



Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

US Water Scarcity & Sanitation

Team 20

Bowen Chen
Jeremy Watkins
Zixin Wang
Sophia Zhou

Background Information

- Human life is impossible without water, covering roughly 70% of the Earth's surface and accounting for 60% of an adult's body by weight. Clean drinking water is necessary for survival, but access to fresh water is also essential for irrigation of crops, basic hygiene, and medical care.
- Continuous efforts have been made to increase the percentage of global population that has access to better quality water but hundreds of millions still live without access to clean water.

Questions to Solve

Public Health

- Key drivers of poor health
- Correlation among different pollutants

Water Scarcity

- How to define?
- Human Activity Correlation
- How to Capture Nonlinearity?

Water Sanitation

- Industry correlation?
- Education level relevance?
- Geographical distribution?

Data Sets

- Chemicals: mean concentration of a particular chemical
- Droughts: particular percentage of various range of drought severities
- Earnings: industry specific median earnings
- Education_Attainment: educational attainment of the US population
- Industry_Occupation: estimated working population for the various industries
- Water_Usage: information about particular water usage
- County_Health_Rankings: population percentage of good/poor health



Public Health

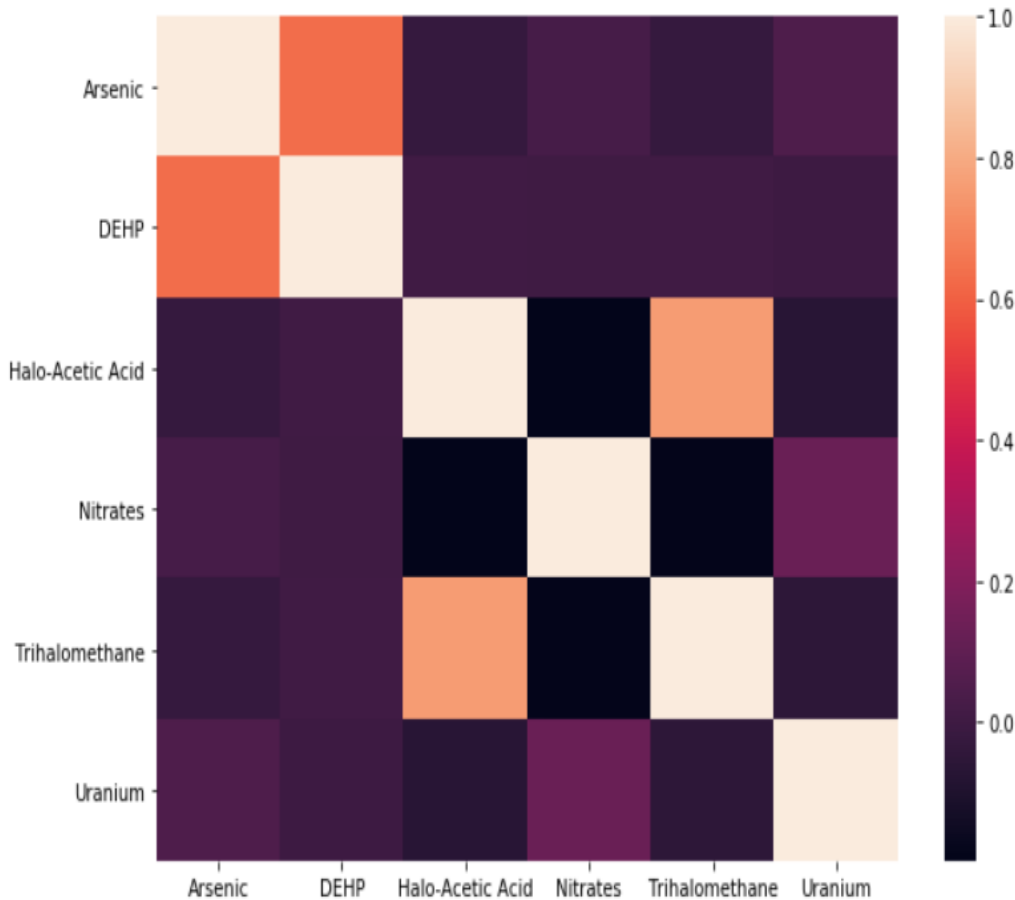


Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

Main Findings

- We looked at six major types of chemical contaminant or particulate that can lead to potential severe health problems: Uranium, Arsenic, DEHP, Nitrates, Halo-Acetic Acid, Trihalomethane
- Most Dangerous pollutants: Nitrates and Trihalomethane because they make significant contribution to poor health condition
- Arsenic and DEHP separately are not harm substance, however, putting them together can cause significant health problems. Correlation matrix shows that they tend to appear at nearby locations.
- Can potentially use the model to predict future disease distribution.

