ENGR 301
Engineering Applications of Data Science

Laboratory 1
Topic: Environmental Engineering; Subtopic: Wind Energy

Introduction: As the world looks to reduce CO_2 emissions and combat global warming, renewable sources are being turned to as alternatives to the burning of fossil fuels to provide energy. One major challenge in environmental engineering is that different regions require different renewable energy platforms depending on their weather and climate. One such promising technology in areas that can support it is wind energy, which, as may be expected, requires a lot of wind on a regular basis. The underpinning of wind energy technology is the use of wine turbines to convert the kinetic energy of wind to electrical energy *via* a generator; a simplified schematic of this is shown in Figure 1.

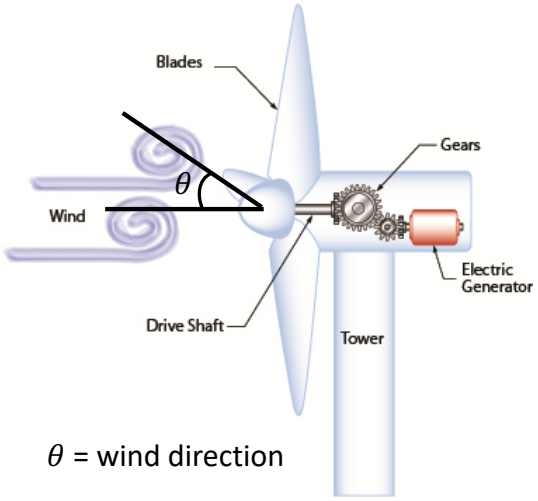

Figure 1. (left) Photograph of an offshore wind turbine. (right) Basic schematic of converting wind kinetic energy into electrical energy using a wind turbine.

Data is constantly collected from wind turbines and used to improve their capabilities. One such data set is provided here and consists of real measurements taken from a single wind turbine in Turkey over the course of a year. In this laboratory, you will use this data set and practice data management and cleaning, exploratory analysis, correlation between variables, and basic visualization.

Discipline Specific Information: The data set contains the following information:

Column Name	Description
Date/Time	Date and time each measurement was taken. Measurements were taken in 10 minute intervals.
Wind Speed (m/s)	The wind speed at the hub height of the turbine (units of meters per second).
Wind Direction (°)	The wind direction at the hub height of the turbine (units of degrees).
LVActivePower (kW)	The *actual* power generated by the turbine at the specified point in time under the given conditions (units of kilowatts).
Theoretical_Power_Curve (KWh)	The theoretical power the wind turbine *should* generate under the given conditions at each point in time (units of kilowatt hours).

ENGR 301
Engineering Applications of Data Science

Tasks

Note: For full credit, all plots must have axis labels and units on the x- and y-axes, and the discussion must be in complete and full sentences.

Part 1 – Basic Data Exploration

1. First, import the appropriate packages (numpy, pandas, and matplotlib) into a Jupyter notebook using the import ... as ... command functionality.

2. The data is provided in a CSV file called turbines.csv. Use *pandas* to import this data set into a Jupyter notebook as a dataframe (using df = read_csv(...)).

How many measurements are there in this data set?

3. Calling each column individually (i.e., df['Column Name'].command()), find the maximum value, minimum value, mean, median, and standard deviation of each variable (besides Observation Number and Date/Time).

In which features are the mean and median the same in which are they significantly different (more than 10%)?

By plotting each variable as a histogram, discuss why the mean and median are the same or different in terms of the shape of the data distribution. Next, use df.describe() and summarize what you see.

4. Create two boxplots: one for (a) wind speed and (b) actual power.

How does the shape of these boxplots compare to the shape of the histogram for these two variables?

5. Copy the code provided in the "month_mapping.ipynb" notebook into your notebook. This maps each observation onto a specific month and creates a new column called "Month".

Use the pandas groupby function to group by month and compute the mean of each variable.

Which months have the highest average wind speed and power output?

Part 2 – Correlation

1. Create two scatterplots: (a) one of the theoretical power versus wind speed and (b) one of the theoretical power versus wind direction.

What feature should be on the x-axis and y-axis and why (think about dependent/independent variables)?

What do these two plots tell you about power generation in a wind turbine?

2. Create two new scatterplots: (a) one of the actual power versus wind speed and (b) one of the actual power versus wind direction.

Compare these plots with the two from the last part.

What do these two plots tell you about power generation in a wind turbine?

3. Using a correlation table (do you remember how to do it in pandas?), describe the correlation between:

(a) wind speed and theoretical power curve
(b) wind direction and theoretical power curve
(c) wind speed and actual power curve
(d) wind direction and actual power curve

Do these observations agree with the four scatterplots you created previously? Why or why not?

4. How well correlated is the theoretical power and the actual power? Does this make sense to you? Why?

Part 3 – Feature Analysis and Interpretation

1. Create two new columns in your data frame: (a) one called "Loss" which contains the difference between the actual power and theoretical power and (b) one called "Loss percentage" which contains the percentage loss between the actual power and theoretical power. (Hint: how can you manipulate existing columns to get this information?).

What do these two new columns represent in terms of power produced by the wind turbine?

2. Create two scatterplots: (a) wind speed versus loss and (b) wind speed versus loss percentage.

How does the wind speed affect the loss?

3. What do you think is creating this loss? How could you overcome it?