PREPROCESSING THE DATASET

PREPROCESSING OF ENERGY CONSUMPTION DATASETS

INTRODUCTION:

This document outlines the preprocessing steps undertaken to prepare a comprehensive dataset for the analysis of energy consumption. The dataset is derived from multiple sources, and this document provides a detailed description of the data integration and preprocessing procedures.

DATA SOURCES:

The dataset comprises information from the following sources:

- AEP_hourly.csv
- COMED_hourly.csv
- DAYTON_hourly.csv
- DEOK_hourly.csv
- DOM_hourly.csv
- DUQ_hourly.csv
- EKPC_hourly.csv
- FE_hourly.csv
- NI_hourly.csv
- pjm_hourly_est.csv
- PJM_Load_hourly.csv
- PJME_hourly.csv
- PJMW_hourly.csv

DATA INTEGRATION:

The initial step involved importing data from each source using the Pandas library in Python. The data files were read and stored in separate DataFrames. These DataFrames were then merged horizontally (column-wise) to create a consolidated dataset, ensuring that duplicate columns were removed to avoid redundancy.

DATA PREPROCESSING

THE PREPROCESSING OF THE DATASET INVOLVED SEVERAL ESSENTIAL TASKS:

1.HANDLING MISSING VALUES:

Missing values, if any, were addressed by using appropriate techniques such as interpolation, forward-fill, or backward-fill, to ensure a complete dataset.

2. FEATURE ENGINEERING:

Additional features were created to enhance the dataset's predictive power. This may include transformations, scaling, or the creation of new derived features

.

3.DATA TYPE CONVERSION:

Data types were checked and modified to ensure consistency. In particular, non-numeric data types were converted to numerical types to make them compatible with machine learning algorithms.

PROGRAM:

```
import pandas as pd

dataset_filenames = [ 'AEP_hourly.csv', 'COMED_hourly.csv', 'DAYTON_hourly.csv', 'DEOK_hourly.csv',

'DOM_hourly.csv','DUQ_hourly.csv','EKPC_hourly.csv','FE_hourly.csv','NI_hourly.csv',

'pjm_hourly_est.csv','PJM_Load_hourly.csv', 'PJME_hourly.csv','PJMW_hourly.csv']

datasets = []

for dataset_filename in dataset_filenames:

    data = pd.read_csv(dataset_filename)

    datasets.append(data)

merged_dataset = pd.concat(datasets, axis=1)

merged_dataset = merged_dataset.loc[:, ~merged_dataset.columns.duplicated()]

print(merged_dataset)
```

OUTPUT:

```
              Datetime   AEP_MW  COMED_MW  DAYTON_MW  DEOK_MW  DOM_MW  \
0      2004-12-31 01:00:00  13478.0    9970.0     1596.0   2945.0  9389.0
1      2004-12-31 02:00:00  12865.0    9428.0     1517.0   2868.0  9070.0
2      2004-12-31 03:00:00  12577.0    9059.0     1486.0   2812.0  9001.0
3      2004-12-31 04:00:00  12517.0    8817.0     1469.0   2812.0  9042.0
4      2004-12-31 05:00:00  12670.0    8743.0     1472.0   2860.0  9132.0
...                ...      ...       ...        ...      ...     ...
178257             NaN      NaN       NaN        NaN      NaN     NaN
178258             NaN      NaN       NaN        NaN      NaN     NaN
178259             NaN      NaN       NaN        NaN      NaN     NaN
178260             NaN      NaN       NaN        NaN      NaN     NaN
178261             NaN      NaN       NaN        NaN      NaN     NaN

       DUQ_MW  EKPC_MW   FE_MW   NI_MW  ...    DUQ    EKPC      FE  NI  \
0      1458.0   1861.0  6222.0  9810.0  ...    NaN     NaN     NaN NaN
1      1377.0   1835.0  5973.0  9001.0  ...    NaN     NaN     NaN NaN
2      1351.0   1841.0  5778.0  8509.0  ...    NaN     NaN     NaN NaN
3      1336.0   1872.0  5707.0  8278.0  ...    NaN     NaN     NaN NaN
4      1356.0   1934.0  5691.0  8089.0  ...    NaN     NaN     NaN NaN
...       ...      ...     ...     ... ...     ...     ...     ... ..
178257    NaN      NaN     NaN     NaN  ... 1962.0  2866.0  9378.0 NaN
178258    NaN      NaN     NaN     NaN  ... 1940.0  2846.0  9255.0 NaN
178259    NaN      NaN     NaN     NaN  ... 1891.0  2883.0  9044.0 NaN
178260    NaN      NaN     NaN     NaN  ... 1820.0  2880.0  8676.0 NaN
178261    NaN      NaN     NaN     NaN  ... 1721.0  2846.0  8393.0 NaN

        PJME    PJMW  PJM_Load  PJM_Load_MW  PJME_MW  PJMW_MW
0        NaN     NaN   29309.0      29309.0  26498.0   5077.0
1        NaN     NaN   28236.0      28236.0  25147.0   4939.0
```

```
2          NaN     NaN   27692.0     27692.0  24574.0   4885.0
3          NaN     NaN   27596.0     27596.0  24393.0   4857.0
4          NaN     NaN   27888.0     27888.0  24860.0   4930.0
...        ...     ...       ...         ...      ...      ...
178257  44284.0  8401.0      NaN         NaN      NaN      NaN
178258  43751.0  8373.0      NaN         NaN      NaN      NaN
178259  42402.0  8238.0      NaN         NaN      NaN      NaN
178260  40164.0  7958.0      NaN         NaN      NaN      NaN
178261  38608.0  7691.0      NaN         NaN      NaN      NaN

[178262 rows x 25 columns]
```

RESULT:

The result is a well-structured, cleaned, and consolidated dataset, free of duplicate columns and missing values. This dataset is now ready for further analysis, modelling , and evaluation of energy consumption patterns.

CONCLUSION:

The successful preprocessing of the dataset is a critical step in any data analysis project. By combining data from multiple sources and ensuring data quality, we have created a solid foundation for future research and analysis in the domain of energy consumption.