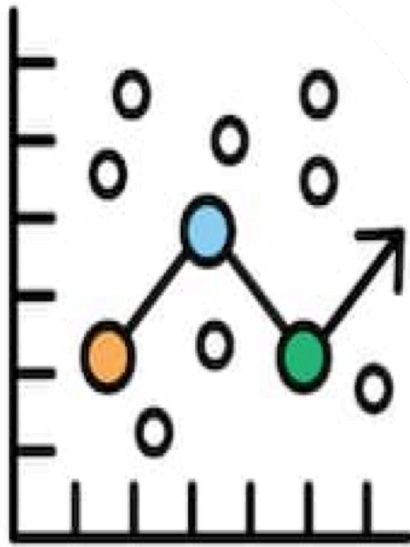
Correlation among Pollutants



Correlation between pollutants

- Statistically, with very little correlation between these chemicals, it will allow us to build a robust regression model
- Arsenic, DEHP separately won't cause big issues, however, combination of those two can lead to severe health issue. High correlation appears in the correlation matrix (0.76), which means they tend to appear at nearby locations.

Does Those Chemicals REALLY Affect People's Health?



Regression Analysis

- Linear regression analysis
- Interpretable, though no casual effects
- Uncover the relationship between pollutants and %population with poor health
- Discover significant, individual relationships between populations poor health and each pollutants

Regression Model - Formulation

Features

Arsenic
DEHP
Halo-Acetic Acid
Nitrates
Trihalomethane
Uranium'

The 6 chemicals

TheData  pen

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Responses

% Fair/ Poor Health

*Total population that has
poor health*

The Most Significant Chemicals

	coef	std err	t	P> t	[0.025	0.975]
Population	4.327e-06	4.7e-07	9.202	0.000	3.41e-06	5.25e-06
Arsenic	0.2475	0.028	8.691	0.000	0.192	0.303
DEHP	-0.1186	0.021	-5.558	0.000	-0.160	-0.077
Halo-Acetic Acid	0.1076	0.011	9.550	0.000	0.086	0.130
Nitrates	2.3981	0.081	29.727	0.000	2.240	2.556
Trihalomethane	0.2906	0.007	38.925	0.000	0.276	0.305
Uranium	0.1301	0.026	5.015	0.000	0.079	0.181

- Nitrates and Trihalomethane are the two most dangerous pollutants in terms of their high risk in causing health problems, such as rectal, bladder and breast cancers.
- Largest positive coefficient – every 1% increase in nitrate concentration leads to 2% of population increase in poor health condition.
- Statistically significant – close to truth.
- Can potentially use this regression model to predict future disease distribution.

Water Scarcity



Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

Main Findings

- Low level droughts may be driven by water consumptions (easy to predict) while severe droughts mostly driven by climate (hard to model).
- Although domestic consumption is low compared with total usage, it reflects people's sense of water saving
- Random Forest did remarkably better than traditional linear/logistic regressions

How Do We Define?

- Use the drought.csv data and compute each county's yearly average drought population percentage for different levels
- The more population percentage involved in high drought levels, the more severe the water scarcity does the county face

Response Variables	Description
Drought Ratio	Yearly Average Percentage of Drought Population
D0 Ratio	Yearly Average Percentage of D0 Drought Population
D1 Ratio	Yearly Average Percentage of D1 Drought Population
D2 Ratio	Yearly Average Percentage of D2 Drought Population
D3 Ratio	Yearly Average Percentage of D3 Drought Population
D4 Ratio	Yearly Average Percentage of D4 Drought Population

Unmask the Link Between Droughts and Human Behavior Patterns

- We designed several county-level indicators to model the behavior

Indicators	Description
Ground Water Ratio	Total Ground Water Withdrawal/Total Water Withdrawal
Fresh Water Ratio	Total Fresh Water Withdrawal/Total Water Withdrawal
Industry Water Ratio	Total Industry Withdrawal/Total Water Withdrawal
Irrigation Water Ratio	Total Irrigation Withdrawal/Total Water Withdrawal
Livestock & Aqua Ratio	Total Livestock & Aqua Withdrawal/Total Water Withdrawal
Mining & Electric Ratio	Total Mining & Electric Withdrawal/Total Water Withdrawal
Domestic Use Per Capita	(Domestic Water Supply + Public Deliver)/Population
Median Earnings	Median Earnings of County
Education Level	Percentage of Population With College Degree or Higher



Individual Tests of Indicators

- Response variables are county-level yearly drought
- We firstly use cross-sectional linear regression framework to test the individual predictabilities of indicators
- Most of indicators showed strong t-stats, but poor R squares
- Some indicators showed non-linear relationships with response variables, which demands machine learning algorithms to capture nonlinearity

