots of high earnings in the northeast and pacific region.

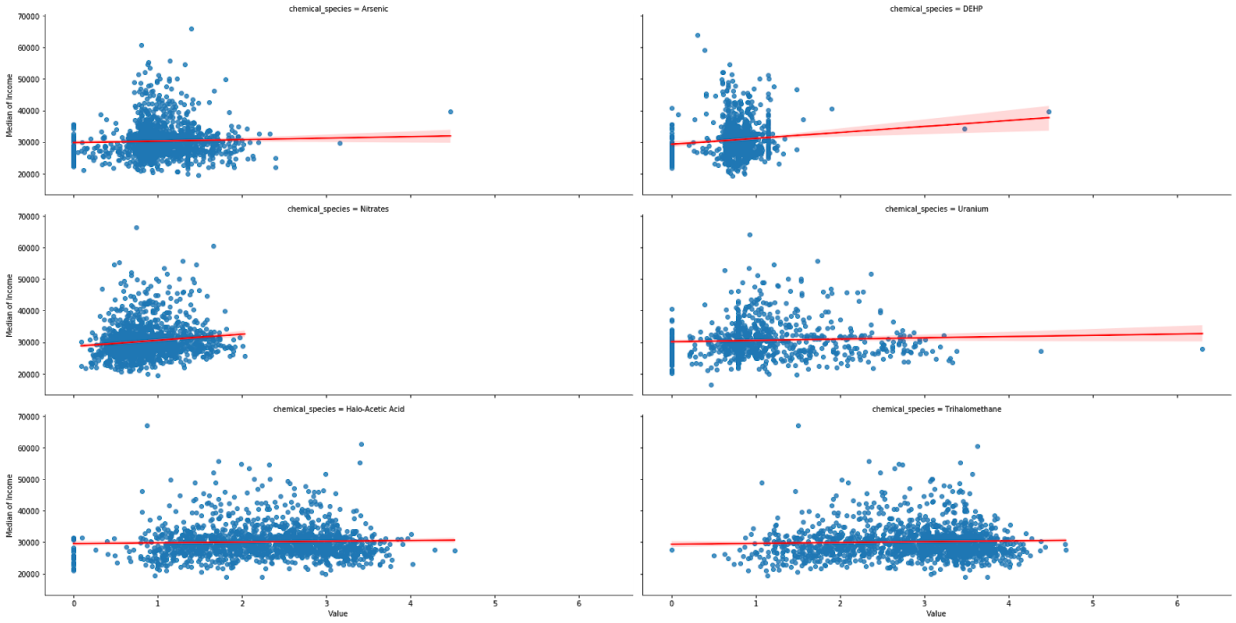
Graph 3. Total median of earnings per county map.



Source: own elaboration.

The next step consisted in relating the filtered data of median earnings with the values of water contaminants. Figure 3 shows scatter plots of median earnings with the values of several contaminants. It seems that high values of pollutants presented in water resources are associated with low values of mean earnings. This is more evident in Arsenic, Haloacetic Acids and Trihalomethane pollutants.

Graph 4. Scatter plot between median earnings and water contaminants/pollutants.



Source: own elaboration.

Since no clear relationship was found in Figure 2, a series of scatter plots by year and by pollutant are plotted with hue according to the contaminant level. A new scale was proposed to divide the original range in greater than MCL, upper bound less than MCL, lower bound less than MCL and non-detected. There is a clear trend of red dots (greater than MCL values) to lie in zones where the median earning is low.

Although the primary focus of the project lies in water quality and money, it seems that money is a more complex variable that is not only dependent on water quality, that is why other variables such as political tendency (e.g. democrat or republican), demographic (population) and geographic (region name) were included to test if they can explain our target variable.

Graph 5. Scatter plot between median earnings and water contaminants by contaminant category.

Source: own elaboration.

Through EDA, the defined final hypothesis was:

- Check if there is a significant impact of the water pollutants in the median earnings of the counties of the United States of America.

