Project Description

PJM Interconnection, as the first Independent System Operator (ISO) in the U.S., coordinates the movement of wholesale electricity in parts of the U.S. East and Midwest, including 13 states and the District of Columbia. Headquartered in Valley Forge, Pennsylvania, PJM has grown to become a critical nexus in the U.S. electricity market, overseeing an expansive transmission system that ensures a reliable electric supply to over 65 million people.

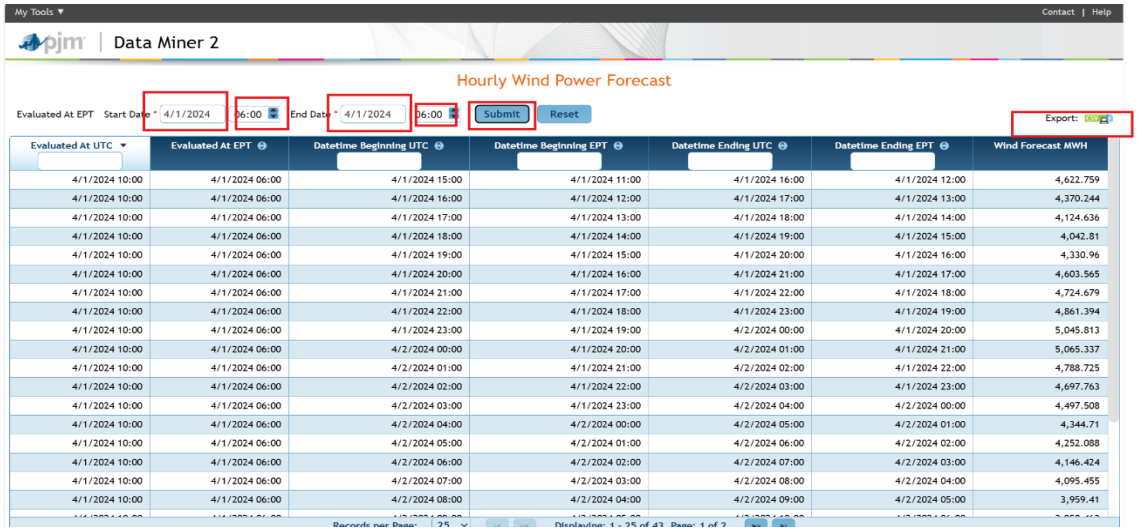
Each day, PJM publishes crucial information related to the market. For your task, you will be responsible for predicting the power generation by wind plants in the RTO Area of PJM.

- Under the project email, you will find two files with each the name *Wind_Actual.csv* and *Wind_Forecast.csv*:
  - *Wind_Actual* contains actual power generation by wind plants.
  - *Wind_Forecast* contains forecasted power generation by wind plants made 1 day prior to the actual power generation. (For example, the forecast value for 2024-03-02 was predicted on 2024-03-01).
  - Both files contain data from 2023-01-01 to 2024-03-31 inclusive
  - Both files contain the following columns:
    - DATE: Target date of forecast/actual wind generation
    - HE: Hour Ending. It is a way to reference the end of an hourly period. For example, *HE 1* represents the time interval from 2023-03-31 00:00:00 to 2023-03-31 01:00:00.
    - MW: Total MW of forecast/actual power generation regardless of specific area
  - The *Wind_Actual* table includes an additional column labeled *Area*. Please note that all values in this column will be *RTO*, which represents the RTO Area within PJM.

- **Submission Guideline:**
  - Your project should be completed using Python, and we strongly recommend utilizing ipynb format so that we can track your intermediate steps. A good place to format ipynb is Google CoLab where you will also have access to their shared GPU if you are using a NN model.
  - **This project is due by 2024-04-22 11:59PM EST.** Late submissions will not be accepted.
  - While prediction accuracy will be considered during our recruitment process, we value more on the cohesion and quality of your code. Therefore, focus more on developing clear, well-structured code rather than stressing excessively over the modeling aspect.
  - To submit your project, please send us an email to **Yuwen.Zhang@5DimensionsEnergy.com** with the subject line: *2024-04-22 Project [First Name] [Last Name]*. Attach a zip-file that includes the following:
    - Your Python code file
    - The final input CSV file that you use to train model and generate your prediction
    - The final model that generates your prediction (.h5, .pkl or relevant format)
    - Your prediction result file
      - The prediction CSV should have the name *2024-04-22_Pred_[First Name]_[Last Name].csv*, and should follow the same format of the *Wind_Actual* file (Same columns with 24 rows representing each HE of 2024-04-22)
  - If your file size exceeds the maximum attachment size for an email, you can also provide us with a google drive link to your zip-file, make sure to grant access to the zip-file.