t-stats Results

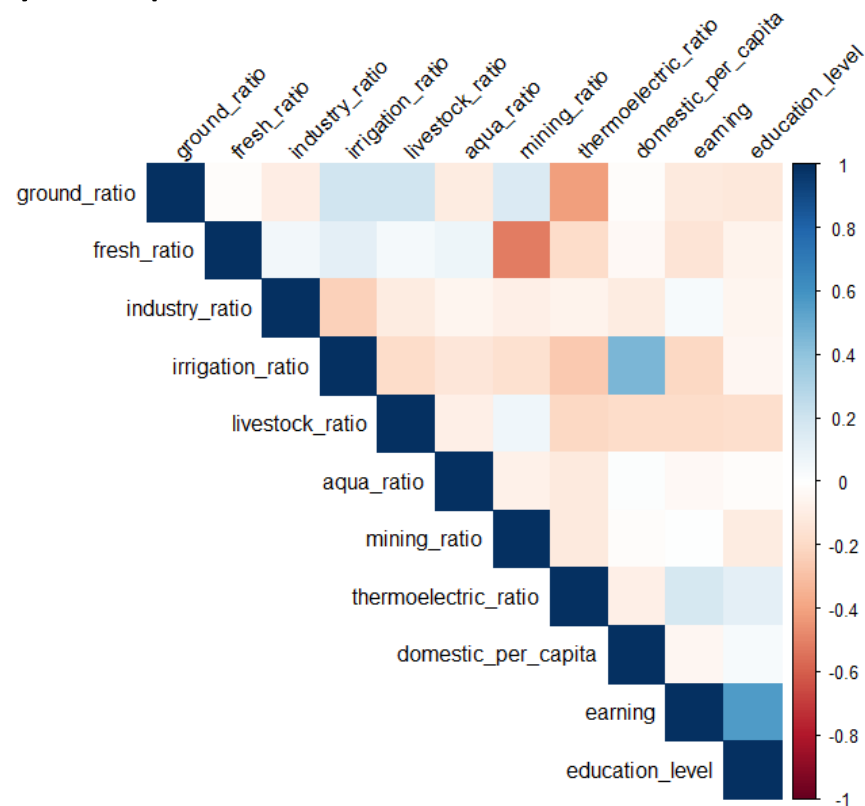
Indicators	Drought Ratio	D0	D1	D2	D3 & D4
Ground Water Ratio	-6.10	-10.99	-3.39	4.26	4.65
Fresh Water Ratio	2.25	0.81	1.36	2.97	4.35
Industry Water Ratio	3.31	1.50	2.22	2.71	3.58
Irrigation Water Ratio	6.21	5.01	4.76	2.24	1.39
Livestock & Aqua Ratio	-17.89	-16.57	-10.36	-6.10	-3.75
Mining & Electric Ratio	-4.06	-2.93	-3.06	-1.78	-1.91
Domestic Use Per Capita	15.46	18.14	9.30	-0.55	-0.60
Median Earnings	-1.38	2.06	-2.41	-3.60	-3.38
Education Level	-4.98	-0.48	-4.64	-6.17	-4.69

R Square Results

Indicators	Drought Ratio
Ground Water Ratio	1.3%
Fresh Water Ratio	0.13%
Industry Water Ratio	0.32%
Irrigation Water Ratio	1.2%
Livestock & Aqua Ratio	9.2%
Mining & Electric Ratio	0.49%
Domestic Use Per Capita	7.3%
Median Earnings	0.28%
Education Level	0.74%

Indicator Correlation Matrix

- We can see relatively low correlations between our indicators, which prevents the collinearity of inputs



Some Interesting Features

- Indicators have different predictabilities to different level of drought
 - Domestic Use Per Capita is strong when predicting low level droughts, but useless when predicting severe droughts. **It indicates that low level droughts may be driven by water consumptions while severe droughts mostly driven by climate.**
 - Ground Water Ratio changes sign of slope when predicting different levels of droughts, which indicates different water consumption patterns.
- Nonlinearity & heteroskedasticity for most indicators