Ad hoc hypotheses:

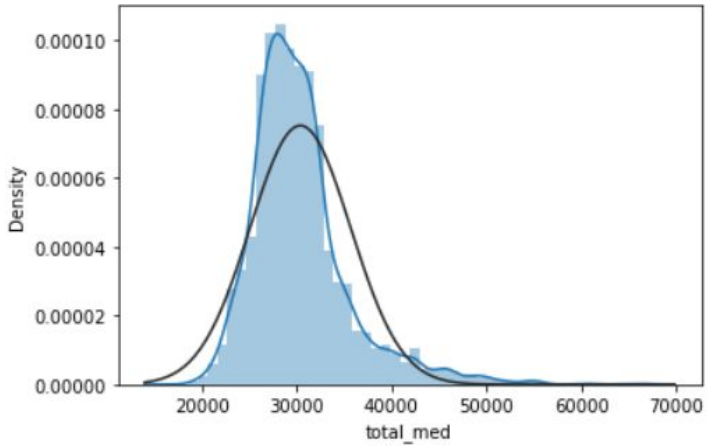
- Has the political tendency of the county a significant impact on the median earnings of the counties of the United States of America.
- Has the region belonging to the county a significant impact on the median earnings of any of the counties of the United States of America.

3. MODELING

Road to Linear Regressions:

First, it's needed to evaluate the normality of our dependent variable. To do so, the distribution of the data for the median of income was plotted, finding it had negative skewness and some kurtosis, as seen in the figure below:

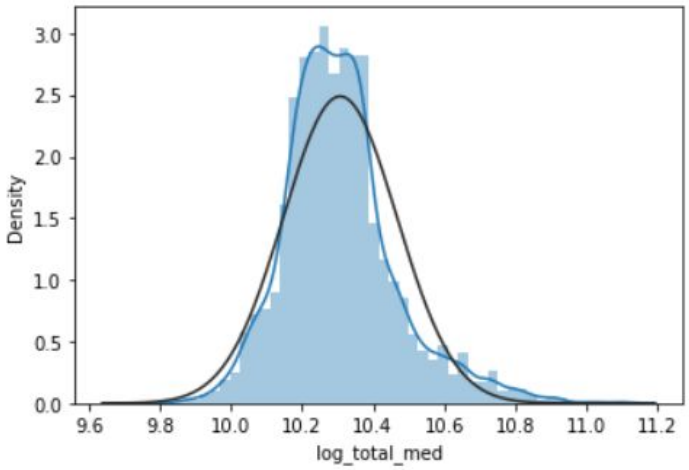
Graph 6. Hist plot of the median income.



Source: own elaboration.

A Box-Cox transformation was performed, in which the lambda value was negative, though a logarithmic transformation was used. Results can be seen in the graph below, there is seemingly still some work to be done, but on comparison, it looks like there is an overall correction.

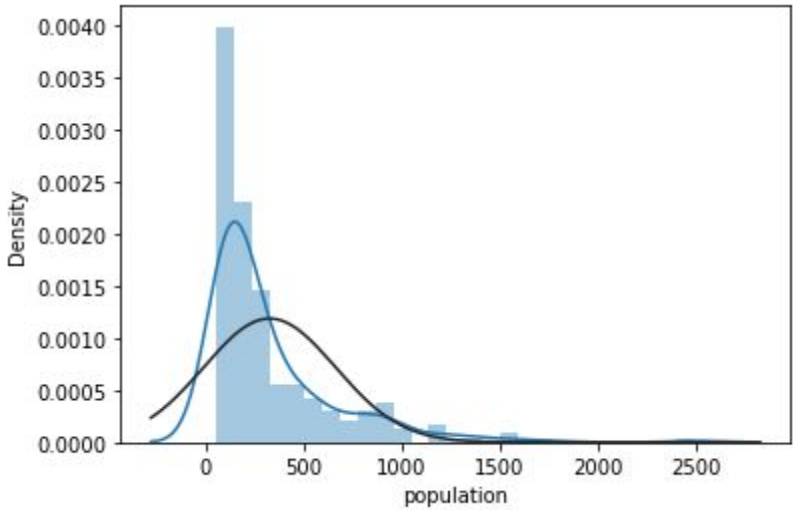
Graph 7. Hist plot of the median income - Box-Cox transformation.



Source: own elaboration.

Another variable worth looking up for its distribution is the population. The resultant distribution that can be seen upon sight in the figure below is a Chi-Squared. However for the purposes of normality required in Linear Regressions, doing a transformation is a must.

