Summer Invitational Datathon 2024 Problem Statement

Welcome to the Summer Invitational Datathon 2024! 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

There are approximately seven hundred and fifty thousand restaurants in the United States representing myriad countries and cultures. Despite this range of dining options, American cuisine is still synonymous with greasy, oversized hamburgers, towering milkshakes, and sweet, syrupy apple pie. And food isn't just a hallmark of American culture, it's a crucial part of the economy in the United States. From first jobs in food service, to careers at processing plants, the country's relationship with comfort foods has endured generations.

The relationship isn't necessarily a healthy one. Evidence continues to mount against sugar-laden, processed products as being bad for our overall health. In 2022, 22 states had an adult obesity rate at or above 35%. This is up from 19 states in 2021, and just ten years ago no state had an obesity rate at that level. Politicians have tried to introduce legislation to ban soft drinks with limited success, despite the fact that it's been associated with weight gain and an increased risk of diabetes. And it isn't just sugary drinks. Processed meats have also been linked to obesity and heart disease. Meanwhile McDonalds stock has tripled in value over the last ten years.

So why can't we quit junk food? Or should we attempt to quit at all? After all, America is "the land of the free" – to eat whatever you want, whenever you want, wherever you want – to have an entire meal served to you through your driver's side window. And while that Big Mac is always served up quickly, it doesn't appear from thin air. Multiple people play a role: servers, plant workers, farmers, truck drivers, even the engineers who design the machines to produce the food at scale. Not all of the jobs are glamorous, but they all put food on someone's table.

While healthy eating slowly makes inroads, from Erewhon smoothies and Michelin star restaurants going vegan, processed foods don't appear to be going away anytime soon, nor the economy that supports it. With all of the data available, is there a way to better understand the true costs associated with this lifestyle?

Your Task

Your goal is to analyze the provided datasets, potentially in combination with supplementary datasets, in order to better understand the footprint of processed food in the United States.

We have partially pre-cleaned several supplementary datasets for your use. In addition to the meat production numbers, we are including obesity data, as well as information around domestic meat production and supply in the United States. In addition we have pulled historic high/low/open/close prices for stocks in sectors related to processed foods or their production.

You are asked to pose your own question and answer it using the available datasets as well as any supplementary datasets you may find. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is more important than breadth of the question posed.**

Submissions may be predictive, using machine learning to classify or predict patterns. Submissions may also be illuminating through use of data visualizations or through sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: Are meat production yields predictive of restaurant stock prices?
Sample Question 2: Does the price of sugar impact youth consumption of sugar drinks, and are all regions impacted equally?
Sample Question 3: Do periods with low meat production track with overall unemployment numbers?

Datasets

The provided datasets are stored in the “Datathon Materials”, and are spread across seven tables. Your team should only use the tables that are relevant to your chosen question. 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.

Nutrition Physical Activity and Obesity Data

Information on obesity and reported physical activity in the US at a state level from 2001-2022.
~ 133K rows and 34 columns. Size: 48.6 MB. Source: CDC Behavioral Risk Factor Surveillance System

Meat Statistics
Production, slaughter, and supply information for domestic livestock and meat in the United States
Meat Production ~*13.6K rows & 8 columns.* Size: 962 KB
Slaughter Counts ~*15.2K rows & 8 columns.* Size: 1.1 MB
Slaughter Weights ~*9.3 rows & 9 columns.* Size: 835 KB
Cold Storage ~*4K rows & 7 columns.* Size: 218 KB
Source: Livestock & Meat Domestic Data

Processed Food Stocks and Commodities (HLOC)
A collection of daily high/low/open/close prices for stocks associated with processed foods (from restaurants to farm and plant equipment manufacturers) and four exchange-traded funds (ETF) that track major indices (S&P 500, Nasdaq, Dow Jones Industrial Average), some general information about these companies, and daily prices for relevant commodities.
Stocks and ETFs ~*163K rows & 7 columns.* Size: 7.7 MB
Commodities ~*1K rows & 4 columns.* Size: 48 KB
Descriptions *32 rows & 6 columns.* Size: 9 KB
Source: Alpha Vantage and Yahoo! Finance

American Community Survey - Selected Economic Characteristics (5 Year Estimates)
The American Community Survey (ACS) is an ongoing survey conducted by the US Census Bureau that provides vital information on a yearly basis about various social and economic characteristics across demographics.
~98K *rows & 8 columns.* Size: 12.1 MB. Source: American Community Survey

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.

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, the 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 technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

<u>Submissions: Evaluation</u>

The competition will have multiple rounds of evaluation. Your Report will be judged as follows:

- **Non-Technical Executive Summary**
    - o *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**
    - o *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.
    - o *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis 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?
    - o *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 LaTeX source file.

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

Submissions MUST be received by 5PM EST on Sunday, August 4th, 2024. Any submissions received after that time will NOT be evaluated by the judges.

Tips & Recommendations

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 their results. **This cannot be stressed enough – quality data analysis that is incompletely presented will NOT win one of the top prizes**.

Things to Consider:

- Did you account for significant exogenous influences (e.g. Covid)?
- Is your report structured and complete (i.e. does it have a conclusion or does it end abruptly)?
- Is there intention behind your approach, or are you presenting the results of various testing methods with no meaningful takeaways?

- As much as this is an exercise in data analysis it is also a competition – are you setting yourself apart from other submissions with a creative approach or presentation?

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.