- **Goal 1 (Optional): Scrap actual and forecasted wind generation from 2024-03-31 onwards**
  - Please note that the file we provide only contains data up to 2024-03-31, and your final goal is to predict the power generation by wind plants on 2024-04-22. We encourage you to develop a data scraping code to download the most recent data. Please note this is the only source from which you should download data to enhance your dataset with the most recent forecast and actual data information. No additional external or public data should be included in your project.

  - **For Forecast Data:**
    - Go to **https://dataminer2.pjm.com/feed/hourly_wind_power_forecast**. The *Hourly Wind Power Forecast* dashboard contains the hourly Wind Forecast within PJM Area. The forecast is updated every 10 minutes and covers roughly the next 42 to 46 hours, five hours from the forecast evaluation time. For example, at 2024-04-01 06:00 EPT, PJM will forecast the wind generation from 2024-04-01 11:00 EPT to 2024-04-03 05:00 EPT.
    - For consistency and fairness, reference only forecast made at 06:00 EPT for next day's forecasted generation in your project. For example, use forecast evaluated at 2024-04-01 06:00 EPT for the 24-hour wind generation forecast on 2024-04-02.
    - To query from this table, you need to enter the *Start Date* and *End Date* of the forecast evaluated time in EPT. This parameter stands for EPT timestamp that the forecast is made. In our project, for example, to access forecast data on 2024-04-02, enter 4/1/2024, as both the *Start Date* and *End Date*, and append 06:00 as the time. After setting your dates and times, click *Submit* to access the data. Keep in mind that data is stored for only one month, so information older than 30 days is not available.