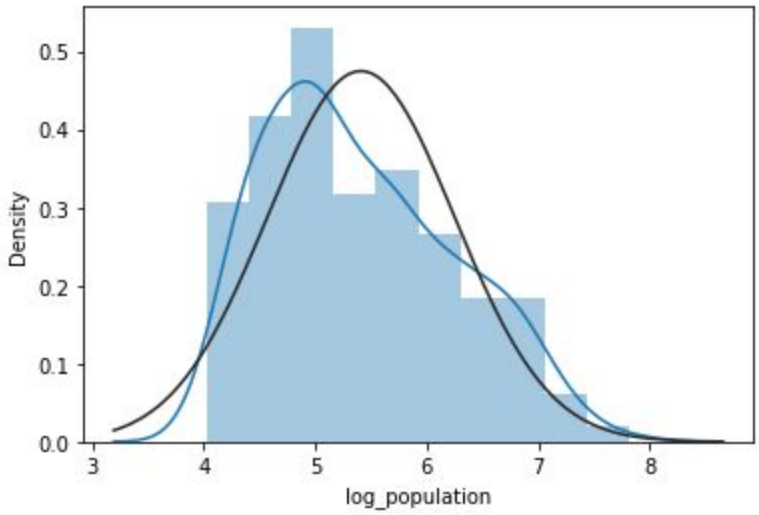
Graph 8. Hist plot of the population.



Source: own elaboration.

Again, a Box-Cox transformation was held with a negative value. Hence, logarithms were required and the results appear below. Still a slight negative skewness can be seen, but the adjustment seems to be enough.

Graph 9. Hist plot of the population - Box-Cox transformation.



Source: own elaboration.

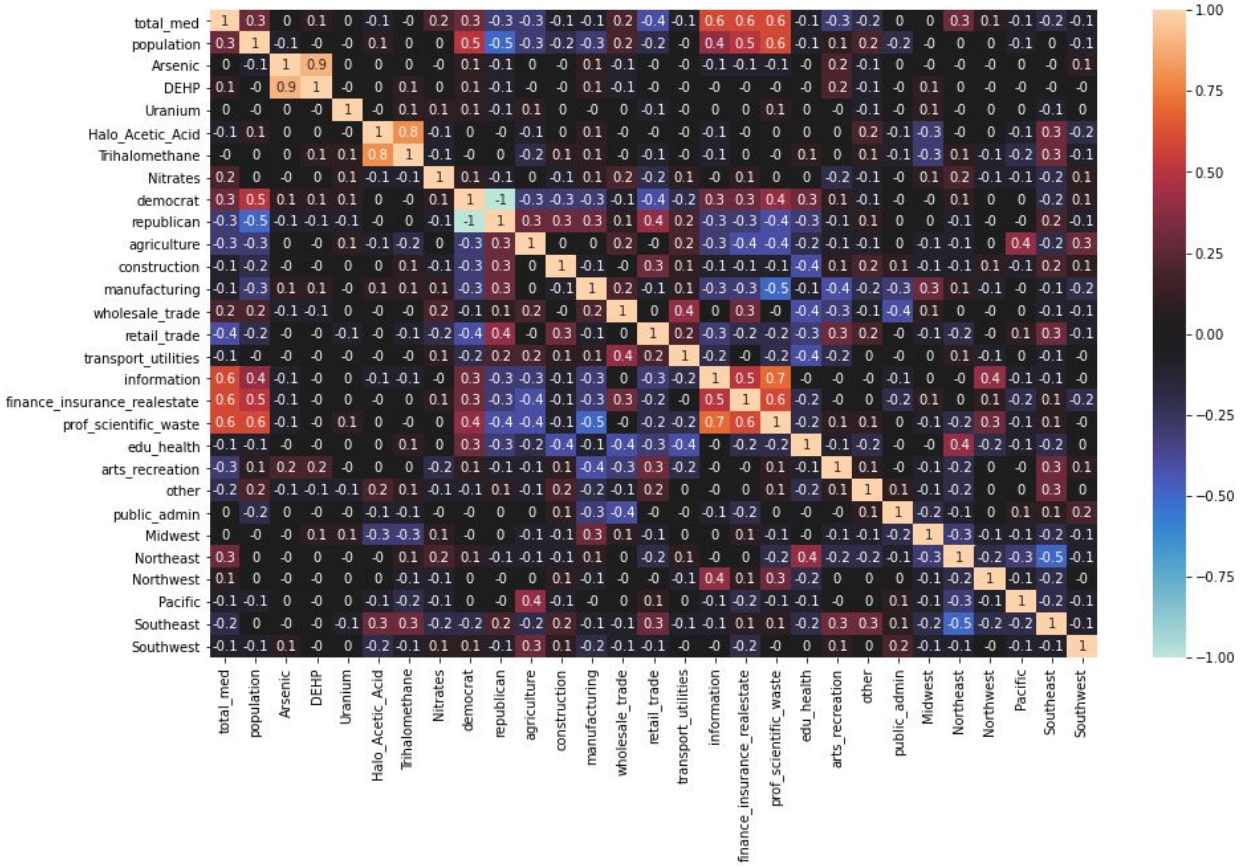
Again, a Box-Cox transformation was held with a negative value. Hence, logarithms were required and the results appear below. Still a slight negative skewness can be seen, but the adjustment seems to be enough.

