Phase - 3 Report

Building Our project by loading and preprocessing the dataset.

1. Loading the Dataset:

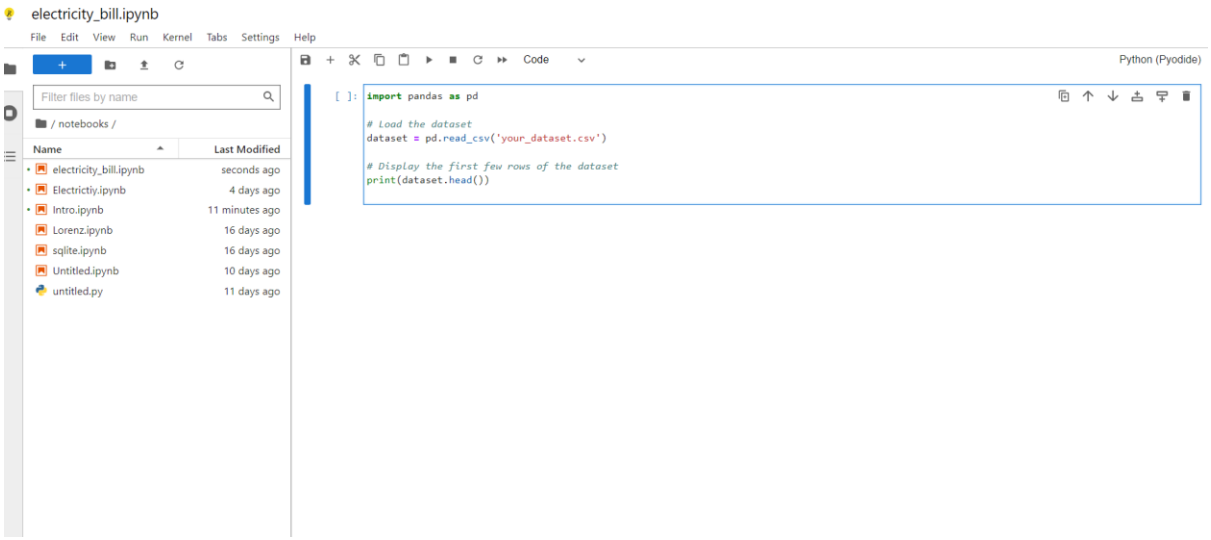
- Use appropriate libraries in your programming language (e.g., pandas for Python) to load the dataset into your project.

- Verify that the dataset has been loaded correctly by displaying the first few rows. This helps in understanding the structure and format of the data.

