TheDataOpen

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

Problem Statement

Welcome to Correlation One's 2018 Dublin 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

Among many other factors, technological developments and scientific advances in the fields of medicine and healthcare have led to progressive increases in average human life expectancy. The emergence of vaccines and antibiotics are prime examples of such technologies. However, to this day humans still remain extremely susceptible to many diseases that likely result from poor living, eating habits and environmental conditions.

There are myriad factors that are responsible for determining an individual's health, but many studies have honed in on the importance of particular life choices. Compelling research (Behavioral Risk Factor Surveillance System - BRFSS) has shown that tobacco use and smoking are a leading cause of cardiovascular diseases, claiming roughly 30,000 adult lives per year in the New York State. Additional factors that have been shown to negatively affect health include lack of physical activity and poor dieting from frequent outings to fast food chains and restaurants.

Inadequate environmental and living conditions are also fundamentally important to consider and have been linked to health deficiencies. Air pollution from industrial factories and transportation, radiation from nuclear power-stations, water contamination from various treatment plants, and poor housing conditions can also lead to further increases in the probabilities of developing disease or sickness.

Since 2010, the Centers for Disease Control and Prevention (CDC) implemented the Health People 2020 Initiative, which outlined a set of leading health indicators monitored throughout the span of 10 years to assess the health of individuals in the United States. The ultimate goal is to analyze the data subject to the health indicators to design novel government strategies and efforts to improve the health and well-being of the nation.

These initiatives resonate with the New York State Department of Health, which implemented similar policies and objectives in the Prevention Agenda 2013-2018 to address and promote healthy living. As we slowly transition to a more conscientious mindset regarding health, these

next 10 years will prove to be an exciting era, as they will showcase how effective these various measures and actions have impacted residential health in the state of New York.

Your Task

Your goal is to analyze the 311 service requests and food establishment inspections data (described below), potentially in combination with supplementary datasets, in order to increase understanding of how governmental initiatives and socioeconomic factors in New York State have influenced residential health.

We have partially pre-cleaned several supplementary datasets for your use. Additional data is available, including details about environmental radiation, community health indicators, and restaurant and grocery store locations. We also provide demographics information on income and health insurance benefits.

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 public health 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 socioeconomic factors such as income or availability of health insurance influence the living quality of a community or county?

Sample Question 2: In a particular community, is there a relationship between proximity and/or concentration of restaurants or grocery stores and a reportedly low quality of health?

Sample Question 3: How do environmental or living conditions impact the overall health of individuals? Is there a relationship between frequency of service requests and poor health indicator measurements in a given county?

Datasets

The provided datasets are stored in your team's USB drive that you received at registration, and are spread across eight tables. (Alternatively, if you do not have a USB portal, they are also stored in the "Datathon Materials" folder on your team's Box account (described later).) 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.

311_service_requests
Service requests made to the 311 hotline in the New York City metropolitan area (agency, complaint type, descriptor, location, etc.).
~1.1 million rows & 25 columns. Size: ~70MB zipped, ~425MB unzipped. Source: NYC Open Data - 311.

food_service_esablishment_inspections
Information about food service establishment (facility, address, date of inspection, location etc.), the inspection description, and any violations that were found since 2005 in New York State.
~1 million rows & 23 columns. Size: ~90MB zipped, ~425MB unzipped. Source: New York State Department of Health.

community_health
Monitored values of various health indicators (health topic, county, indicator, event count, location, etc.) for every county in the State of New York, taken from periods throughout 2008 – 2015.
24,564 rows & 13 columns. Size: ~5MB. Source: NY Department of Health.

demographics_city
New York City demographic data (population, age, income, etc.) organized alphabetically by Neighborhood Tabulation Area (NTA).
188 rows & 33 columns. Size: ~0.1MB. Source: NYC Metropolitan Transportation Authority.

demographics_state
Demographic data (households, income level, health insurance coverage etc.) organized by New York State counties, taken from 2011 – 2016. All income and earnings values are denominated in that particular year's dollars.
378 rows & 26 columns. Size: ~0.1MB. Source: U.S. Census.

environmental_radiation
Environmental samples and measurements of radioactive materials (sample type, sampling frequency, location, isotope, value, etc.) taken at various power plants in New York State from 2001 – 2016.
20,485 rows & 10 columns. Size: ~2MB. Source: New York State Department of Health.

food_establishments_inspections_city
Information about food service establishment (name, address, type of cuisine, date of inspection, location etc.), the inspection description, and any violations that were found at the time for particularly NYC (all boroughs).
~375,000 rows & 15 columns. Size: ~17MB zipped, ~135MB unzipped. Source: NYC Open Data.

food_venues
Information about restaurants and grocery stores (business name, street address, city, state, location, etc.) across the United States.
~750,000 rows & 9 columns. Size: ~17MB zipped, ~85MB unzipped. Source: Leads Deposit (private).

geographic
Data about the shape of each NTA (latitude and longitude coordinates, in order) in New York City.
9,302 rows & 195 columns. Size: ~4MB. Source: NYC Open Data.

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). It will also provide you helpful tips for the competition, so please read it carefully!

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:

- **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

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.