Generating Dummies

To make a correlation matrix, it's precise to make categorical variables into dummies. Actually, two categorical variables are available:

- Region
    - Midwest
    - Northeast
    - Northwest
    - Pacific
    - Southwest
    - Southeast
- Party
    - Republican
    - Democrat

Graph 10. Correlation matrix - all variables



Source: own elaboration.

Conclusions:

- There's almost no correlation between the median of income of a county and the value of contamination for each chemical.
- There are industries highly correlated with income.
- Population could be important to explain the median income of a County.
- High correlation between Trihalomethane and Halo-Acetic Acid is explained because they happen because of the same chemical released in water sources to clean them from bio-hazards for humans: chlorine.
    - If these variables were significant probably we must leave just one, because of collinearity.
- High correlation between DEHP and Arsenic is not immediately found.
    - If these variables were significant probably we must leave just one, because of collinearity.
- Industries with high correlation with the median of income are also highly correlated among them (Information, financial/insurance/realState, professional/scientific).
- Democrats appear to have some positive correlation with median of income.

Building Linear Regression

The first OLS model includes all variables previously analyzed, where our dependent variable is the median income of the counties. When estimating the model by OLS, it is assumed that the median income performance of county i can be explained based on each of the listed characteristics of that county i.

The proposed econometric model is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_k X_{ik} + \varepsilon i$$

Where k is the n-th independent variable of the model and i is a given county.

Thus:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_{27} X_{i27} + \varepsilon i$$

Where:

$Y_i := Logarithm\ of\ median\ income$
$\beta_0 := Intercept$
$X_{i1} := Dummy - Democrats\ party\ as\ a\ base\ category$
$X_{i2} := Logarithm\ of\ population$
$X_{i3} := Arsenic$
$X_{i4} := DEHP$
$X_{i5} := Uranium$
$X_{i6} := Haloacetic\ Acid$
$X_{i7} := Trihalomethane$
$X_{i8} := Nitrates$
$X_{i9} := Agriculture$
$X_{i10} := Construction$
$X_{i11} := Manufacturing$
$X_{i12} := Wholesale\ trade$
$X_{i13} := Retail\ trade$
$X_{i14} := Transport\ utilities$
$X_{i15} := Information$
$X_{i16} := Finance - Insurance - Real\ Estate$
$X_{i17} := Prof\ scientific\ waste$
$X_{i18} := Edu\ Health$
$X_{i19} := Art\ recreation$

$X_{i20} := Other$
$X_{i21} := Public\ Admin$
$X_{i22} := Northeast\ (Midwest\ as\ a\ base\ category)$
$X_{i23} := Northwest\ (Midwest\ as\ a\ base\ category)$
$X_{i24} := Pacific\ (Midwest\ as\ a\ base\ category)$
$X_{i25} := Southwest\ (Midwest\ as\ a\ base\ category)$
$X_{i26} := Southeast\ (Midwest\ as\ a\ base\ category)$
$X_{i27} := Southwest\ (Midwest\ as\ a\ base\ category)$
$i := County$
$\varepsilon i := Residual$

Given this multiple regression model, an attempt is made to explain the median income of a given county based on the polluting chemicals present in its water resources, the most preferred political party (democrats or republicans) in the county, the percentage of economically active population of each industry over the general total of the county, and the region to which said county belongs.

Full Linear Regression

In this first approximation we include all the variables mentioned so far. Table X shows the results of this first regression. Here we seek to focus on three test statistics; the first, the f-test, where we verify the proof that the model is globally significant; the second, the R squared, which is a measure that shows the percentage of the total of the dependent variable explained by the independent variable; and third, the significance of each of the betas of the independent variables.

F-Test

The F-test for overall significance has the following two hypotheses:

- The null hypothesis states that the model with no independent variables fits the data as well as your model.
- The alternative hypothesis says that your model fits the data better than the intercept-only model.