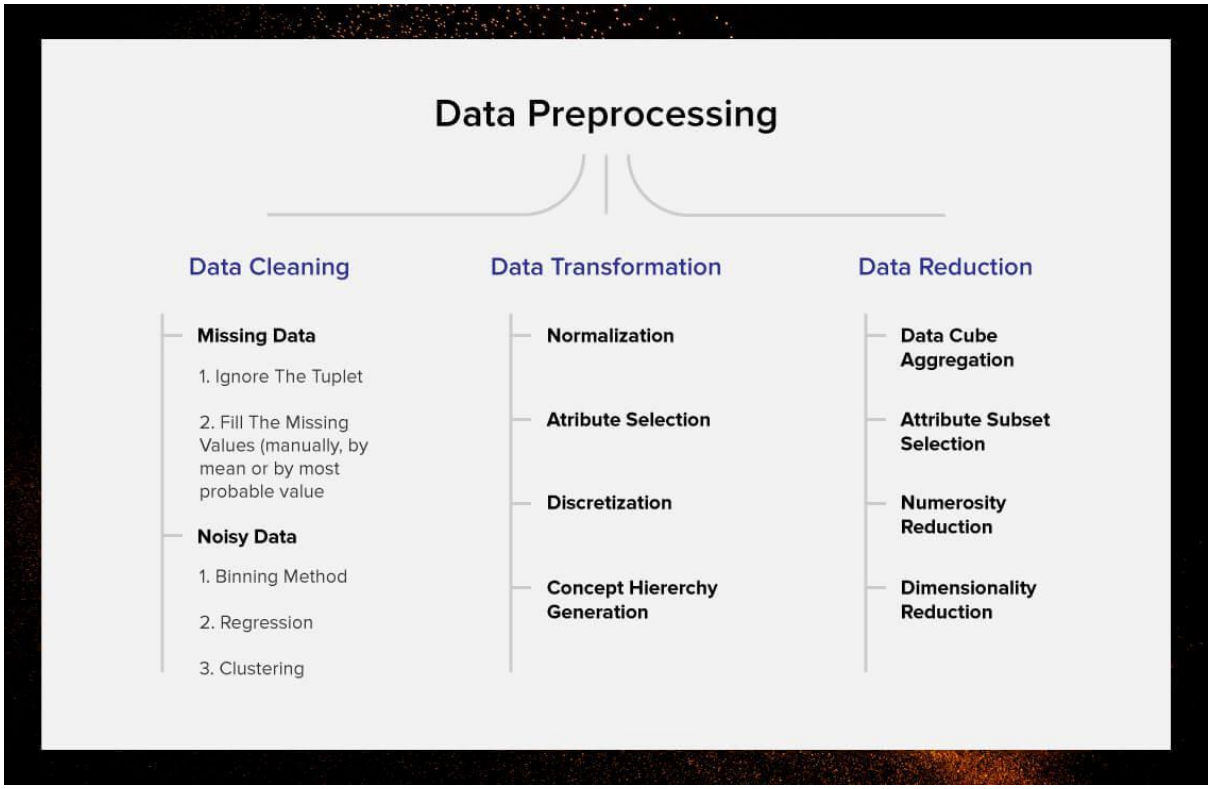
Pandas

Pandas is a popular open-source data analysis and manipulation library for Python. It provides easy-to-use data structures, such as DataFrame and Series, to efficiently handle and manipulate structured data.

Pandas simplifies tasks like cleaning, transforming, and analyzing data, making it a fundamental tool in data science and data analysis workflows.



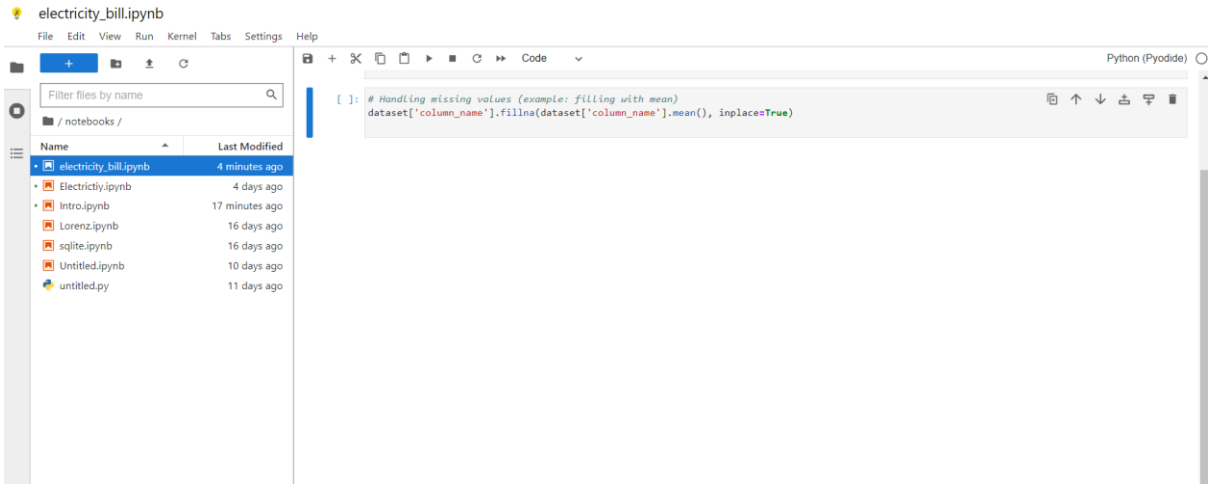
2. Data Preprocessing:

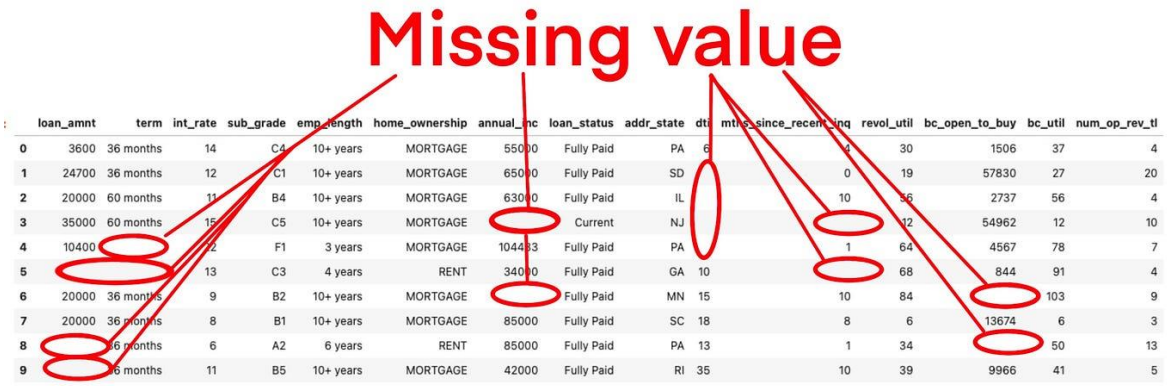


1. Data Cleaning:

➢ Handling Missing Values:

Identify and handle missing data points in the dataset. This can involve removing rows with missing values or filling in missing values using imputation techniques (mean, median, mode).





➢ Dealing with Duplicates:

Identify and remove duplicate records from the dataset to maintain data integrity.

➢ Correcting Inconsistencies:

Address any inconsistencies in data representation, such as typos, different spellings, or variations in categorical values.

➢ Outlier Detection and Removal:

Identify outliers using statistical methods (like Z-score) and remove or adjust them if they significantly affect the analysis.

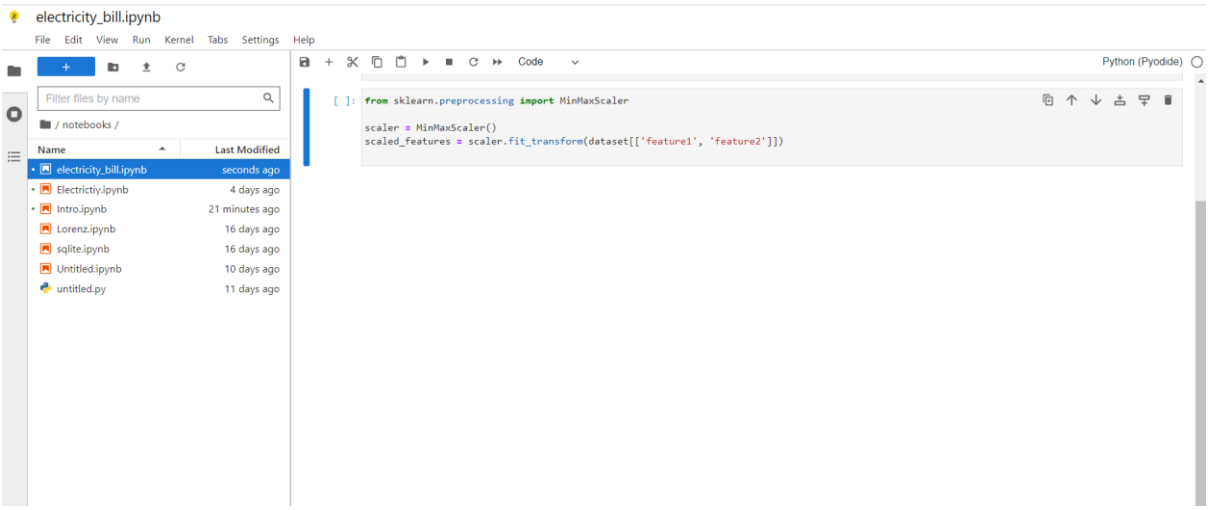
2. Data Transformation:

➢ Handling Categorical Data:

Convert categorical variables into numerical representations using techniques like one-hot encoding (for nominal data) or label encoding (for ordinal data).

➢ Feature Scaling:

Scale numerical features to ensure all features contribute equally to the analysis. Common methods include Min-Max scaling and Standardization (Z-score normalization).



➢ Feature Engineering:

Create new features from existing ones to capture relevant information. This can involve mathematical transformations, interaction terms, or domain-specific transformations.

- Datetime Conversion:

If your dataset contains date and time information, convert them into a usable format. Extracting features like day, month, or year can be valuable.

3. Data Integration:

- Merge or Join Data:

If your data is spread across multiple sources, merge or join datasets based on common identifiers to create a comprehensive dataset for analysis.

