

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.

electricity_bill.ipynb

File Edit View Run Kernel Tabs Settings Help

Python (Pyodide)

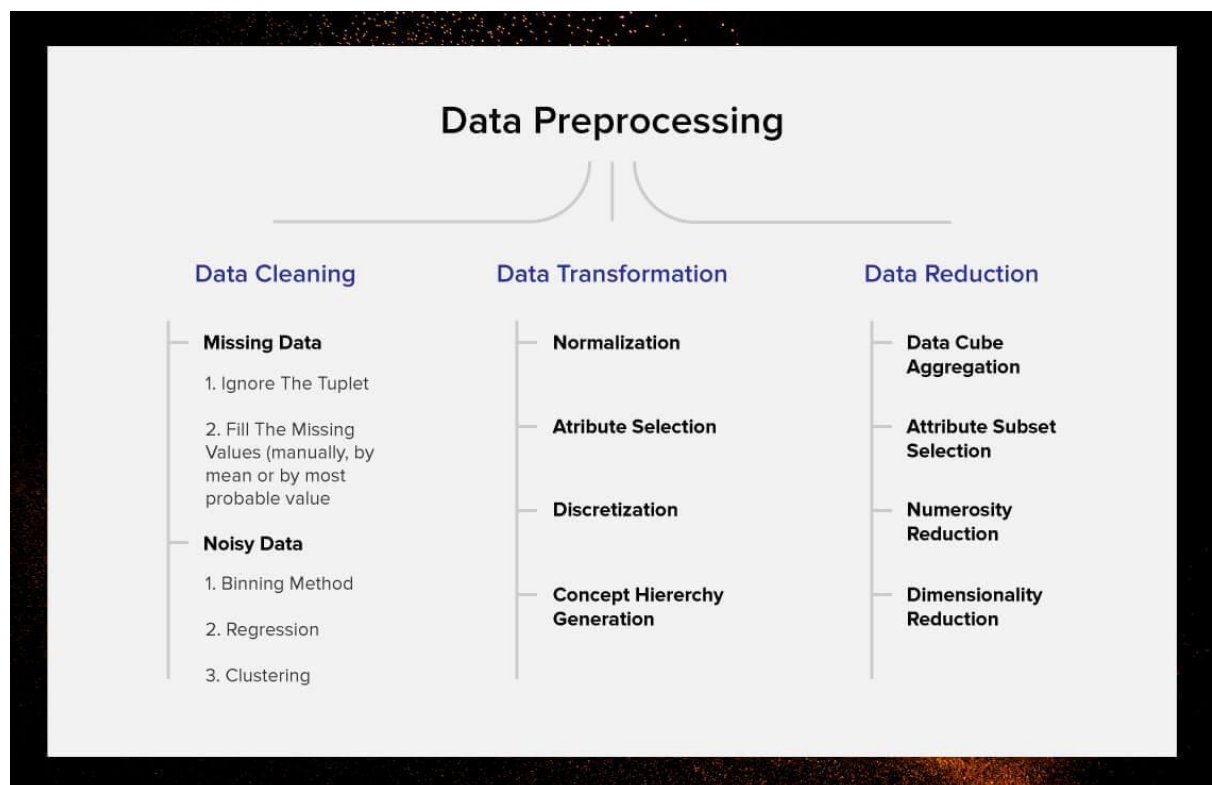
Filter files by name

/ notebooks /

Name	Last Modified
electricity_bill.ipynb	seconds ago
Electricity.ipynb	4 days ago
Intro.ipynb	11 minutes ago
Lorenz.ipynb	16 days ago
sqlite.ipynb	16 days ago
Untitled.ipynb	10 days ago
untitled.py	11 days ago

```
[ ]: import pandas as pd
# Load the dataset
dataset = pd.read_csv('your_dataset.csv')
# Display the first few rows of the dataset
print(dataset.head())
```