At this moment we compare the p-value for the F-test with our significance level (alpha=0.05), and we can conclude that our sample data provide sufficient evidence to conclude that our regression model fits the data better than the model with no independent variables.

R-squared

R-squared measures the strength of the relationship between our model (all variables) and the dependent variable (median income). In this case our R-squared is equal to 0.786, which means that the median income is explained in 78.6% for our model.

Hypothesis tests on β coefficients

The T-test has the following two hypotheses:

- The null hypothesis states that the β coefficient is equal to zero.
- The alternative hypothesis says that the β coefficient is different to zero.

At this moment we compare the p-value for the T-test of each one of our variables with our significance level (alpha=0.05). If the p-value is lower than our significance level we can reject the null hypothesis and conclude that we observe sufficient evidence to say that the determined variable of our regression model helps us to predict the median income. If the p-value of a certain variable leads us not to reject the null hypothesis then we can dispense with said variable.

Table 4. OLS Regression Results - all variables

Dep. Variable:	np.log(total_med)	R-squared:	0.786
Model:	OLS	Adj. R-squared:	0.763
Method:	Least Squares	F-statistic:	34.13
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	8.82e-64
Time:	23:33:40	Log-Likelihood:	288.15
No. Observations:	258	AIC:	-524.3
Df Residuals:	232	BIC:	-431.9
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.2018	0.103	99.144	0.000	9.999	10.405
party[T.republican]	-0.0037	0.015	-0.241	0.810	-0.034	0.026
np.log(population)	-0.0312	0.010	-3.137	0.002	-0.051	-0.012
Arsenic	0.0018	0.002	0.789	0.431	-0.003	0.006
DEHP	0.0014	0.002	0.662	0.509	-0.003	0.005
Uranium	-0.0008	0.001	-0.888	0.375	-0.003	0.001
Halo_Acetic_Acid	0.0005	0.001	0.460	0.646	-0.002	0.002
Trihalomethane	-0.0005	0.001	-0.749	0.454	-0.002	0.001
Nitrates	0.0071	0.006	1.109	0.268	-0.005	0.020
agriculture	-0.2829	0.288	-0.981	0.328	-0.851	0.285
construction	1.3657	0.435	3.138	0.002	0.508	2.223
manufacturing	0.3979	0.170	2.339	0.020	0.063	0.733
wholesale_trade	4.4428	0.930	4.778	0.000	2.611	6.275
retail_trade	-1.7546	0.457	-3.842	0.000	-2.655	-0.855
transport_utilities	-0.6273	0.489	-1.282	0.201	-1.591	0.337
information	2.8344	1.103	2.569	0.011	0.661	5.008
finance_insurance_realestate	2.7095	0.450	6.020	0.000	1.823	3.596
prof_scientific_waste	2.9611	0.365	8.108	0.000	2.242	3.681
edu_health	-0.0667	0.177	-0.376	0.707	-0.416	0.283
arts_recreation	-0.9575	0.267	-3.589	0.000	-1.483	-0.432
other	-2.1475	0.926	-2.320	0.021	-3.971	-0.324
public_admin	1.3271	0.243	5.452	0.000	0.848	1.807
Northeast	0.0852	0.021	4.089	0.000	0.044	0.126
Northwest	-0.0785	0.031	-2.524	0.012	-0.140	-0.017
Pacific	-0.0060	0.029	-0.209	0.835	-0.063	0.051
Southeast	-0.0482	0.024	-1.999	0.047	-0.096	-0.001
Southwest	-0.0765	0.043	-1.782	0.076	-0.161	0.008

Source: own elaboration.

- The model is globally significant.
- The model explains the 76% of the errors.
- There are several variables that are non significant that we must extract from our model.

Reducing the Linear Regression

Once we have debugged each of the independent variables (one at a time) whose p-value leads us not to reject the null hypothesis, we can conclude that this given model is the final model.

The variables debugged were:

- party
- edu_health
- Halo_Acetic_Acid
- Trihalomethane
- DEHP
- agriculture
- Uranium
- Nitrates
- transport_utilities
- Other
- Arsenic

Arsenic was ripped off last, because the coefficient has an impact of exp(0.003) or 1.0 USD in the median of income in a County, per each microgram/L of Arsenic found in water sources.

Because the max. amount in our dataset is 89.7 and the minimum is 0.0; means the maximum amount of USD this variable contributes is USD 90.0.

