Explanation of the solution and python code

1. **Load the data using the data type pandas DataFrame**
    a. Downloaded data by changing the separator to semicolon, data was irregularly separated if downloaded with commas.
    b. Columns names and unit names contained many empty places. So, removed those and added units into the column names so that only the data can be handled.
    c. Replaced all the ‘,’ by ‘.’ in the values and converted them to float.
    d. Converted date time to standard datetime format.
2. **Train any ML Model that predicts the power (single target) by a given windspeed (single feature)**
    a. Windows of 6 hours data used to predict immediately next time step.
    b. Windows shuffled before training to avoid any bias related to time of the year.
    c. Data was split as follows –-> 70% training, 20% validation, 10% testing (**Turbine 1**)
    d. Data for both turbines was studies and it was observed that both turbines have similar data. Hence, turbine 1 was used for training and turbine 2 was used to check model performance.
    e. Data was standardized using the mean and standard deviation values from training dataset
    f. LSTM based model was used.
3. **Choose a good metric to measure the model performance**
    a. Mean Absolute Percent Error (MAPE) was used to measure model performance.
    b. As the task is of regression type, to measure the model performance the error between predicted and actual value must be calculated. MAPE provides that error.
    c. As it is percentage of actual value it provides some reference to analyze error value.
4. **Explain why you chose this model architecture and what the limitations of this architecture might be**
    a. LSTM based models learn long-term dependencies in the data better than the normal RNN models and hence, LSTM is normally used for time dependent data.
    b. Two three model architectures were tried out with one/two/three LSTM layers and a dense layer.
    c. LSTMs requires large amount of data for training and do not perform well with non-stationary data. So, this might cause the deterioration in performance as the season changes.
5. **Try out additional features. Which features did you choose? How did the results change and why?**
    a. All the KH features have 0 values so removed them.
    b. Logically time of the day and day of the year should also have an effect of wind patterns so those were converted to sine curve to see their effect.
    c. Correlation plots and PCA was used to find relevant features.
    d. So, following features were considered – Gen1, Lager, Aussen, GetrT, and {PCA1, PCA2} →{Wind, Rotor, Strom, strom1, strom2}
    e. Logically day, time, Betrieb Stunden should also have effect on power but due to limitation of time they are not considered here and can be studied if time permits.
    f. Model performance was marginally better after considering other features. This is because, now model has much more relevant data to learn from.
    g. Only marginal improvement can be explained by model limitations or the possibility of not considering some relevant features due to time limitations for now.
6. **Based on this model, where would you suspect turbine anomalies? Please list time frames and visualize the anomalies.**
    a. To find anomalies standardized Euclidean distance between actual and predicted values is calculated and averaged over the window of 3 hours. Windows where standardized value is more than 3 is called anomalous.
    b. List of anomalies for turbine 2 (Considering 3 hours windows) is as follows
        i. 08/01/2016 – from 3.00 to 6.00 and from 9.00 to 15.00
        ii. 13/01/2016 – from 12.00 to 15.00
        iii. 26/01/2016 – from 21.00 to 24.00
        iv. 01/02/2016 – from 12.00 to 15.00
        v. 02/02/2016 – from 12.00 to 18.00
        vi. 07/02/2016 – from 15.00 to 18.00
        vii. 10/02/2016 – from 00.00 to 3.00
        viii. 21/02/2016 – from 3.00 to 6.00 and from18.00 to 21.00
        ix. 21/03/2016 – from 18.00 to 21.00
        x. 30/03/2016 – from 12.00 to 15.00
    c. Scatter plot of standardized error was plotted.

Theoretical Questions

1.
   a. In general LSTM, GRU, CNN, TCN based models can be used for time series forecasting.
   b. LSTM, GRU based models are good at finding long term patterns in the data and thus can be used for finding and learning long term seasonality in the data. For example, there can be yearly seasonality in the wind patterns.
   c. Performance of LSTM and GRU based models decreases as the sequence length increases and so attention layers can be used to improve performance of these models.
   d. CNN can be good at finding short term patterns in the data. TCNs are the modified version of 1D convolutional networks which increases its capacity of learning long term patterns. But, in general these networks can be used for learning short-term patterns like daily patterns or hourly patterns.
   e. All these architectures can be implemented in tensorflow with keras. Keras provides different methods to implement all of these as layers. They also have tutorials online which can be used while implementation.

2.
   a. Handling missing data can be very situation dependent. It also depends on for what duration the data is missing and which attributes are missing.
   b. If the data for short period is missing for the attributes where data remains close to mean, then it can be replaced by mean (e.g. data in Rotor(rpm) column).
   c. There are some columns where data can be predicted based on the available data at the boundaries of missing time interval. E.g. in case of BtrStd1, BtrStd2 columns (Considering it is Betrieb Stunden) values keep on increasing with time and so missing values can be guessed by looking at the values before and after missing value interval.
   d. But no technique can be commonly used for all the attributes as it will lead to wrong data.
   e. Also, as it is time series data completely removing the in between time steps should be avoided as far as possible. Because it will lead to loss of time dependencies. If it is necessary to remove some time steps in between then required actions must be taken to consider for the time dependencies.

3.
   a. Particularly, considering the case provided in question, there can be few causes behind this anomaly.
   b. Temperature in gear box can increase due to high speeds of rotor, reduced coolant pressure, mechanical problems like reduced performance of bearing, gears, or lubricants.
   c. To find the exact reason behind this anomaly, different parameters like coolant pressure, lubricant pressure, rotor speed, mechanical components operational hours can be separately monitored and if anomalous behavior is observed in any of these parameters, then the right cause can also be easily found out.
   d. However, this will require inputs from subject experts and specific solutions for each type of common anomalies must be designed.
   e. In general, the area of explainable AI deals with this type of use cases. So, this area can be researched.