Random Forest Classification

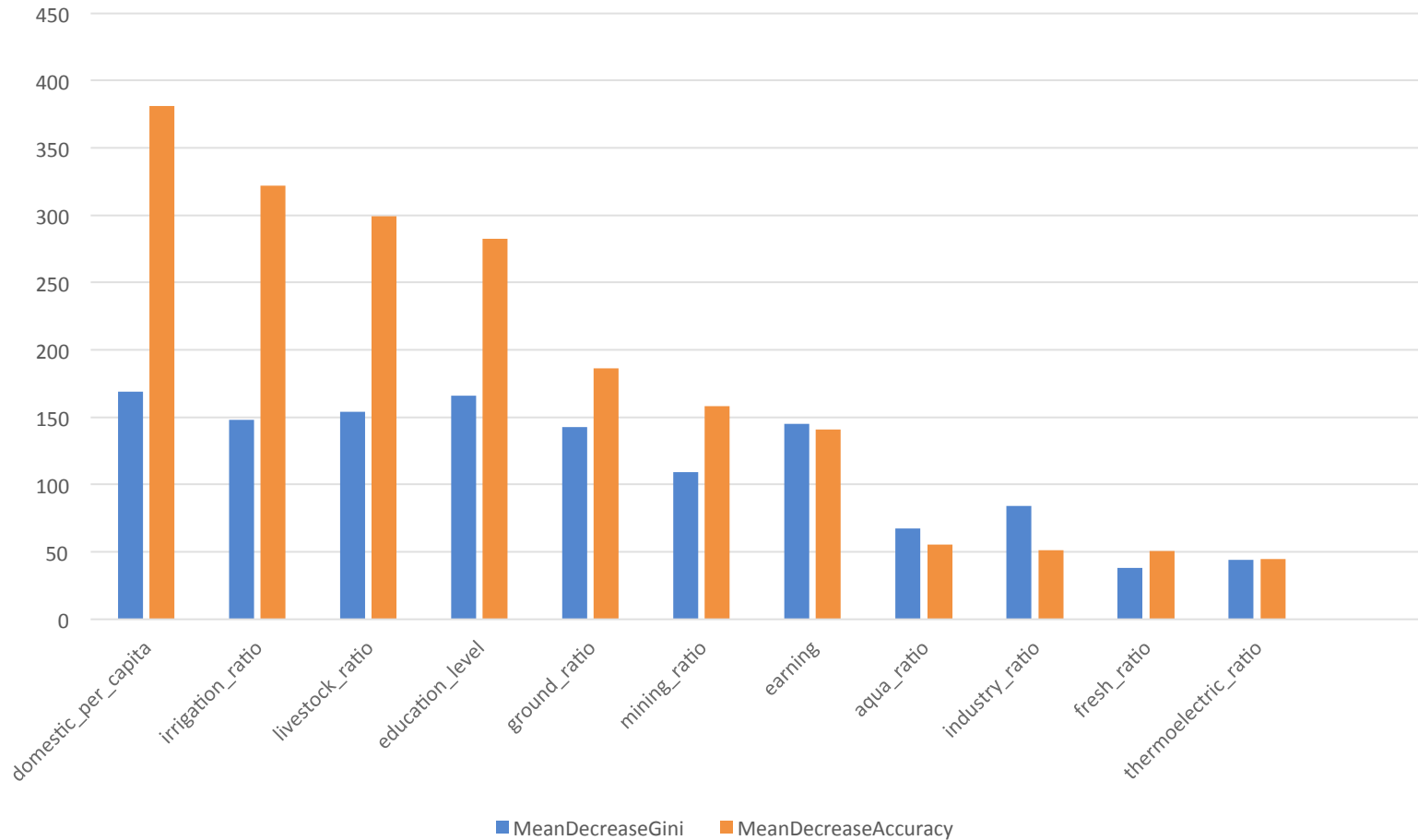
- Since the data showed strong nonlinear relationships, we use Random Forest to classify the drought level
- We classify response variable into 2 factors:
 - Moderate droughts or below: if the county doesn't have droughts greater than D1 level
 - Severe droughts: if the county has droughts greater than D1 level
- For each decision tree, choose 4 splits and aggregate 500 decision trees. Randomly choose 4 indicators as each tree's input
- Split 1/5 of data as testing data, remaining others to train the model

Random Forest Result

- The over-all test error converges to 20% as more trees generated



Indicators' Importance Rank



What does the RF say?

- Much lower error rate compared with linear/logistic regression
- Similar importance rank with linear regression's R square rank
- Robust error convergence

