# Economics 4803/8803: Machine Learning for Economics Problem Set 1

Due by: **Feb 5, 5 PM ET**

This problem set is designed to get you started with the dataset for the group project. A secondary objective is to get you more comfortable with analyzing and visualizing data with R.

You will work with three datasets - wind turbines, wind ordinances, and wind resource quality, clean it, and perform some simple data summary exercises. You are allowed to work in groups of up to four students, but you must disclose the members of your groups. Individual submissions are required. The code you submit may be iden- tical to the one of the other group members, but I expect comments and answers to the questions to be your own.

Your submission should consist of:

1. a pdf file with responses to the questions marked with * (please do not attach any code snippets)
2. a .R document with code
3. Please upload these separately on Canvas (please do not email me or the TA your home work submission).

Each of the following four sections below summarize the required tasks, with the individual lists detailing the steps needed to complete them.

# 1 Wind turbine Data

1. The United States Wind Turbine Database (USWTDB) provides the locations of land-based and offshore wind turbines in the United States, corresponding wind project information, and turbine technical specifications. Wind turbine records are collected and compiled from various public and private sources, digitized and position-verified from aerial imagery, and quality checked.
   Download the US Wind Turbine Database (USWTB) from the following link: https: //eerscmap.usgs.gov/uswtdb/data/
   - Download the 'Tabular Data: CSV format' dataset under **Raw Data & Meta- data Downloads** headline. This dataset is also on Canvas with the name uswtdb_v6_0_20230531.csv.
   - You should also download the full codebook (data dictionary) for your reference. The relevant file is 'Codebook V6 (2023-05-31) XLS' under Related files, here: https://emp.lbl.gov/publications/us-wind-turbine-database-files

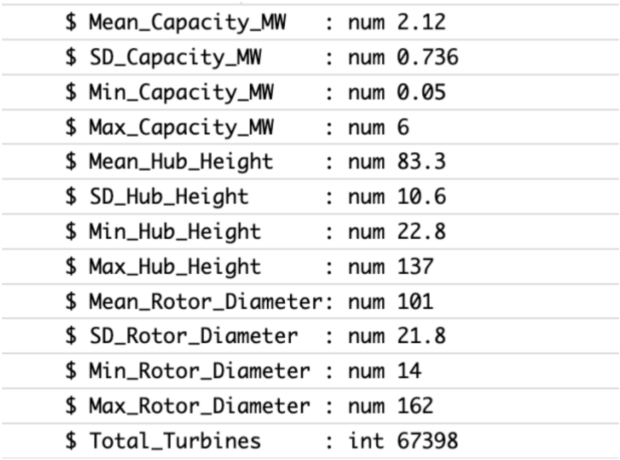
2. Bring this data into R. You can use read.csv().

3. Remove observations with missing information on operation year, project capacity, turbine capacity, turbine hub-height, turbine rotor diameter, longitude, and latitude

4. Remove observations for projects that started operating before 2001

5. Remove observations from Alaska and Hawaii (we would only focus on contiguous US).

6. Create a summary table and show the mean, standard deviation, min, and max of turbine capacity (in MW, not 1 MW = 1000 kW), turbine hub-height, turbine rotor diameter. Also report the total number of turbines (i.e. N) in the table? *