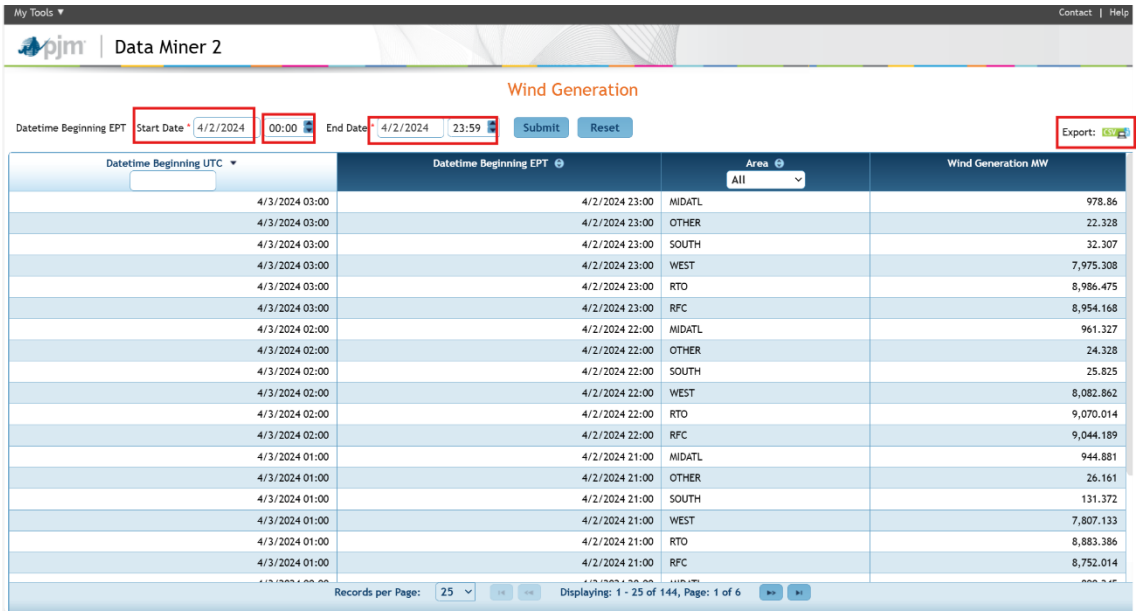
Evaluated At UTC	Evaluated At EPT	Datetime Beginning UTC	Datetime Beginning EPT	Datetime Ending UTC	Datetime Ending EPT	Wind Forecast MWH
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 15:00	4/1/2024 11:00	4/1/2024 16:00	4/1/2024 12:00	4,622.759
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 16:00	4/1/2024 12:00	4/1/2024 17:00	4/1/2024 13:00	4,370.244
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 17:00	4/1/2024 13:00	4/1/2024 18:00	4/1/2024 14:00	4,124.636
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 18:00	4/1/2024 14:00	4/1/2024 19:00	4/1/2024 15:00	4,042.81
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 19:00	4/1/2024 15:00	4/1/2024 20:00	4/1/2024 16:00	4,330.96
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 20:00	4/1/2024 16:00	4/1/2024 21:00	4/1/2024 17:00	4,603.565
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 21:00	4/1/2024 17:00	4/1/2024 22:00	4/1/2024 18:00	4,724.679
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 22:00	4/1/2024 18:00	4/1/2024 23:00	4/1/2024 19:00	4,861.394
4/1/2024 10:00	4/1/2024 06:00	4/1/2024 23:00	4/1/2024 19:00	4/2/2024 00:00	4/1/2024 20:00	5,045.813
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 00:00	4/1/2024 20:00	4/2/2024 01:00	4/1/2024 21:00	5,065.337
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 01:00	4/1/2024 21:00	4/2/2024 02:00	4/1/2024 22:00	4,788.725
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 02:00	4/1/2024 22:00	4/2/2024 03:00	4/1/2024 23:00	4,697.763
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 03:00	4/1/2024 23:00	4/2/2024 04:00	4/2/2024 00:00	4,497.508
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 04:00	4/2/2024 00:00	4/2/2024 05:00	4/2/2024 01:00	4,344.71
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 05:00	4/2/2024 01:00	4/2/2024 06:00	4/2/2024 02:00	4,252.088
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 06:00	4/2/2024 02:00	4/2/2024 07:00	4/2/2024 03:00	4,146.424
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 07:00	4/2/2024 03:00	4/2/2024 08:00	4/2/2024 04:00	4,095.455
4/1/2024 10:00	4/1/2024 06:00	4/2/2024 08:00	4/2/2024 04:00	4/2/2024 09:00	4/2/2024 05:00	3,959.41

    - Next, click the *Export* CSV button located at the upper right corner of the table to download the data. Below is a preview of the csv file, note that you only need to reference these columns:
      - datetime_beginning_ept and datetime_ending_ept: The values in these two columns are consistently one hour apart and are your target date and hour. For instance, if datetime_beginning_ept shows 2024-04-01 1:00:00 AM, then datetime_ending_ept should correspondingly show 2024-04-01 2:00:00 AM, which corresponds to *Date* = 2024-04-01 and *HE* = 2 in the format from the *Wind_forecast* file. You may need to perform a transformation to convert the downloaded data to match the *Wind_forecast* data.

- wind_forecast_mwh: Equivalent to the MW column in the *Wind_Forecast* file.

evaluated_at_utc	evaluated_at_ept	datetime_beginning_utc	datetime_beginning_ept	datetime_ending_utc	datetime_ending_ept	wind_forecast_mwh
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 15:00	4/1/2024 11:00	4/1/2024 16:00	4/1/2024 12:00	4622.759
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 16:00	4/1/2024 12:00	4/1/2024 17:00	4/1/2024 13:00	4370.244
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 17:00	4/1/2024 13:00	4/1/2024 18:00	4/1/2024 14:00	4124.636
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 18:00	4/1/2024 14:00	4/1/2024 19:00	4/1/2024 15:00	4042.81
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 19:00	4/1/2024 15:00	4/1/2024 20:00	4/1/2024 16:00	4330.96
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 20:00	4/1/2024 16:00	4/1/2024 21:00	4/1/2024 17:00	4603.565
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 21:00	4/1/2024 17:00	4/1/2024 22:00	4/1/2024 18:00	4724.679
4/1/2024 10:00	4/1/2024 6:00	4/1/2024 22:00	4/1/2024 18:00	4/1/2024 23:00	4/1/2024 19:00	4861.394

