TheDataOpen

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

Problem Statement

Welcome to Correlation One's UK Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

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. The World Health Organization (WHO) reports that with access to safe water, children demonstrate much better health outcomes, enabling them to focus on education and achieve more in life. Water can be thought of as not just essential for human health, but for the health of a society as a whole.

Various global initiatives such as the Millennial Development Goals (MDG) led by UNICEF and the WHO have been implemented to improve water access and quality. These efforts have resulted in a dramatic increase in the percentage of the global population that has access to "improved water sources", which keep water supplies free from contamination, from 76% in 1990 to 91% in 2015. However, hundreds of millions still live without access to clean water, especially in rural areas.

While the populations with the least access to good water sources are largely clustered in developing countries, recent events such as the Flint water crisis in Michigan and the California drought remind us that even in economically advanced countries such as the United States, careless public planning and negligent environmental regulations can still threaten the public's access to clean and safe public water supplies. In Flint, scientific reports indicate that the inadequate addition of the appropriate corrosion chemicals to the water sources, led to increased lead exposure of many residents and particularly doubling that in children. In this study, they outlined that children have higher water-soluble lead absorption than that of adults, posing a significant threat to their development.

Industrial and technological revolutions which have resulted in advanced manufacturing processes, have slowly resulted in water pollution of nearby areas from the contaminants and by-products released into the natural water sources. As we proceed with the development of novel

由 扫描全能王 扫描创建

technologies, it is imperative that we continue to improve the quality of life and access to safe drinking water. The effects of pollution in water sources will trickle down and potentially lead to unprecedented changes of the local/global environment, ecology of species, and human development.

Your Task

Your goal is to analyze the chemicals, droughts, water usage and industry occupational data (described in detail below), potentially in combination with supplementary datasets, in order to increase the understanding of how various factors such as natural events and governmental initiatives have influenced the environment through time, specifically pertaining to the quality of water in residential areas.

We have pre-cleaned several supplementary datasets for your use. Additional data is available, including details about educational attainment of the population and earnings by industry.

You are asked to pose your own question and answer it using the available datasets in the available time. What is important is the insightfulness and depth of your conclusions and analysis. **You need not be comprehensive; quality data analysis will be rewarded over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to predict or model water supply trends. Submissions may also be illuminating, through use of thoughtfully chosen data visualizations or sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question generally has a positive effect on judges' assessment of your submission; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: How do water quality measures correlate with quality of life measures/SES factors?

Sample Question 2: What counties are most vulnerable in the event of a drought? Do droughts have an effect on industry specific earnings?

Sample Question 3: Does a relationship exist between the major type of industries (i.e; manufacturing) and the quality of water? Are there greater concentrations of potentially hazardous contaminants in areas of more industrial manufacturing?

由 扫描全能王 扫描创建

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Box and are spread across six tables. Your team should only use the tables that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to "play nice" with each other.

chemicals
Data containing the measured mean concentration of a particular chemical or substance in community water systems throughout the counties in the United States from 2000 – 2016.
~882,000 rows & 12 columns. Size: ~100MB. Source: Centers for Disease Control and Prevention.

droughts
Data containing the particular percentage of various range of drought severities, indexed by counties for particular start-end periods throughout the United States.
~1.35 million rows & 11 columns. Size: ~100MB. Source: U.S. Drought Monitor.

earnings
Information about the industry specific median earnings (in that specific year's USD, inflation adjusted) indexed by counties for all of the United States, taken from 2010 – 2016.
21,999 rows & 31 columns. Size: ~5MB. Source: U.S. Census.

educational_attainment
Data containing the educational attainment of the US population by county from 1970 – 2000 and estimates of 2012 – 2016.
16,416 rows & 12 columns. Size: ~2MB. Source: U.S. Department of Agriculture.

Industry_occupation
Data containing the estimated working population (16 years and over) for the various industries indexed by counties, taken from 2010 – 2016.
5,712 rows & 18 columns. Size: ~0.7MB. Source: U.S. Census.

water_usage
Information about particular water usage (irrigation, public supply, crop, etc.) and thermoelectric power generated for counties that were found for the year 2010.
3,225 rows & 117 columns. Size: ~2MB. Source: U.S. Department of the Interior.

由 扫描全能王 扫描创建

Additional Datasets

You are welcome to scour the Web for custom datasets to supplement your analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's technical product team if you believe your idea is worthy of an exception).

Other Materials

We will provide you the schema for each of the data tables in another packet.

We will also provide you a Datathon manual at registration, which contains a section on using Box. This will show you how to download the datasets (described above) and upload your submissions (described below).

Submissions: Content

Submissions should have two components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
    b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your work without your team there to explain it; therefore, **your submission must "speak for itself"**. It need not be polished to the level of a final product, but do ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

You will be evaluated based on your Report, as follows:

由 扫描全能王 扫描创建

- **Non-Technical Executive Summary**
  - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
- **Technical Exposition**
  - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypotheses tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

However, please also include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your raw LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be provided a sheet with your team's Box account login details when the hacking session begins; you will be using the account to download the datasets as well as to upload your submission content. We recommend that you wrap up your work by 3:15 PM and begin uploading your submission at that time. **Submissions MUST be received by 3:30 PM. Any submission received after 3:30 PM will NOT be evaluated by the judges**.