Water Sanitation

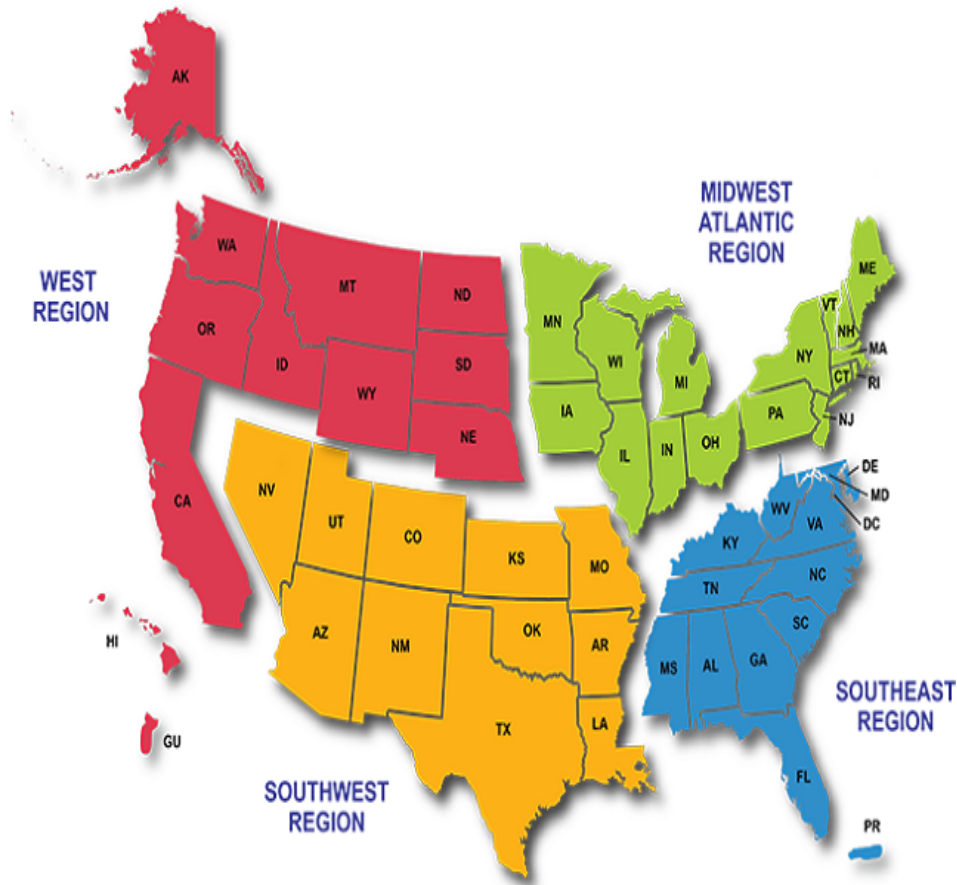


Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

Main Findings

- Less wealthy states, such as Alabama, Kentucky and West Virginia have higher water pollution caused health problems.
- Certain industries such as agriculture, retail and arts/recreation make significant contribution to water pollution.
- Water pollutions are highly related to educational level. Areas concentrated with less educated people tend to have higher water pollution level.

Location Effect – The location health issues

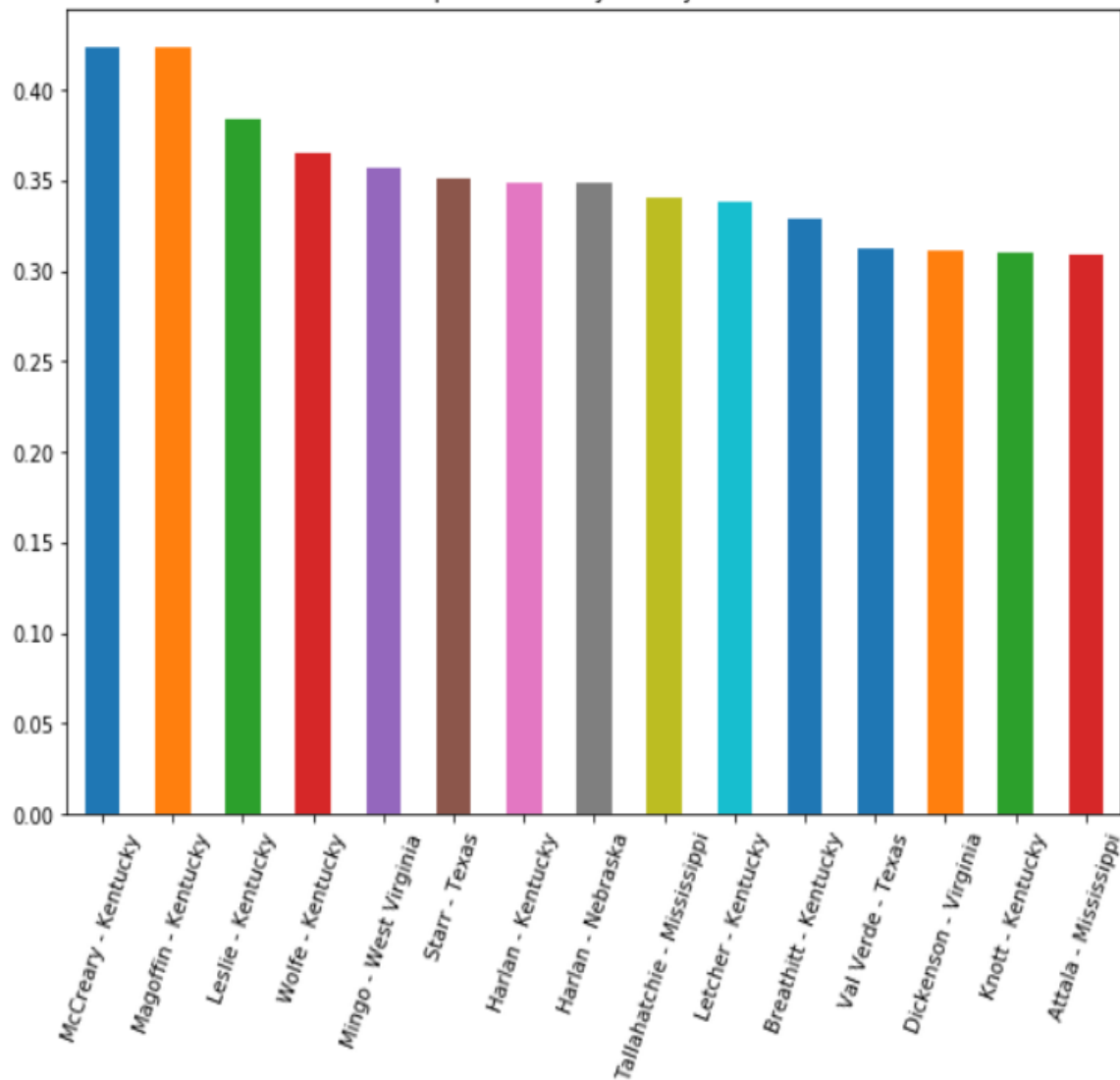


- We built several visualizations to investigate if **the location of the population is an important factor** of population's health.
- We ranked 15 states by their county population health conditions.
- We looked at time trend of state conditions.