- **For Actual/Realized Data:**
  - Go to https://dataminer2.pjm.com/feed/wind_gen. The *Wind Generation* dashboard contains the hourly wind generation amounts in PJM. The data is updated daily at around 7:00 AM for the realized wind generation two days ago.
  - Similar to the forecast table, to query from this table, you need to set both a *Start Date* and an *End Date* in Datetime Beginning EPT. For example, to access actual wind generation data on 2024-04-02, you should enter 4/2/2024 00:00 EPT as the *Start Date* and 4/2/2024 23:59 EPT as the *End Date*. Once entered the dates and times, click *Submit* to retrieve the latest data.
  - Please be aware that actual wind generation data incorporates a two-day lag for the measurement readings to be recorded and posted. For example, after 2024-04-22 7 AM, you will be able to access data up to and including 2024-04-20. There will still be, however, entries for 04-21 and 04-22, but all values in the *Wind Generation MW* column will be 0. This indicates that the table is merely a placeholder and lacks actual measurement readings.



  - Next, click on the *Export* CSV button located at the upper right corner of the table to download the data. Below is a preview of the csv file, note that you only need to reference these columns:
    - datetime_beginning_ept: This column corresponds to the *Date* and *HE* column in the provided *Wind_Actual* file. For example, datetime_beginning_ept = 2024-04-10 1:00:00 AM corresponds to *Date* = 2024-04-10 and *HE* = 2.
    - wind_generation_mw: Equivalent to the MW column in the *Wind_Actual* file.

- area: Equivalent to the Area column in the *Wind_Actual* file. Keep only area == *RTO*

datetime_beginning_utc	datetime_beginning_ept	area	wind_generation_mw
4/9/2024 4:00	4/9/2024 0:00	MIDATL	227.603
4/9/2024 4:00	4/9/2024 0:00	OTHER	0
4/9/2024 4:00	4/9/2024 0:00	SOUTH	161.022
4/9/2024 4:00	4/9/2024 0:00	WEST	4445.173
4/9/2024 4:00	4/9/2024 0:00	RTO	4833.798
4/9/2024 4:00	4/9/2024 0:00	RFC	4672.776
4/9/2024 5:00	4/9/2024 1:00	MIDATL	226.894
4/9/2024 5:00	4/9/2024 1:00	OTHER	0
4/9/2024 5:00	4/9/2024 1:00	SOUTH	181.524
4/9/2024 5:00	4/9/2024 1:00	WEST	4464.872
4/9/2024 5:00	4/9/2024 1:00	RTO	4873.29

- We recommend you develop a web-scraping tool to automatically load the page, navigate to the previous mentioned two tables, input the dates, and download the CSV files, as data scraping is an essential skill at 5DE. However, we recognize that there might be time constraints for some. Therefore, if you feel unable to complete the auto-scraping process, you can either rely solely on the data we have provided, which should be sufficient for this project, or simply manually download the most recent data.
- Given that the project is due on 2024-04-22 11:59PM EST, the most recent actual wind generation data available will be from 2024-04-20. Therefore, if you opt to download data from 2024-04-01 onwards, whether automatically or manually, you have the flexibility to choose the date range up to whatever date you believe is necessary for completing the project. In other words, there is no strict requirement to use the latest data published immediately before the project's due date.
- Please ensure that your code clearly indicates if you utilize any recent data, specifying the exact dates of the data you have downloaded. Additionally, if you have developed a web-scraping tool, make sure to distinctly indicate this portion of your code.

- **<u>Goal 2: Predict actual power generation by wind plants of the RTO Area on 2024-04-22.</u>**
  - Train a model to predict the 24-hour wind generation for the RTO Area on 2024-04-22.
  - Feel free to use either actual data, forecast data, or both for modelling.
  - Choose any model you believe is appropriate for this task.
  - While it is not mandatory, you are encouraged to incorporate any steps you find beneficial, such as visualization, data preprocessing, or model tuning processes. Feel free to explore various models. However, please clearly select and specify the model you intend to use for the final prediction, along with the reasons for your choice. Ensure to include necessary and brief comments in your code to explain your decision-making process and methodology.