**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 15 March 2024 |
| Team ID | Team-740034 |
| 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\_Energy** | | --- | --- | --- | --- | --- | --- | | **count** | 4447.000000 | 4447.000000 | 4447.000000 | 4447.000000 | 4447.000000 | 4447.000000 | | **mean** | 7.357389 | 140.667803 | 8.535192 | 78.648874 | 1019.491652 | 1176.785881 | | **std** | 4.361162 | 93.616266 | 3.034301 | 9.004574 | 5.154328 | 1374.223232 | | **min** | 0.000000 | 0.000000 | 4.000000 | 54.125000 | 1004.541667 | -1.077131 | | **25%** | 3.669025 | 53.272396 | 6.000000 | 74.000000 | 1015.875000 | 0.000000 | | **50%** | 6.717962 | 143.424896 | 8.000000 | 80.041667 | 1020.833333 | 482.019714 | | **75%** | 10.197950 | 206.816154 | 12.000000 | 84.708333 | 1023.458333 | 2289.062988 | | **max** | 21.621000 | 359.942291 | 14.000000 | 93.958333 | 1028.208333 | 3602.782959 | |
| Bivariate Analysis | Relationships between two variables (correlation, scatter plots).   | **Wind Speed (m/s)** | **Wind Direction** | **maxtempC** | **humidity** | **pressure** | **Output\_Energy** | | --- | --- | --- | --- | --- | --- | | **Wind Speed (m/s)** | 1.000000 | 0.017336 | 0.339107 | -0.151853 | -0.234967 | 0.882457 | | **Wind Direction** | 0.017336 | 1.000000 | 0.080762 | -0.313542 | -0.020962 | 0.122913 | | **maxtempC** | 0.339107 | 0.080762 | 1.000000 | -0.065329 | -0.597324 | 0.403382 | | **humidity** | -0.151853 | -0.313542 | -0.065329 | 1.000000 | -0.129295 | -0.251067 | | **pressure** | -0.234967 | -0.020962 | -0.597324 | -0.129295 | 1.000000 | -0.249726 | | **Output\_Energy** | 0.882457 | 0.122913 | 0.403382 | -0.251067 | -0.249726 | 1.000000 | |
| Multivariate Analysis | Patterns and relationships involving multiple variables. |
| Outliers and Anomalies |  |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |
| Handling Missing Data |  |
| Data Transformation |  |
| 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 |