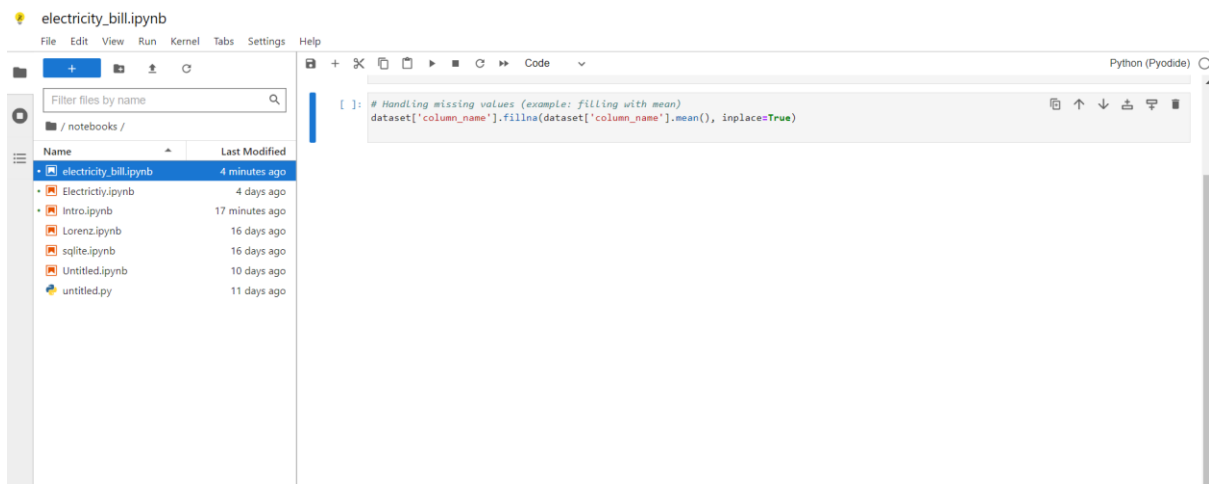
2. Data Preprocessing:



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).



Missing value

	loan_amnt	term	int_rate	sub_grade	emp_length	home_ownership	annual_inc	loan_status	addr_state	dti	months_since_recent_inq	revol_util	bc_open_to_buy	bc_util	num_op_rev_tl
0	3600	36 months	14	C4	10+ years	MORTGAGE	55000	Fully Paid	PA	6	2	30	1506	37	4
1	24700	36 months	12	C1	10+ years	MORTGAGE	65000	Fully Paid	SD	0	0	19	57830	27	20
2	20000	60 months	11	B4	10+ years	MORTGAGE	63000	Fully Paid	IL	10	10	36	2737	56	4
3	35000	60 months	15	C5	10+ years	MORTGAGE	Current	NJ	0	12	54962	12	10		
4	10400	36 months	12	F1	3 years	MORTGAGE	104433	Fully Paid	PA	64	1	64	4567	78	7
5	20000	36 months	13	C3	4 years	RENT	34000	Fully Paid	GA	10	68	844	91	4	
6	20000	36 months	9	B2	10+ years	MORTGAGE	85000	Fully Paid	MN	15	10	84	13674	103	9
7	20000	36 months	8	B1	10+ years	MORTGAGE	85000	Fully Paid	SC	18	8	6	50	13	3
8	16000	36 months	6	A2	6 years	RENT	85000	Fully Paid	PA	13	1	34	9966	41	5
9	42000	36 months	11	B5	10+ years	MORTGAGE	42000	Fully Paid	RI	35	10	39			

➤ 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.