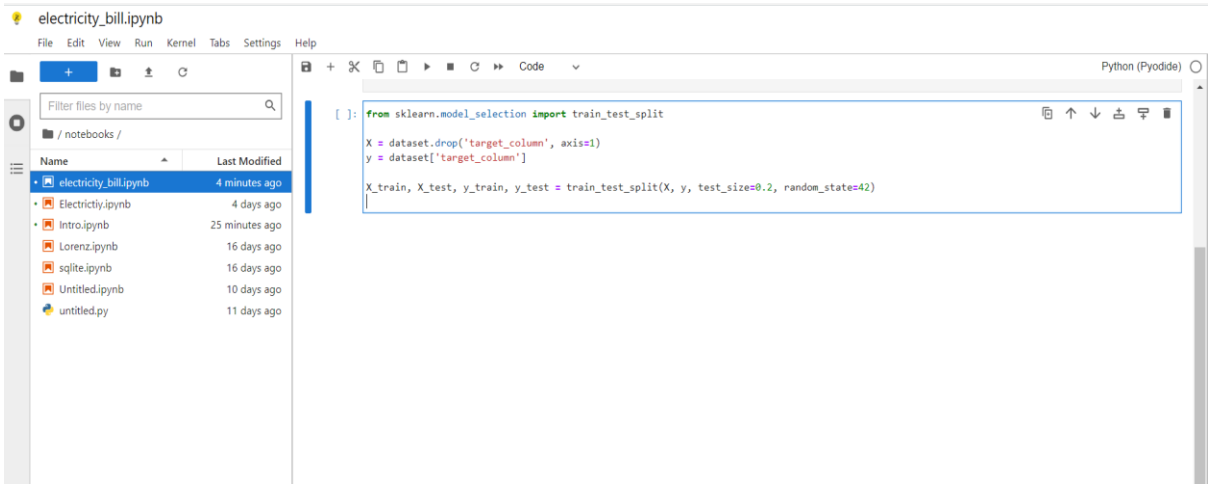
- External Data Integration:

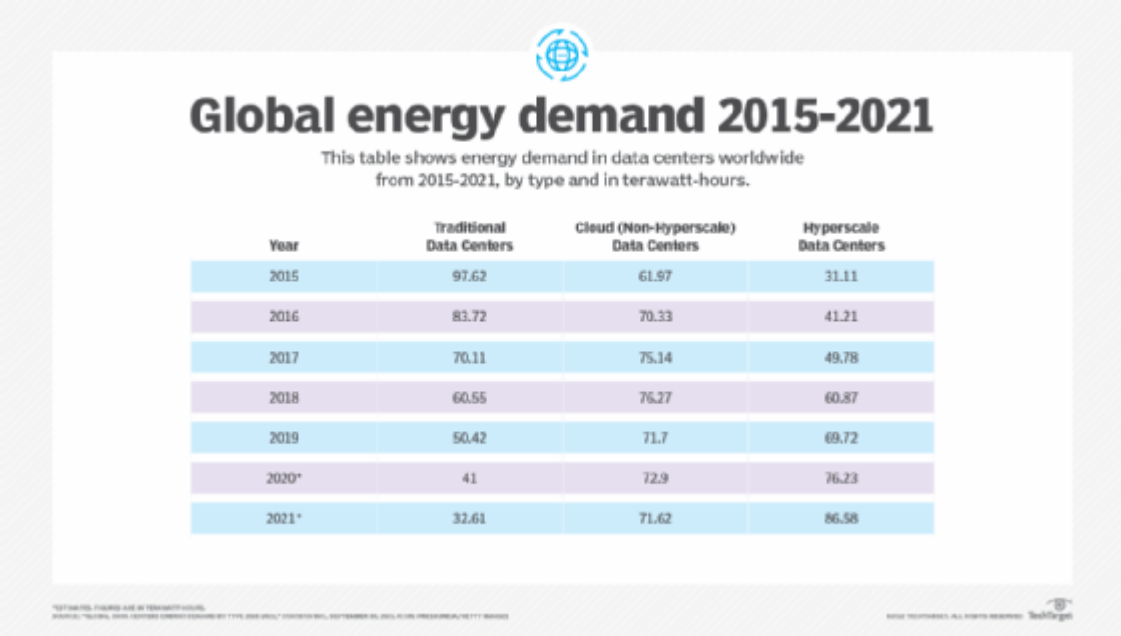
Integrate external datasets if they provide additional context or features that can enhance the analysis.

4. Data Organization:

Data Splitting:

Split the dataset into training and testing sets for model development and evaluation.





Global energy demand 2015-2021

This table shows energy demand in data centers worldwide from 2015-2021, by type and in terawatt-hours.