Other indications that also led us to debug the arsenic variable:

* Correlation between the variables of contaminants in Water per chemical was very close to zero.
* There's a tiny impact this variable has in our dependent variable.
* The value of contamination of the rest of the Chemicals was none.

Final Linear Regression

Evaluating the model whose betas are significant we have:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{17} X_{i17} + \varepsilon i$$

Where:

$$Y_i := Logarithm\ of\ median\ income$$

$\beta_0 := Intercept$
$X_{i1} := Logarithm\ of\ population$
$X_{i2} := Construction$
$X_{i3} := Manufacturing$
$X_{i4} := Wholesale\ trade$
$X_{i5} := Retail\ trade$
$X_{i6} := Information$
$X_{i7} := Finance - Insurance - Real\ Estate$
$X_{i8} := Prof\ scientific\ waste$
$X_{i9} := Art\ recreation$
$X_{i10} := Public\ Admin$
$X_{i11} := Northeast\ (Midwest\ as\ a\ base\ category)$
$X_{i12} := Northwest\ (Midwest\ as\ a\ base\ category)$
$X_{i13} := Pacific\ (Midwest\ as\ a\ base\ category)$
$X_{i14} := Southwest\ (Midwest\ as\ a\ base\ category)$
$X_{i15} := Southeast\ (Midwest\ as\ a\ base\ category)$
$i := County$
$\varepsilon i := Residual$

In the final linear regression model we have the following variables:

Table 4. OLS Regression Results - final variables

Dep. Variable:	np.log(total_med)	R-squared:	0.765
Model:	OLS	Adj. R-squared:	0.751
Method:	Least Squares	F-statistic:	52.62
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	2.67e-67
Time:	23:33:40	Log-Likelihood:	276.12
No. Observations:	258	AIC:	-520.2
Df Residuals:	242	BIC:	-463.4
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.9581	0.105	95.126	0.000	9.752	10.164
np.log(population)	-0.0309	0.009	-3.269	0.001	-0.050	-0.012
construction	1.2918	0.427	3.024	0.003	0.450	2.133
manufacturing	0.6808	0.179	3.798	0.000	0.328	1.034
wholesale_trade	4.1187	0.884	4.659	0.000	2.377	5.860
retail_trade	-1.8644	0.436	-4.277	0.000	-2.723	-1.006
information	3.2389	1.191	2.720	0.007	0.894	5.584
finance_insurance_realestate	2.8742	0.434	6.625	0.000	2.020	3.729
prof_scientific_waste	3.1512	0.343	9.196	0.000	2.476	3.826
arts_recreation	-0.5491	0.290	-1.894	0.059	-1.120	0.022
public_admin	1.5325	0.279	5.495	0.000	0.983	2.082
Northeast	0.0896	0.019	4.638	0.000	0.052	0.128
Northwest	-0.0850	0.030	-2.866	0.005	-0.143	-0.027
Pacific	-0.0099	0.026	-0.388	0.699	-0.060	0.041
Southeast	-0.0576	0.021	-2.740	0.007	-0.099	-0.016
Southwest	-0.0754	0.041	-1.848	0.066	-0.156	0.005

