TOPIC: PREDICTING ENERGY CONSUMPTION

The primary goal of this project is to develop a model that can accurately forecast future energy consumption based on historical consumption data. This model will help energy providers manage supply, optimize grid operations. and plan for future energy needs. I have downloaded the dataset from **kaggle. The** name of the dataset is energy consumption, the dataset has metadata explaining what is all about the dataset, the dataset had 2160 columns and 15 rows.

Loading the dataset

I have loaded the dataset using the pandas package that has the method pd.read_csv() used to read data that are in comma separated value. After loading the dataset I have checked the format of it by viewing five rows using head() function and the dataset is in wide format since every column represents a unique observation. Then I checked the number of rows and number of columns using the shape attribute of pandas. Then after I checked the data types of each column using the info() function of pandas, the dataset has integer data types, string data types and float data types.

Checking for missing values and filling them

I used the isnull() function and sum() function that return the number of missing values in each column and I used fillna() and the mean of each column to fill in the missing values.

Dropping other columns that were not needed

I have dropped a column named "CONSUMER_TYPE" from the dataset, this column will not affect anything if it is there in the dataset or not there in the dataset.

Removing outliers

Firstly i have plot the graphs of numerical column to see if there are outliers, the boxplot is the best at showing outliers and spread of dataset

I have removing outliers using scatter plot technique here I have calculated interquartile range using code and also I calculated lower bound and upper bound and here is I have removed outliers, using code I checked every number in column if it is greater than upper bound i have removed that number and when a number is less than lower bound I have also removed it. Outlines prevent deep insights of the dataset. After removing outliers I plotted graph again to show if outliers are still exit but the plotted boxplots are not showing any outliers

Dealing with duplicate rows that are multiple in the dataset

I have checked for duplicate rows using duplicated() function in pandas and after finding row that are similar in the dataset I removed them using drop_duplicates() function in pandas

Transforming dataset

I have changed all the numerical columns to have a mean of 0 and standard deviation of 1 using z-score normalization. To transform numerical columns I used standardscaler class from sklearn package I had to import it to use

Creating new variables

I have introduced two new variables in the dataset using other existing columns in the dataset here are new introduced variables:

1. **Total Energy Consumption for the Year (TOTAL_CONSUMPTION)**
   This variable will be the sum of the monthly consumption values for each row.

2. **Average Monthly Consumption (AVERAGE_CONSUMPTION)**
   This variable will be the average of the monthly consumption values for each row

Encoding categorical variables

This is changing variable values or string to represent numbers this increase generalization of model, I have changed the column called “states into numbers.

Exploratory data analysis

1. **Finding summary statistics**

   This is finding mean, average, minimum value, maximum value, standard deviation, 25 percent, and 75 percent of the dataset. To do this I use describe() function in pandas

2. **Data visualization**

   - **Histograms**

     Histograms are useful for understanding the distribution of individual column

     Histograms are showing energy consumption in each month and it is showing that in november and december energy has been consumed than other months

   - **Boxplot**

     It is used to show spread or skewness of the column and it is also used to show if a column has outliers**.** After viewing boxplot it shows that all the columns of months are showing long whiskers in the left side this mean that there are some months where small days had exceptionally low energy consumption compared to the rest of the days in a month in column

   - **Correlation matrix**

     It helps to understand relationship between numerical columns by showing numbers between -1 ,0 to +1

     When it shows -1 it means there's a perfect negative correlation between those two columns, when it shows 0 it means there is no correlation and when it shows +1 it means there is perfect positive correlation of the column. From the graph number it is showing positive correlation because every shown number from the graph is greater than 0 so it just indicates that energy consumption in any two months tend to increase or decrease together. All the correlation is greater than zero suggesting energy consumption in any month is highly associated with high consumption in other months

- **Pair plots**

Pair plots provide scatter plots for each pair of numerical columns, useful for identifying relationships and patterns of numerical columns. From the plotted graph it is showing points forming upward sloping line for any column pair meaning positive correlation for any columns pair and if point would form downward sloping line I would say it's negative correlation and if also points would be in randomly scattered I would say there's no any relationship between columns

All graphs in pair plot are showing points in upward direction mean all column are positive correlated

Saving transformed data

I have saved the transformed data into csv using to_csv() function in pandas used to save file opened in pandas to comma separated values

Challenges I faced when doing this project

1. Energy consumption data was missing
2. Some graphs was difficult to interpret and finding relationship between columns
3. The dataset I found is too small I needed big dataset but it was not found