Data Analyst InternTest
Level 3

Congrats **Vatsal Mehta (candidate no -A256)** on Clearing the first two rounds on Data Science Test. We wish you luck for this final round.

• **Test Instructions:**

The test is focused on data scraping, cleaning and structuring. An example dataset is shared for a better understanding of the task. Carefully read the following instructions properly and review the shared data example before you start the test.

1. **The test solution code should be shared in a python file.** (Any other submissions will be disqualified)
2. The result datasets for the test can be shared in **XLSX** or **CSV.**
3. Write a readme file containing a brief description of your approaches and explaining your thought process.
4. Submit your test solution with the required files to hr@mindworksglobal.com with '**Data Analyst Internship Submission**' as the subject line (Other submissions may not be considered)
5. **The test solution must be submitted by within 24 hours** (from time the test was given)

6. For any query regarding the test, you can write to hr@mindworksglobal.com

• Tasks

1. Data Scraping using Scrapy Framework (dataset_energygov_scrapy.xlsx):

1. Use scrapy for creating a web scraper for the the news articles listing on the site below: https://www.energy.gov/listings/energy-news

   The scraper should:

   - Be able to crawl to next pages on the sites (ex: crawling till Page 4)
   - Scrape news article details such as article date, article headline, article url, article short description available from the listing

2. Clean and Structure the scraped data as shown in the given in the excel file (**energygov_scrapy.xlsx**)
3. Save the data as a separate excel file.

Note: scrapping should be implemented through code (.py).

2. Data Scraping using Selenium Web Driver (energygov_scrapy.xlsx):

1. Use selenium web driver for creating a web scraper for the the news articles listing on the site below: https://www.energy.gov/listings/energy-news

   The scraper should:

   - Be able to click to **next page** button on the site, to extract more news listings. (ex: till Page 4)
   - Scrape news article details such as article date, article headline, article url, article short description (if available) from the listing.

2. Clean and Structure the scraped data as shown in the given in the excel file (**energygov_scrapy.xlsx**)
3. Save the data as a separate excel file.

Note: scrapping should be implemented through code (.py).