Tips & Recommendations

由 扫描全能王 扫描创建

You will have ~11.5 hours total to work on the problem statement. However, you will not have access to the actual data until the morning of the competition. As such, we recommend you split your time as follows:

- Friday evening, ~7:30PM – 12:00AM: You will receive a copy of the problem statement, data table schema, and data table heads. This gives you the opportunity to study the available data fields, think about suitable questions to tackle, and plan out your exploration process. Additionally, the data table heads should be sufficient for you to begin putting together some data wrangling & cleaning scripts.
- Saturday, 8:30AM – 3:30PM: You will receive the actual data. If you set up your data munging scripts already, you should be able to quickly apply them and immediately begin working with the data. You should spend most of your day investigating the data, performing qualitative & quantitative analysis, and writing up your process & results.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

Finally, **we STRONGLY encourage you to start typing up your final submission AT LEAST three to four hours before the submission deadline**. In the past, many teams have spent a lot of time conducting great analyses, only to realize that they left almost no time for actually writing up and presenting their results. **This cannot be stressed enough – quality data analysis that is incomplete or poorly presented will NOT win one of the top prizes**.

Ask for Help

The Datathon team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.

由 扫描全能王 扫描创建

Data Table Schema

chemicals

Data about the measured mean concentration of a particular chemical or substance in community water systems throughout the counties in the United States from 2000 – 2016. *~882,000 rows & 12 columns.* Size: ~100MB. Source: Centers for Disease Control and Prevention.

Field	Type	Description
cws_name	STRING	Name of community water system
chemical_species	STRING	Type of chemical contaminant or particulate
contaminant_level	STRING	Classification of contaminant measurement (three types: less than/greater than mean concentration level, non-detect)
county	STRING	County name
pws_id	STRING	Public water supply identification code
pop_served	INTEGER	Total population which public water supply serves
state	STRING	Name of state
unit_measurement	STRING	Measurement units of the contaminant
value	FLOAT	Measure contaminant value, in measurement units
year	INTEGER	Year of data measured
fips	INTEGER	FIPS code for the particular state+county
state_fips	INTEGER	FIPS code for state

droughts

Data containing the particular percentage of various range of drought severities, indexed by counties for particular start-end periods throughout the United States. *~1.35 million rows & 11 columns.* Size: ~100MB. Source: U.S. Drought Monitor.

Field	Type	Description
fips	INTEGER	FIPS code for the particular county
county	STRING	County name
state	STRING	State code (2-letters)
Classification of Droughts (6 total)	FLOAT	Percentage of population affected by severity of droughts by: no drought (none), D0 = abnormally dry, D1 = moderate, D2 = severe, D3 = extreme, D4 = exceptional Additional Information
valid_start	STRING	Start date of event in form of YYYY-MM-DD

由 扫描全能王 扫描创建

valid_end	STRING	End date of event in form of YYYY-MM-DD

earnings

Information about industry-specific median earnings (in that year's USD, inflation adjusted) indexed by counties for all of the United States, taken from 2010 – 2016.
21,999 rows & 31 columns. Size: ~5MB. Source: U.S. Census.

Field	Type	Description
geo_id	STRING	Particular ID used by U.S. Census
fips	INTEGER	FIPS code of county as (YYZZZ)
county	STRING	Name of county and state in form (county, state)
total_med	INTEGER	Total median earnings in the particular county
Median Industry Income (26 total)	INTEGER/ STRING	Specific median earnings for various industries. * There may be instances when the value is "2,500-" or "250,000+" which indicates: '-' following a median estimate means the median falls in the lowest interval of an open-ended distribution. An '+' following a median estimate means the median falls in the upper interval of an open-ended distribution. Additional Description
year	INTEGER	Year of data collected

education_attainment

Data containing the educational attainment of the US population by county from 1970 – 2000 and estimates of 2012 – 2016, based on U.S Census Data & American Community Survey.
16,416 rows & 12 columns. Size: ~2MB. Source: U.S. Department of Agriculture.

Field	Type	Description
fips	INTEGER	FIPS code of county as (YYZZZ)
state	STRING	State code (2-letter) or country (US)
county	STRING	Name of county
year	INTEGER	Year in which data was recorded for of form YYYY.
less_than_hs	INTEGER	Number of people with less than a high school diploma
hs_diploma	INTEGER	Number of people with only a high school diploma
some_college_or_associates	FLOAT	Number of people with some college (1-3 years) or an Associate's degree

由 扫描全能王 扫描创建

college_bachelors_or_higher	INTEGER	Number of people with a college or Bachelor's degree or higher
Percentages of various educations (4 total)	FLOAT	Percentage of each respective educational degree as a total of 100.

industry_occupation

Data containing the estimated working population (16 years and over) for the various industries indexed by counties, taken from 2010 – 2016.
5,712 rows & 18 columns. Size: ~0.7MB. Source: U.S. Census.

Field	Type	Description
geo_id	STRING	Particular ID used by U.S. Census
fips	INTEGER	FIPS code of county as (YYZZZ)
county	INTEGER	County
Various industries (14 total)	INTEGER	Estimated working population in that particular industry
Year	INTEGER	Median household income

water_usage

Information about particular water usage (irrigation, public supply, crop, etc.) and thermoelectric power generated for counties that were found for the year 2010.
3,225 rows & 23 columns. Size: ~2MB. Source: U.S. Department of the Interior.

Field	Type	Description
state	STRING	State code
state_fips	INTEGER	FIPS code for state (1-78)
county	STRING	Name of particular county
county_fips	INTEGER	FIPS code for the particular county (001-840)
fips	INTEGER	Total FIPS of county of form YYZZZ (YY = state_fips, ZZZ = county_fips)
year	INTEGER	Year of data in form of YYYY
population	FLOAT	Total population of county, in thousands
Parameters (110 total)	FLOAT/ INTEGER	Information of various water metrics

water_usage_dictionary

Detailed description of each of the fields with units of measurement.
117 rows & 2 columns Size: 8KB. Source: U.S. Department of the Interior.

由 扫描全能王 扫描创建