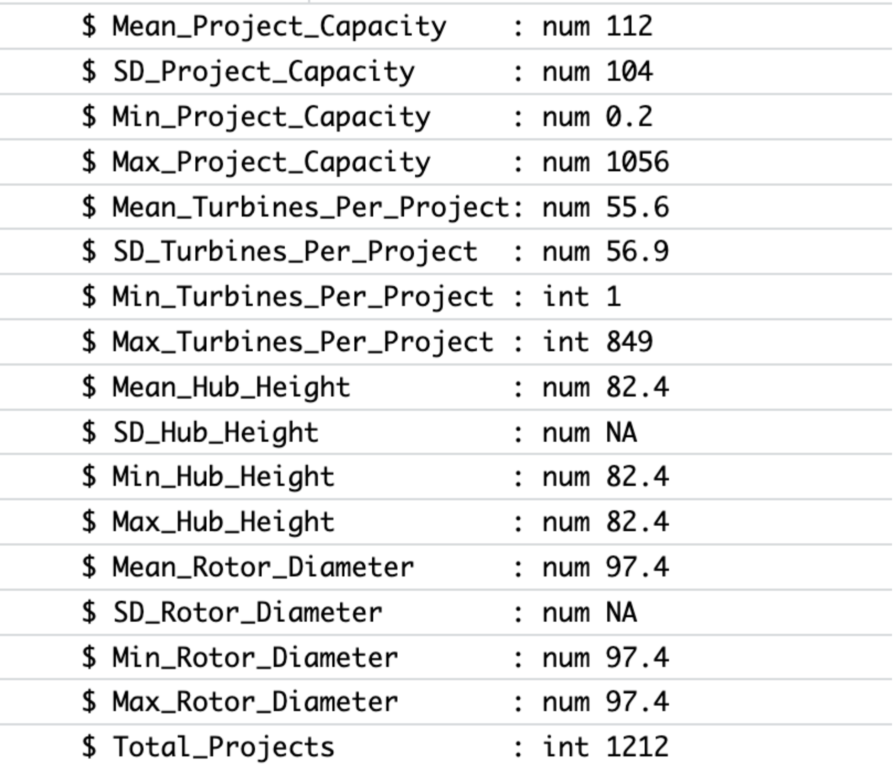


```
 $ Mean_Capacity_MW   : num 2.12
 $ SD_Capacity_MW     : num 0.736
 $ Min_Capacity_MW    : num 0.05
 $ Max_Capacity_MW    : num 6
 $ Mean_Hub_Height    : num 83.3
 $ SD_Hub_Height      : num 10.6
 $ Min_Hub_Height     : num 22.8
 $ Max_Hub_Height     : num 137
 $ Mean_Rotor_Diameter: num 101
 $ SD_Rotor_Diameter  : num 21.8
 $ Min_Rotor_Diameter : num 14
 $ Max_Rotor_Diameter : num 162
 $ Total_Turbines     : int 67398
```

7. Aggregate/collapse the data to the project level. Use EIA ID for the aggregation.

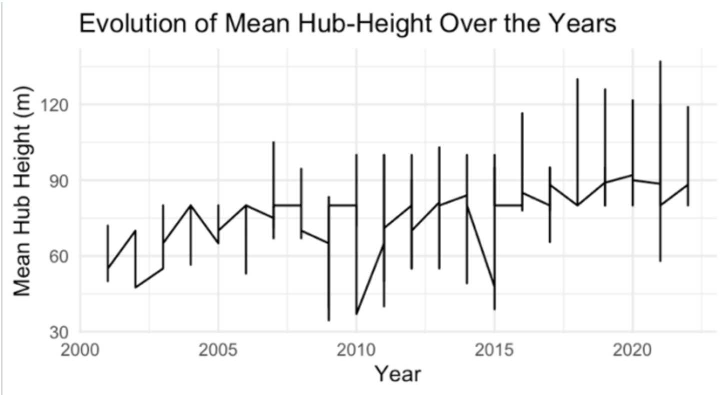
    - You can use several commands for aggregating data. The most common functions are: aggregate()or summarize()in package dplyr.[1]
    - While collapsing/aggregating the dataset to the project level, aggregate (or keep depending on the variable) the following variables: mean project capac- ity (in MW), operating year, total number of turbines in the project, mean capacity of a turbine in the project (in MW), mean turbine hub-height, mean turbine rotor diameter, state, and county where the project is located (for some of these variables, you may use the first occurrence).

8. Create a summary table for the data at the project level. Report the mean, standard deviation, min, and max of mean project capacity, average number of turbines in a project, average hub-height, average rotor diameter. Report the total number of projects (i.e. N)? *

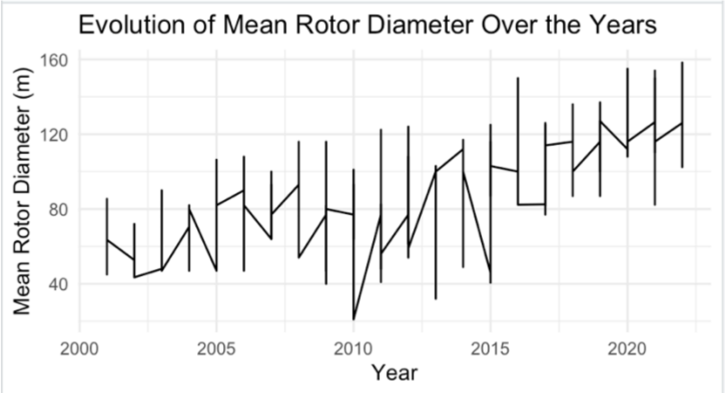


```
$ Mean_Project_Capacity    : num 112
$ SD_Project_Capacity      : num 104
$ Min_Project_Capacity     : num 0.2
$ Max_Project_Capacity     : num 1056
$ Mean_Turbines_Per_Project: num 55.6
$ SD_Turbines_Per_Project  : num 56.9
$ Min_Turbines_Per_Project : int 1
$ Max_Turbines_Per_Project : int 849
$ Mean_Hub_Height          : num 82.4
$ SD_Hub_Height            : num NA
$ Min_Hub_Height           : num 82.4
$ Max_Hub_Height           : num 82.4
$ Mean_Rotor_Diameter      : num 97.4
$ SD_Rotor_Diameter        : num NA
$ Min_Rotor_Diameter       : num 97.4
$ Max_Rotor_Diameter       : num 97.4
$ Total_Projects           : int 1212
```