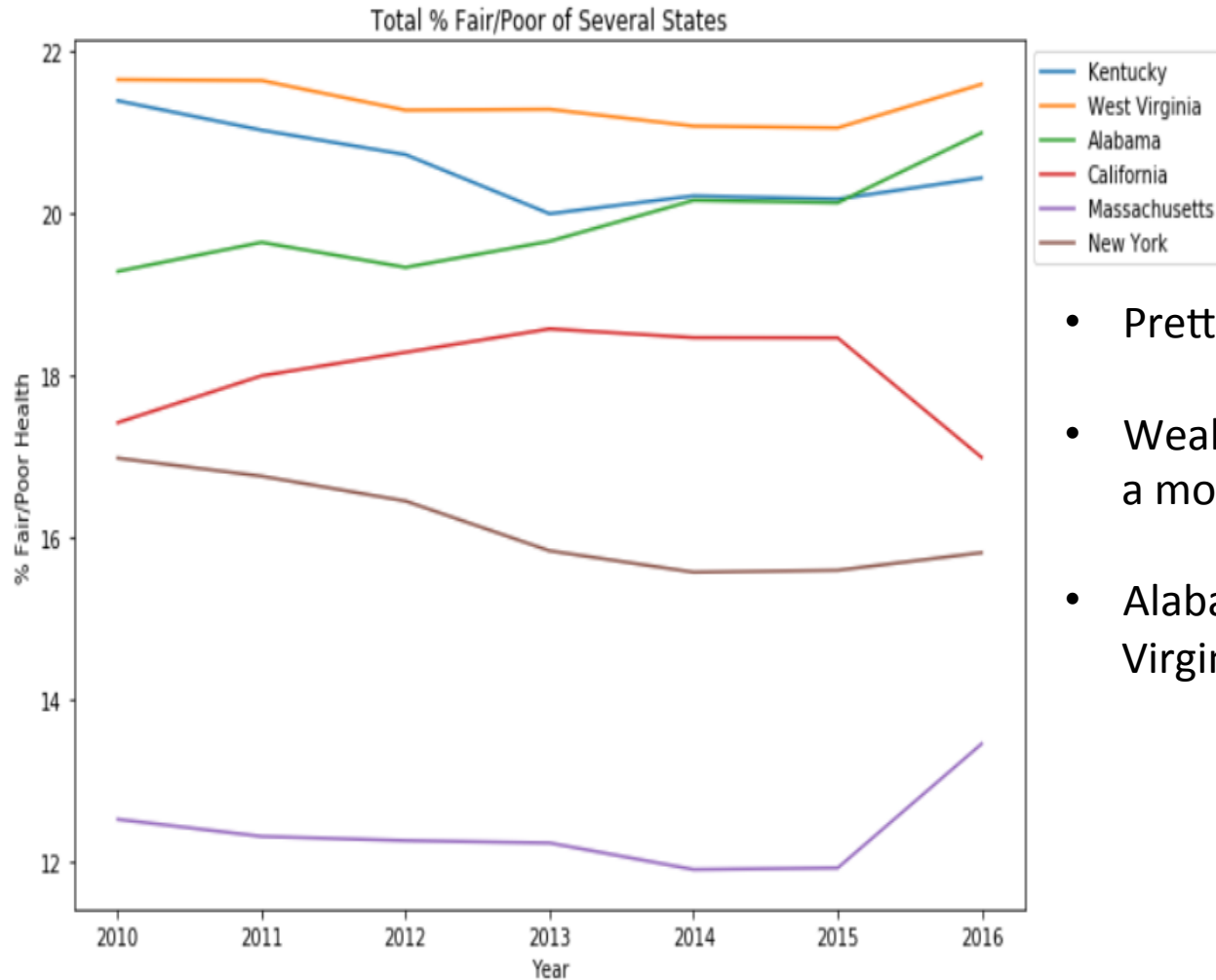
Unhealthy % of Population Per County, Kentucky is Standing Out




Top 15 Unhealthy County, States



Historical Health Trends for Several States



- Pretty stable over time
- Wealthy states seems to have a more healthy population
- Alabama, Kentucky and West Virginia concerning trends



Uncover the Factors of State Specific Health

- To uncover the state level effects, we build an enhanced version of regression that we previously did
- Adding states as dummy variables

State Level Effects – The Medical Conditions

	coef	std err	t	P> t	[0.025	0.975]
Population	-7.532e-07	2.23e-07	-3.380	0.001	-1.19e-06	-3.16e-07
Arsenic	-0.0114	0.013	-0.872	0.383	-0.037	0.014
DEHP	0.0057	0.010	0.594	0.552	-0.013	0.024
Halo-Acetic Acid	0.0093	0.005	1.748	0.080	-0.001	0.020
Nitrates	-0.2957	0.049	-6.084	0.000	-0.391	-0.200
Trihalomethane	0.0020	0.004	0.492	0.623	-0.006	0.010
Uranium	0.0053	0.012	0.444	0.657	-0.018	0.029
State_California	18.0043	0.300	60.025	0.000	17.416	18.592
State_Kentucky	24.2751	0.227	107.124	0.000	23.831	24.719
State_West Virginia	22.1569	0.271	81.781	0.000	21.626	22.688
State_New York	13.9030	0.266	52.310	0.000	13.382	14.424