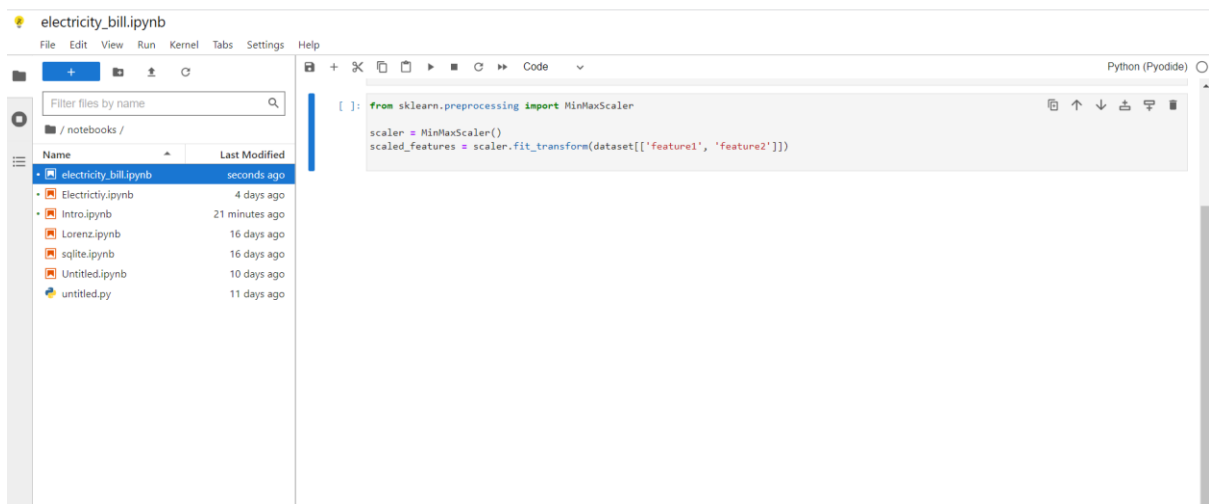
2. Data Transformation:

➤ 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.

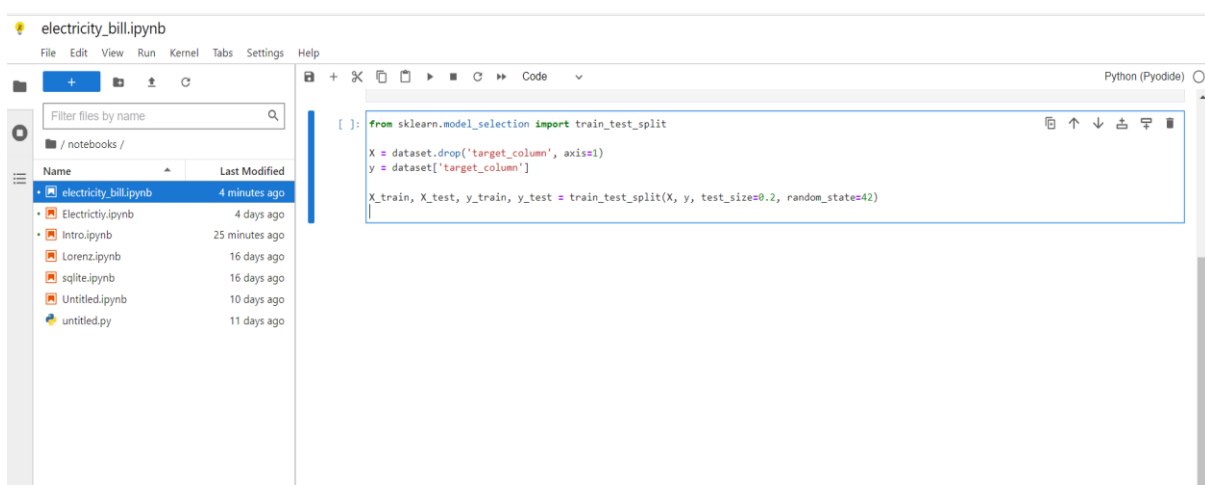
➤ External Data Integration:

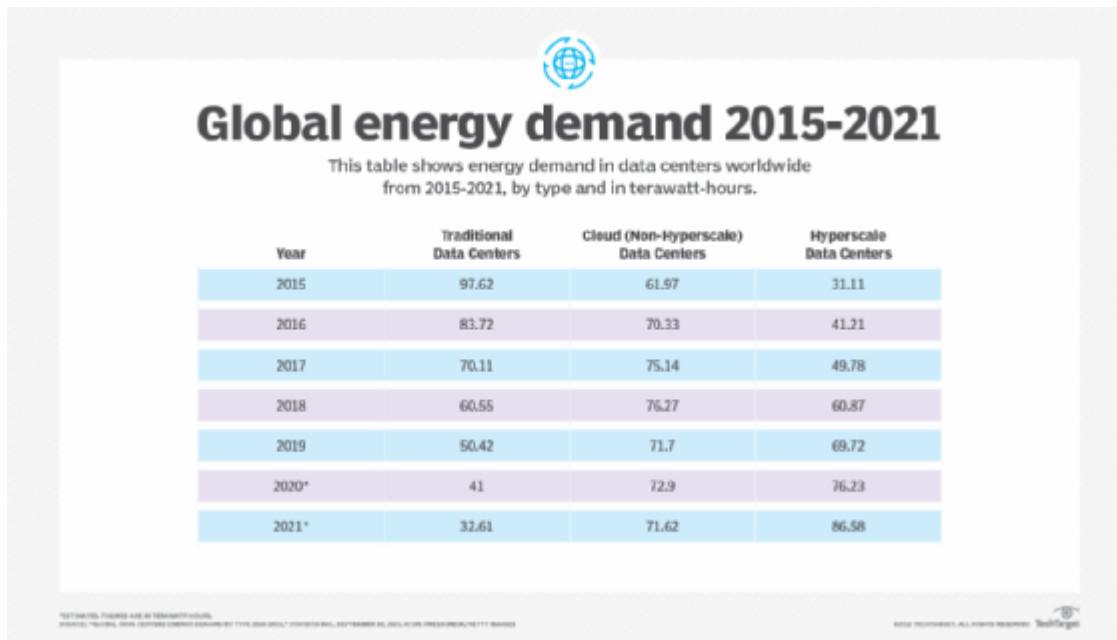
Integrate external datasets if they provide additional context or features that can enhance the analysis.

4. Data Organization:

Data Splitting:

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





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

electricity_bill.ipynb

File Edit View Run Kernel Tabs Settings Help

Python (Pyodide)

Filter files by name

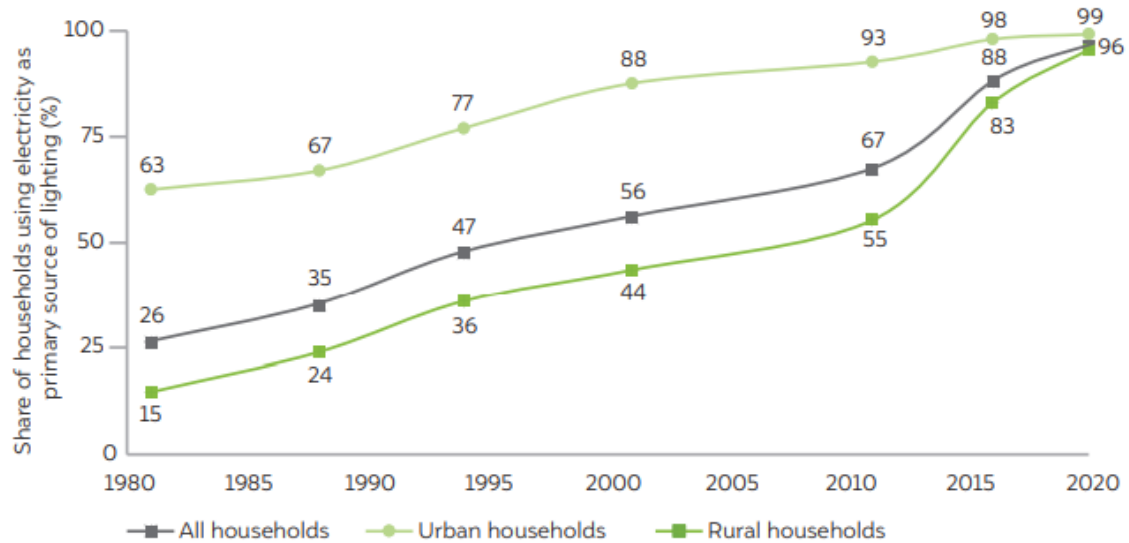
/ notebooks /

Name	Last Modified
electricity_bill.ipynb	2 minutes ago
Electricity.ipynb	4 days ago
Intro.ipynb	27 minutes ago
Lorenz.ipynb	16 days ago
sqlite.ipynb	16 days ago
Untitled.ipynb	10 days ago
untitled.py	11 days ago

```
[ ]: # Example: using seaborn for visualization
import seaborn as sns
import matplotlib.pyplot as plt

# Pairplot for feature relationships
sns.pairplot(dataset, hue='target_column')
plt.show()

# Correlation matrix heatmap
sns.heatmap(dataset.corr(), annot=True, cmap='coolwarm')
plt.show()
```



4. Saving Processed Data :

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