9. While using the dataset at the project level, create a plot(s) (using ggplot) showing the evolution of mean hub-height and mean rotor diameter over the years (2001 - 2022). Provide a brief interpretation of your findings. *



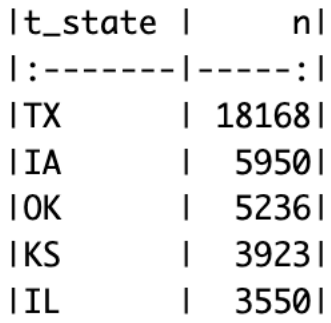
The plot for mean hub-height indicates an overall increasing trend in the height of wind turbine hubs over the years. This suggests that newer wind turbines are being constructed with taller structures, possibly to capture wind at higher altitudes where it is stronger and more consistent. However, there is a dip in the trend, which could be due to several factors such as changes in design preferences, regulations, or the introduction of new technology that allows for efficient operation at lower heights.



The mean rotor diameter plot also generally trends upward, with some fluctuations. Larger rotor diameters indicate turbines with a greater capacity to capture wind energy, which aligns with the push for increased efficiency in wind energy production. The fluctuations might be attributed to technological advancements, market demands, or variations in wind farm locations, where different rotor sizes may be favored.

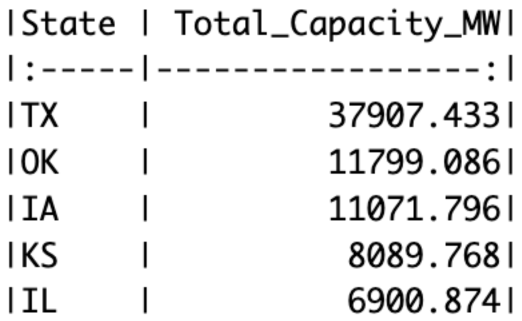
10. Using the cleaned turbine data and the project level data, answer the following questions (report all of these information in a well formatted table): *

   (a) Calculate the top five states with the most wind turbines.



```
|t_state |     n|
|:-------|-----:|
|TX      | 18168|
|IA      |  5950|
|OK      |  5236|
|KS      |  3923|
|IL      |  3550|
```

   (b) Calculate the top five states with the highest capacity of wind projects (in MW). Show the total capacity of that state and the US for comparison.



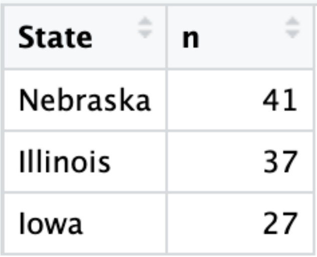
```
|State | Total_Capacity_MW|
|:-----|-----------------:|
|TX    |         37907.433|
|OK    |         11799.086|
|IA    |         11071.796|
|KS    |          8089.768|
|IL    |          6900.874|
```

## 2 Wind Ordinance Data

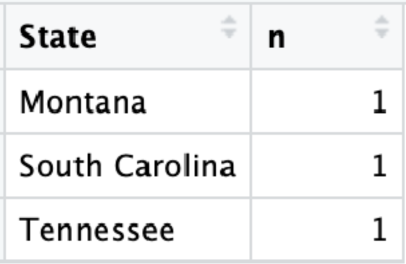
1. Download the wind_ordinance_main.csv dataset from Canvas and bring this data into R. This dataset is already clean, so doesn't require any further cleaning. This data contains information on the year *t* an ordinance was enacted (column ordinance_year) in a County *i* in State *j*.

