1. What data cleaning methods have you applied?

The following methods to clean and manipulate the data:

* Filtering Rows:

Used conditional statements to filter rows based on specific criteria.

* Handling Missing Values:
* Dropped rows or columns with missing values using dropna().
* Changing Data Types:
* Converted data types of columns using pd.to\_datetime() and astype().
* Converted “Calm” to “0” in the column of wind direction.
* Merge the data:

Use the “left on” to merge two tables to one.

* Grouping and Aggregating:

Grouped data by specific columns and aggregated information using functions like groupby().

* Feature Selection:

Selected specific columns or features for analysis.

* Data Visualization:

Used matplotlib for scatter plots and other visualizations.

* Machine Learning:

Utilized machine learning models such as KNN, DecisionTreeCalssifier, DecisionTreeRegression for predictive modeling.

* Error Handling:

Addressed errors related to DataFrame operations and model evaluations.

* Binning:

By dividing continuous data into several discrete intervals bins, the complexity of the data could be simplified.

1. Why have you chosen these methods over other alternatives?

The choice of methods may vary depending on the specific requirement and characteristics of the dataset, general knowledge(“Calm”), etc. Understanding the problem context and explore different approaches will help find the most suitable solution.

**Data Cleaning:**

**Handling Missing Values:** Methods like imputation or removal of missing data are commonly used. For example, imputing missing numerical values with the mean or zero, and imputing categorical values with the mode by grouping date.

**Datetime Conversion:** Converting date and time columns to datetime objects is essential for time-based analysis and is a standard practice.

**Feature Engineering:** Creating new features, such as the Maxi Energy Demand, Demand Category which can provide additional insights into the data.

**Visualization:** Utilizing scatter plots or data analysis and exploration. For example, visualizing the relationship between wind speed and energy demand.

**DecisionTreeClassification/KNN:** Both methods are effective discrete algorithm for classification, especially when we have better patterns in the data. The key point is how to split the dataset to suitable intervals. I have tried to set different section for the categories, which will bring diverse accuracy results.

**DecisionTreeRegression:** It is a linear relationship algorithm for data analysis. The data of energy demand is a continuous data which could use regression algorithm.

1. Explain the process of building your best model. How did you select the features included in the model? How does your model work?

**Define the Problem:**

Review the dataset, understand the type of task, define it's classification, regression, clustering, etc.

**Data Import:**

Import the dataset for data cleaning and preprocessing.

**Data Cleaning and Preprocessing:**

Handle missing values, outliers, and any data quality issues.

Preprocess the data, which may include normalization, encoding categorical variables, etc.

**Visualize data:**

Understand the characteristics of your data through visualization and statistical analysis.

**Feature Selection:**

Identify and select relevant features that contribute to the predictive power of the model.

Use techniques like correlation analysis, recursive feature elimination, or domain knowledge to guide feature selection.

**Split the Data:**

Split the dataset into training and testing sets.

**Iterative Refinement to find a better model:**

Run the machine learning algorithms based on the dataset build. If the model performance is not satisfactory, revisit earlier steps. Adjust features, try different models, or collect more data.

1. How effective is your model? How have you evaluated this?

The performance of a model is to evaluate the effectiveness. The data of Group 4:

**Accuracy:** Accuracy closes to 1.

**MSE:** is low

**R – Square:** close to 1.

1. What insights about weather and/or daily energy usage can you draw from your analysis? Discuss any significant results.

Depending on the graphics:

1. The correlation between temperature and energy demand, especially the maximum temperature.
2. The humidity influences energy demands.
3. With the season of the dataset is for summer of Victoria (Nov 2022 to Apr 2023), the rainfall and windspeed don’t impact the energy a lot.
4. What are the limitations of your results?

The maximum energy demand is a continuous dataset, and theoretically, using regression algorithms would have better results. However, we utilized both discrete and continuous algorithms to predict the maximum energy demand. Surprisingly, the results showed that the discrete algorithm achieved better accuracy. My understanding of this outcome is that the data obtained contains a significant amount of noise. The missing columns in the data have a substantial impact on the final results, leading to these findings.