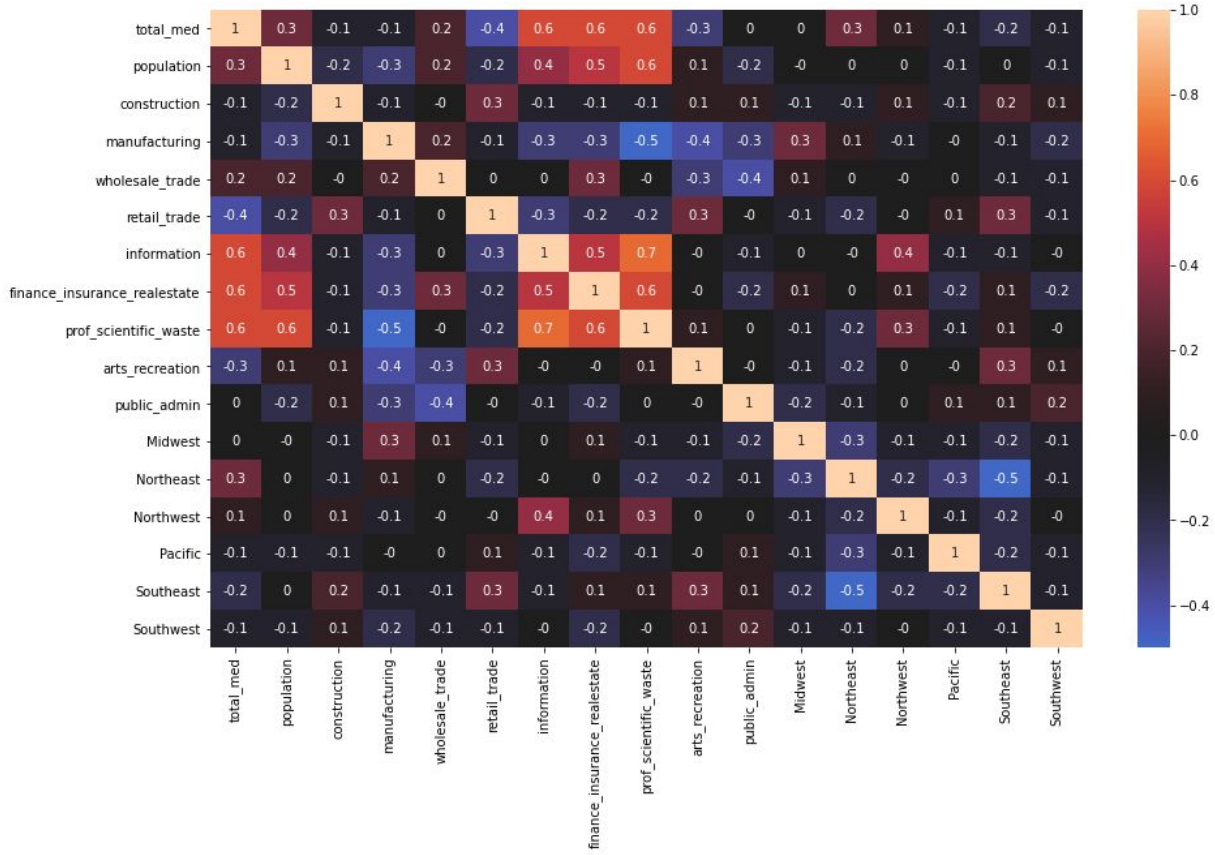
Source: own elaboration.

The reduced model is chosen instead of the the full one, because:

- The AIC is lower than the previous model.
- The adjusted R2 is almost the same.
- We save ourselves the use/measure of several variables.

We present again the correlation matrix of the end variables of our model.

Graph 10. Correlation matrix - all variables



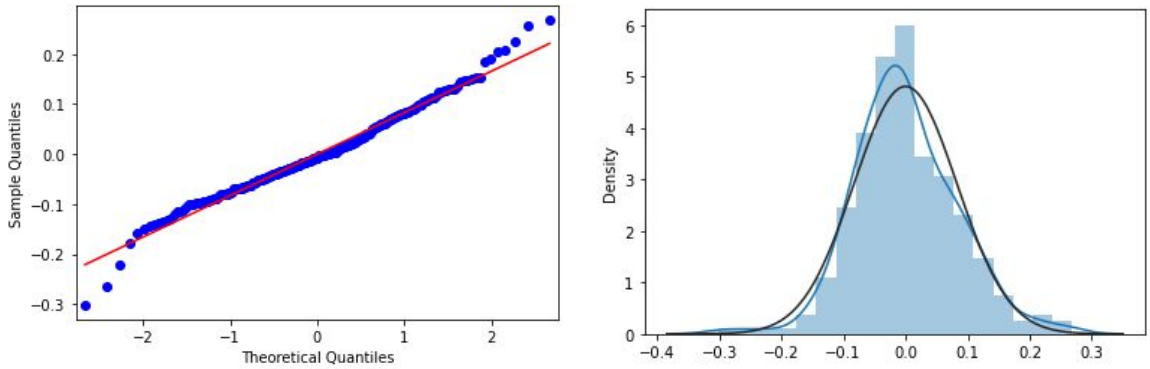
Source: own elaboration.

Conclusions:

- There are industries highly correlated with income.
- Population could be important to explain the median income of a County.
- Industries with high correlation with the median of income are also highly correlated among them (Information, financial/insurance/realState, professional/scientific).

Now the residuals are presented. Though there seems to be some atypic values at both ends, in general the errors seem to follow normality:

Graph 11. QQ Plot and histogram of residuals.