- The table was really long, so we only show a fraction of them
- The t-test shows strong significances on state level effects on population health
- Wealthy states tend to have 6% less unhealthy populations than ordinary states
- The regression coefficients are reflections of the local medical conditions for each state
- Using this regression model, we could effectively allocate medical forces to the high health rate states, such as Kentucky and West Virginia

Water Sanitation Industry Correlation

- We linearly regressed six pollutants station-population-weighted concentration on all industry populations: Agriculture, Construction, Manufacturing, Transportation/Utilities, Finance, Arts/Recreation
- Two most important/dangerous chemical species:
 - Nitrates:
 - Agriculture, Retail and Art/Recreation make significant contribution
 - Significant increase in the use of both inorganic nitrogen and phosphorous fertilisers led to excessive amount of nitrates in waters.
 - Painting materials contain corrosion inhibitors that have high concentration of nitrates.

	Agriculture	Retail	Art/Recreation
t stats	3.50	-2.36	-2.4

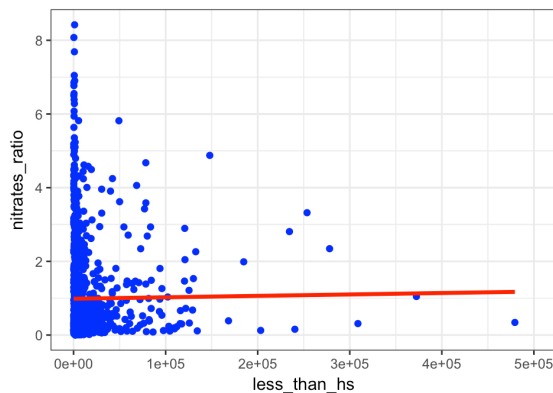
- Trihalomethane:
 - Agriculture, Manufacturing and Art/Recreation make significant contribution

	Agriculture	Manufacturing	Art/Recreation
t stats	-1.83	1.65	2.12

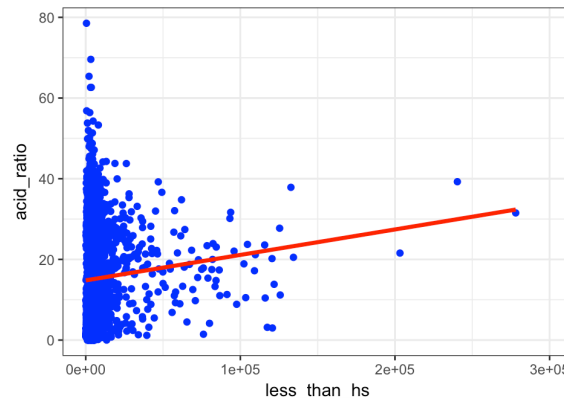
Water Sanitation

Education Level Correlation

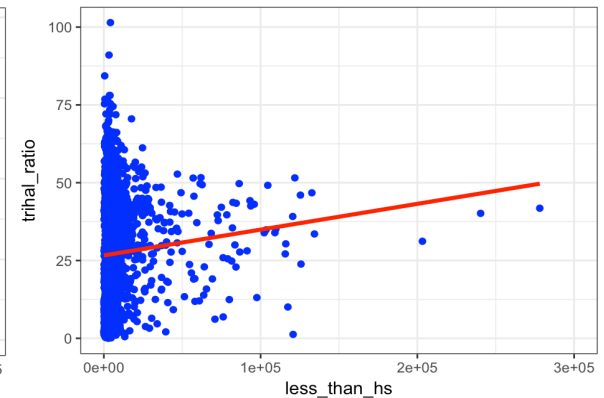
- We linearly regressed six pollutants station-population-weighted concentration three different educational levels: less than high school degree, some college or associate degree, and bachelors or higher degree.
- We found areas with higher density of people with “less than high school degree” are more polluted, especially with substances such as nitrates, acid, and trihalomethane.



Nitrates, $t=2.04$



Acid, $t=4.10$



Trihalomethane, $t=3.13$

Conclusions

- Water scarcity and water sanitation have significant impact on human health conditions.
- Lower level of droughts are driven by human activities, while severe droughts are mostly driven by climates and other effects.
- Nonlinear machine learning models has significantly better performance then linear/logic regression in predicting water scarcity.
- Water sanitation issue is more severe in less wealthy states, such as Alabama, Kentucky and West Virginia.
- Industries such as agriculture, arts/recreation make significant contribution to the water pollution problem especially caused by nitrates.
- More educated people tend to make less water pollution.
- We should closely regulate the industries mentioned above, and educate people about the importance of water sanitation.