Year	Traditional Data Centers	Cloud (Non-Hyperscale) Data Centers	Hyperscale Data Centers
2015	97.62	61.97	31.11
2016	83.72	70.33	41.21
2017	70.11	75.14	49.78
2018	60.55	76.27	60.87
2019	50.42	71.7	69.72
2020*	41	72.9	76.23
2021*	32.61	71.62	86.58

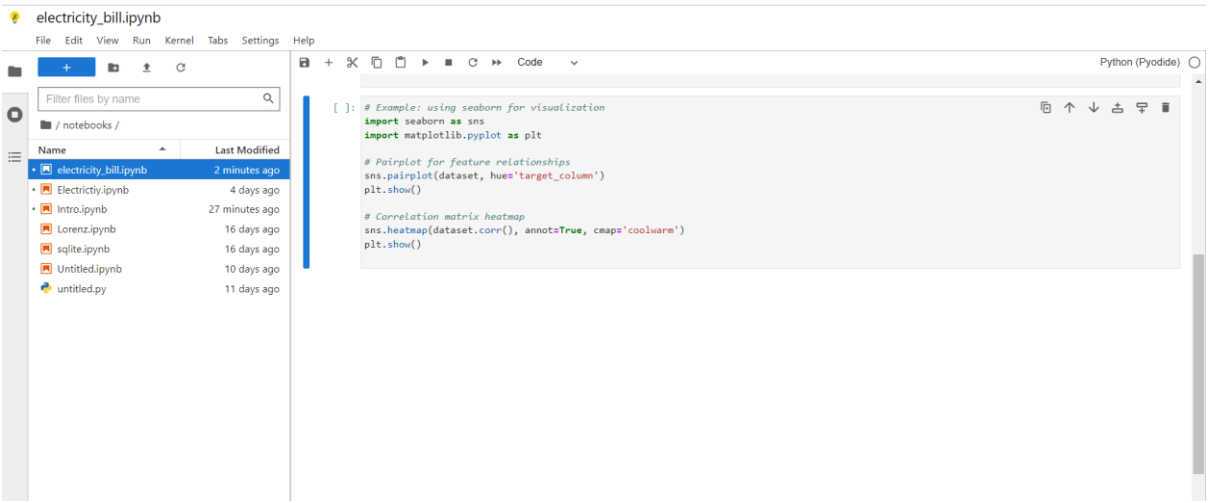
Data Formatting:

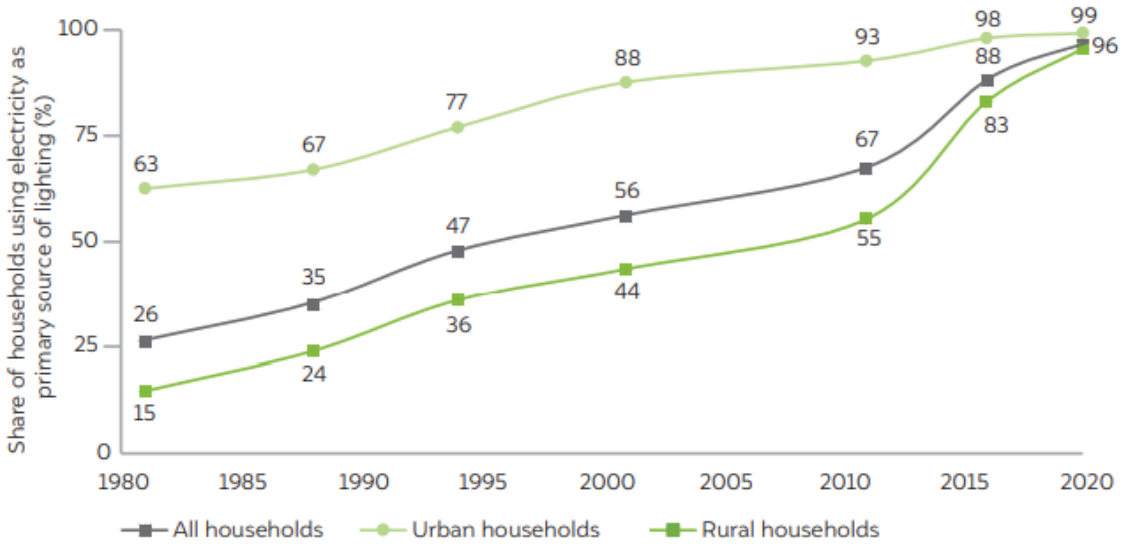
Ensure the final dataset is formatted according to the requirements of the algorithms or tools you'll be using for analysis and modeling.

3. Data Exploration :

Conduct exploratory data analysis (EDA) to gain insights into the dataset, using visualizations and statistical methods.

Explore relationships between features, identify patterns, and detect outliers





4. Saving Processed Data :

I have made significant changes during preprocessing, consider saving the processed data for future use.