Women’s Datathon 2025 Problem Statement

Welcome to the 2025 Women’s 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

Gasoline remains one of the most heavily consumed fuels in the United States and is the primary output of domestic oil refineries. According to the 2020 U.S. Census, approximately 92% of American households own at least one vehicle—underscoring the country’s deeply ingrained car culture. Even with the surge in electric-vehicle adoption, fully electric cars made up less than 10% of new-vehicle sales in Q1 2025, meaning more than 9 out of 10 vehicles still rely on gasoline for at least part of their power.

Yet, the connection between routine driving and broader environmental issues like climate change can often seem intangible. What tends to resonate more immediately with the public are fluctuations in gas prices. These prices are shaped by a variety of factors, including global supply and demand dynamics, crude oil benchmarks, refining and distribution costs, marketing expenses, and government-imposed taxes. Many of these elements are tracked by agencies such as the U.S. Energy Information Administration (EIA) and the Federal Reserve.

Gasoline prices also serve as a highly visible, albeit imperfect, indicator of broader economic conditions. Shifts at the pump can impact consumer sentiment and even correlate with political approval ratings. Low prices typically increase household purchasing power and reduce operational costs in key sectors like transportation and manufacturing. On the other hand, higher prices—while burdensome—can incentivize conservation and investments in sustainable energy alternatives. Regardless of price trends, companies along the fuel supply chain are accountable to their investors.

Given that car travel will continue to dominate U.S. transportation for the foreseeable future, understanding the behavior of gasoline prices remains economically and politically significant. With the wealth of available data, can we better anticipate these shifts—and more importantly—understand their ripple effects across the economy?

Your Task

Analyze the provided U.S.-based datasets—including historical gasoline prices (in $/gallon), transportation statistics, weekly supply estimates, and energy sector stock prices. You may also incorporate international indicators (e.g., Brent crude, OECD data) to enhance your analysis.

You are expected to formulate your own research question and use the data to investigate it. The originality of your question and the depth of your insights are more valuable than covering every dataset. Use as many—or as few—sources as needed to develop a compelling, evidence-backed argument.

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.

You are encouraged to develop your own research question. However, below are some suggestions:

1. Do energy companies see declining stock performance when gas prices fall?

2. Can gasoline prices serve as reliable predictors of broader market shifts?

3. Are there early warning signals consumers can track to anticipate fuel price hikes or drops?

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.

Weekly Gasoline Prices
Weekly reported gasoline prices in the United States from 2000-2023 across fuel types. ~ *514K rows and 7 columns*. Size: 86.3 MB. Source: U.S. Energy Information Administration

Weekly Supply Estimates
Production, Utilization, Imports, Exports, and Supply of Oil and Gasoline in the United States
Definitions for relevant key terms can be found here
1,253 rows & 25 columns. Size: 167 KB. Source: Petroleum & Other Liquids: Weekly Supply Estimates

Monthly Gasoline Makeup Percentages
Monthly percentage attribution of refining, distribution and marketing, taxes, and crude oil.
549 rows & 9 columns. Size: 31 KB. Source: Methodology for Gasoline and Diesel Fuel Pump Components

Energy Sector Stocks and Commodities (HLOC)
A collection of daily high/low/open/close prices for stocks in the energy sector 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 energy commodities.

Stocks and ETFs ~*145K rows & 7 columns.* Size: 7.1 MB
Commodities ~27K *rows & 5 columns.* Size: 1.9 MB
Descriptions 29 *rows & 6 columns.* Size: 14 KB
Source: Alpha Vantage and Yahoo! Finance

Monthly Transportation Statistics
The latest monthly and quarterly data from across the government and the transportation industry.
289 rows & 137 columns. Size: 309 KB. Source: U.S. Dept of Transportation - Bureau of Transportation Statistics

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.

Submissions: Evaluation

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 1:30PM EST on Sunday, May 18th, 2025. 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**.

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.