Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	Team-739770
Project Title	Predicting the energy output of wind turbine based on weather condition
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description

Data Overview	Basic statistics, dimensions, and structure of the data.
	
	`<class 'pandas.core.frame.DataFrame'>`
	`RangeIndex: 4447 entries, 0 to 4446`
	`Data columns (total 6 columns):`
	` #  Column  Non-Null Count  Dtype`
	`---  ------  --------------  -----`
	` 0  Wind Speed (m/s)  4447 non-null  float64`
	` 1  Wind Direction  4447 non-null  float64`
	` 2  maxtempC  4447 non-null  int64`
	` 3  humidity  4447 non-null  float64`
	` 4  pressure  4447 non-null  float64`
	` 5  Output_Energy  4447 non-null  float64`
	`dtypes: float64(5), int64(1)`

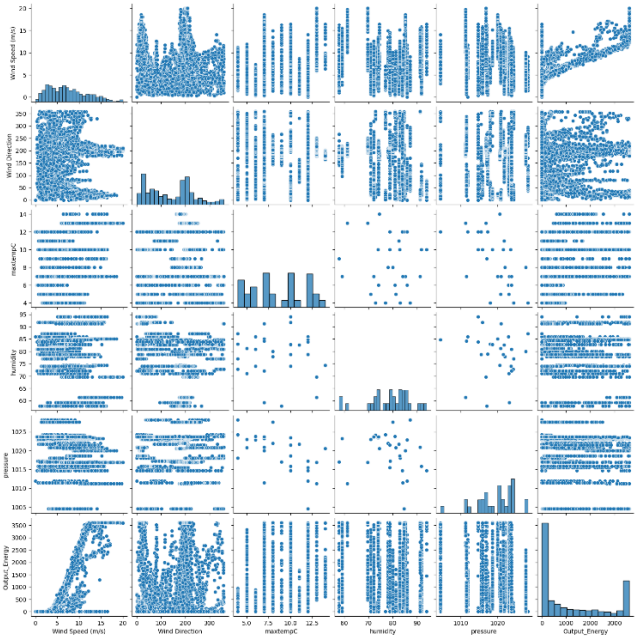
Univariate Analysis	Exploration of individual variables (mean, median, mode, etc.).

		Wind Speed (m/s)	**Wind Direction**	**maxtempC**	**humidity**	**pressure**	**Output_Energ**
	count	4447.000000	4447.000000	4447.000000	4447.000000	4447.00000
	mean	7.357389	140.667803	8.535192	78.648874	1019.49165
	std	4.361162	93.616266	3.034301	9.004574	5.15432
	min	0.000000	0.000000	4.000000	54.125000	1004.54166
	25%	3.669025	53.272396	6.000000	74.000000	1015.87500
	50%	6.717962	143.424896	8.000000	80.041667	1020.83333
	75%	10.197950	206.816154	12.000000	84.708333	1023.45833
	max	21.621000	359.942291	14.000000	93.958333	1028.20833
Bivariate Analysis	Relationships between two variables (correlation, scatter plots).					
		Wind Speed (m/s)	**Wind Direction**	**maxtempC**	**humidity**	**pressu**
	Wind Speed (m/s)	1.000000	0.017336	0.339107	0.15185	
	Wind Direction	0.017336	1.000000	0.080762	0.31354	
	maxtempC	0.339107	0.080762	1.000000	0.06532	

	humidity	-0.151853	-0.313542	-0.065329	1.000000	-0.12
	pressure	-0.234967	-0.020962	-0.597324	-0.129295	1.00
	Output_Energy	0.882457	0.122913	0.403382	-0.251067	-0.24

Multivariate Analysis

Patterns and relationships involving multiple variables.



Outliers and Anomalies

```
for col in df.columns:
    q1 = df[col].quantile(0.25)
    q3 = df[col].quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    df[col]=np.where(df[col]<lower_bound,lower_bound,df[col])
    df[col]=np.where(df[col]>upper_bound,upper_bound,df[col])

for col in df.columns:
    sns.boxplot(df[col])
    plt.show()
```

Data Preprocessing Code Screenshots

Loading Data	`data = pd.read_csv('/content/data.csv')` `target = pd.read_csv('/content/target.csv')`

Handling Missing Data	`Data columns (total 6 columns):`
	` #  Column  Non-Null Count  Dtype`
	`---  ------  --------------  -----`
	` 0  Wind Speed (m/s)  4447 non-null  float64`
	` 1  Wind Direction  4447 non-null  float64`
	` 2  maxtempC  4447 non-null  int64`
	` 3  humidity  4447 non-null  float64`
	` 4  pressure  4447 non-null  float64`
	` 5  Output_Energy  4447 non-null  float64`
	`dtypes: float64(5), int64(1)`
	`memory usage: 208.6 KB`

Data Transformation	`Scaler = StandardScaler()`
	`for col in df.columns:`
	`  if col != 'Output_Energy':`
	`  df[col] = Scaler.fit_transform(df[[col]])`
	
	`df.head()`

Feature Engineering	Code for creating new features or modifying existing ones.
Save Processed Data	Code to save the cleaned and processed data for future use. df = data