Appendix- A Machine Learning Approach to Predict Droughts



Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

Supervised Learning Model

- Binary Classification of drought severity for D0, 4-Weeks in the future
- Model:
 - Logistic Regressor using XGBoost
- Performance Metrics:
 - ROC Curve AUC
 - Precision Recall Curve AUC
- Data
 - Droughts.csv from 2011 to 2015
 - Droughts.csv was merged with 2010 Waterusage.csv

Model Fitting

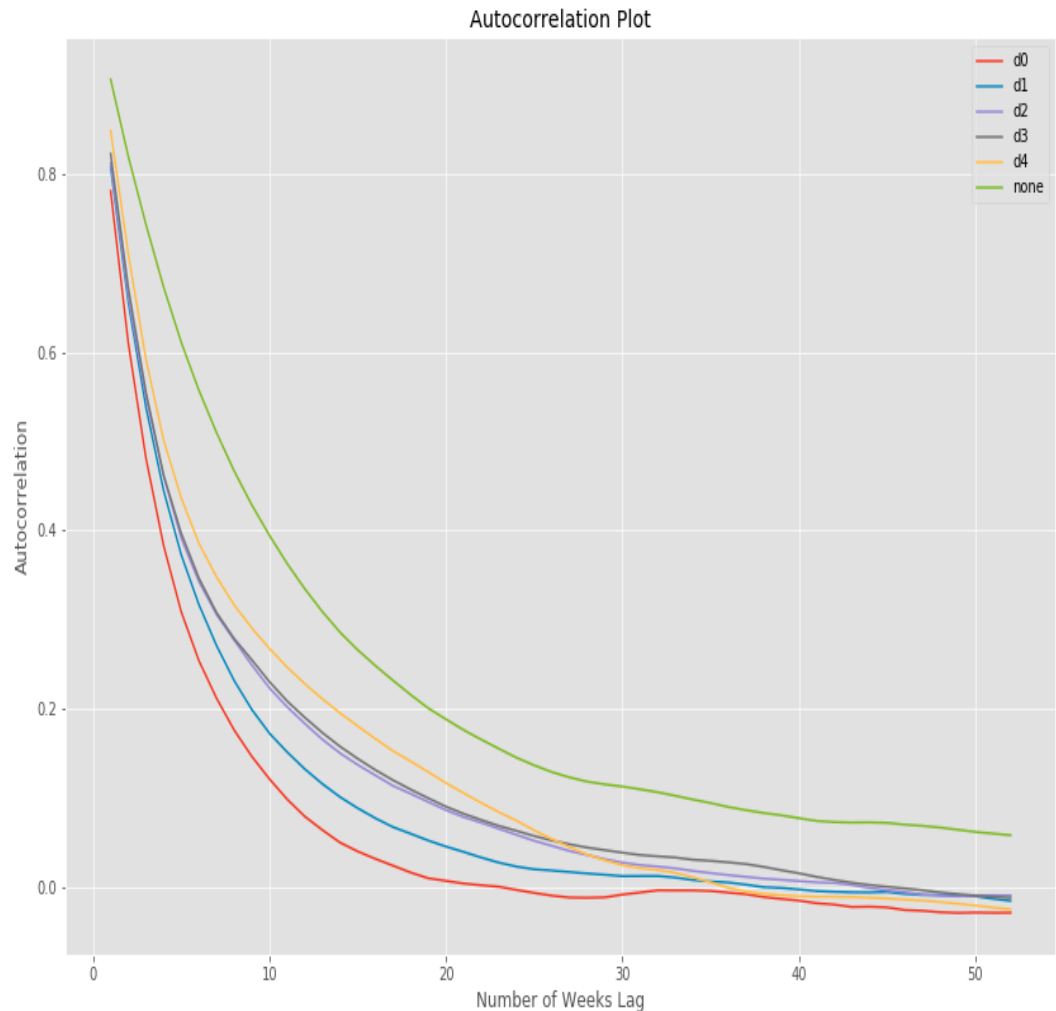
- Hyperparameter tuning using Bayesian Optimization
- Number of Gradient boosted rounds determined by out of sample AUC in CV set
- CV set= random 12.5% split of Training Data
- Training Data: all counties in CA from 2011 to 2014
- Test Data: all counties in CA during 2015
- 165 Features
 - Ratios from water usage data
 - Lagged drought severity

Time Series Analysis

- The autocorrelation of the drought severity categories were calculated for each county in the dataset
 - Then autocorrelation was averaged across each county
- lag



Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**



D0: Top Features From Importance Array

('d1_5_Week_lag', 108)	('d1_6_Week_lag', 28)	('none_7_Week_lag', 20)
('d0_5_Week_lag', 101)	('d1_9_Week_lag', 27)	('none_8_Week_lag', 19)
('d0_11_Week_lag', 64)	('d2_9_Week_lag', 26)	('dom_per_cap', 19)
('d1_8_Week_lag', 54)	('irrigation_1', 25)	('ind_1', 18)
('d2_11_Week_lag', 52)	('none_11_Week_lag', 25)	
('none_5_Week_lag', 44)	('d0_8_Week_lag', 22)	
('dom_sup_7', 44)	('pub_sup_1', 21)	
('none_6_Week_lag', 33)	('d1_7_Week_lag',	
('d0_6_Week_lag', 30)		
('d1_11_Week_lag', 29),		

D0: Prediction Results

	AUC
Precision Recall Curve	0.924
ROC Curve	~ 1