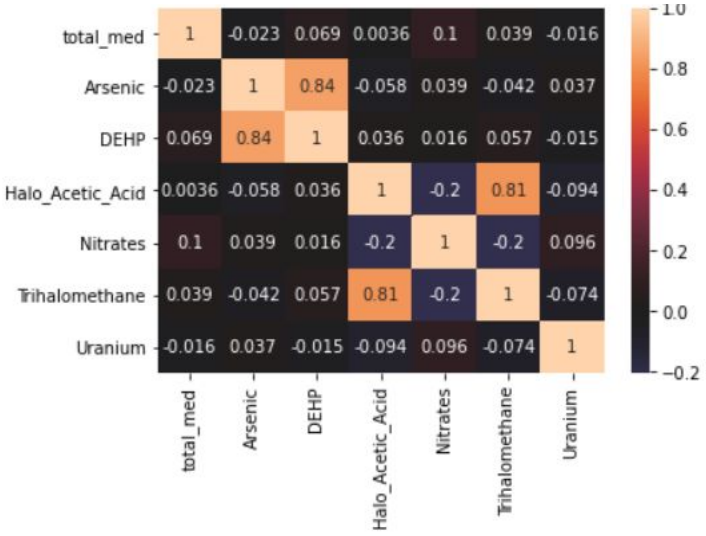
Source: own elaboration.

Mini-case: Impact of chemical contamination explaining median of income

This section seeks to evaluate one of the initial hypotheses, which was that the income of the counties depended directly on the pollution caused by their industries to their water networks.

Correlation Matrix among the chemicals and the dependent variable.

Graph 12. Correlation matrix - mini case



Source: own elaboration.

Linear Regression

Table 6. OLS Regression Results - mini case

Dep. Variable:	np.log(total_med)	R-squared:	0.051
Model:	OLS	Adj. R-squared:	0.042
Method:	Least Squares	F-statistic:	5.486
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	1.52e-05
Time:	17:12:17	Log-Likelihood:	238.60
No. Observations:	619	AIC:	-463.2
Df Residuals:	612	BIC:	-432.2
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.3005	0.017	596.655	0.000	10.267	10.334
Arsenic	-0.0137	0.003	-4.162	0.000	-0.020	-0.007
DEHP	0.0140	0.003	4.408	0.000	0.008	0.020
Halo_Acetic_Acid	-0.0011	0.001	-1.176	0.240	-0.003	0.001
Nitrates	0.0224	0.008	2.981	0.003	0.008	0.037
Trihalomethane	0.0012	0.001	1.694	0.091	-0.000	0.003
Uranium	-0.0002	0.001	-0.303	0.762	-0.001	0.001

Omnibus:	81.250	Durbin-Watson:	1.500
Prob(Omnibus):	0.000	Jarque-Bera (JB):	137.030
Skew:	0.828	Prob(JB):	1.76e-30
Kurtosis:	4.603	Cond. No.	103.

Source: own elaboration.

- Though the model is globally significant, it can explain just 5% of the dependent variable.
- We see there are chemicals that aren't significant for the model.
- The impact of the coefficients of the significant chemicals is about 1 USD.
- The max. value for Nitrates is 5.8, for DEHP 89.9 and Arsenic 89.7; so their impacts are USD 6, USD 90 and USD 90, respectively.
- We can conclude there's little relation between income and contamination in water sources of the 6 chemicals in the study; while being the only explicative variables.

Conclusions of the mini-case:

In this mini-case we include all the contaminant chemicals within the water of each county listed.

First, the f-test, shows us that the model is globally significant; the second, the R squared, who has a value of 0.051, says us that the median income is explained in 5.1% for the contaminants chemicals in the water (a very poor performance); and finally, the significance of half of the contamientates leads us not to reject the null hypothesis. Therefore, this model for predicting the median income with water polluting chemicals does not provide us with strong statements.

END CONCLUSIONS OF THE CASE:

The initial hypothesis of the case was to verify if the mean income of a county had any relationship with the contamination of its water sources. This assumption was based on the fact that the industrialization of a certain county went hand in hand with the contamination of its water.

Thus, the initial purpose was to evaluate whether arsenic contamination of the water had any impact (positive or negative) on the median income of a county.

The statistical evidence shows us that the contamination of the water resources of the counties does not have a relationship (at least not linear) with the median of their income, even with models where more independent variables were included.

The more traditional approaches to predict the median of income seem to be more suited for this task. In the study it was uses the following variables:

* Population
* Industries
* Region