2. Remove observations before 2001.

3. Calculate the 3 states with the *most* number of wind ordinances and the 3 states with *least* number of wind ordinances. Compare this with the total wind capacity (MW) in these states. Is there any meaningful conclusion we can draw from this comparison? Why or why not? *



| State | n |
|---|---|
| Nebraska | 41 |
| Illinois | 37 |
| Iowa | 27 |



| State | n |
|---|---|
| Montana | 1 |
| South Carolina | 1 |
| Tennessee | 1 |

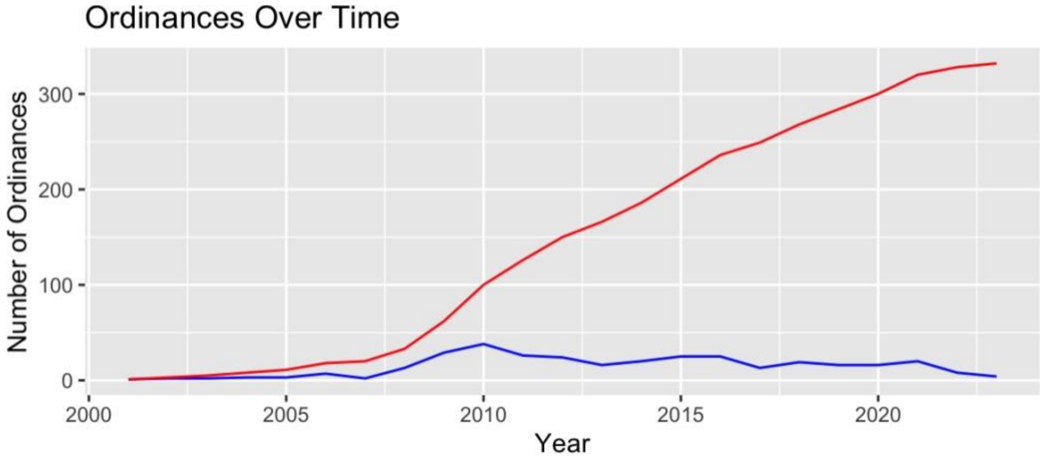
No meaningful conclusions can be drawn because the states themselves do not have a clear pattern when cross comparing wind ordinances and total wind capacity. The states in the tables above are scattered throughout the table containing total wind capacities by state.

4. Collapse this dataset at the year level. Create a graph (using ggplot) that shows two lines: *

   (a) **cumulative** number of ordinances in each year, and

   (b) total **new** ordinances enacted each year

   Compare the two lines and provide a brief interpretation (2-3 sentences).



Red – Cumulative number of ordinances

Blue – Total new ordinances

The graph shows two trends: the cumulative number of wind ordinances over time and the total new ordinances enacted each year. The red line indicates a steady increase in the cumulative number of ordinances, reflecting a growing adoption of wind-related regulations over time. The blue line, however, shows variability from year to year with no clear trend of increase or decrease, suggesting that the enactment of new ordinances is subject to other factors and does not necessarily grow consistently each year. This comparison highlights that while the total number of ordinances grows, the annual rate at which they are enacted can fluctuate significantly.

# 3 Wind Resource Quality Data

1. The WIND Toolkit includes instantaneous meteorological conditions from computer model output and calculated turbine power for more than 126,000 sites in the continental United States for the years 2007–2013. You can read more about this dataset here. Download the csv file wtk_site_metadata.csv from Canvas and bring it in R.

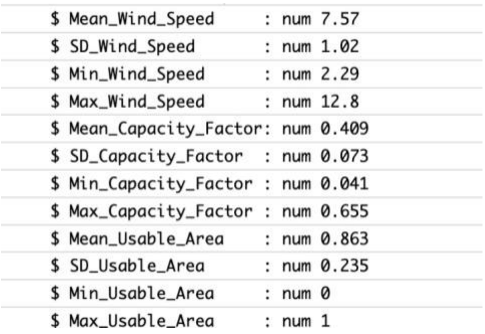
2. This dataset includes the following variables: longitude, latitude, fraction of usable area, average wind speed (m/s), capacity factor (measure of productivity), power curve (class of wind turbine appropriate for the site). This report provides details of this dataset. A brief explanation of key variables is below:

    - Power curve information is based on simulated wind speeds and gusts at 100 m hub-height. It specifies what kind of wind turbines are best suited for the particular location. Sites are classified as: offshore (very high wind speeds), 1 (high winds), 2 (medium winds), and 3 (low winds).
    - Capacity factor is the ratio of actual energy output to the maximum possible output. Average capacity factors of on-shore wind in the US ranges from 25% to 55%.
    - Capacity (MW) is a measure of total capacity of wind turbines that can be accommodated in the 2 km × 2 km grid for the particular location. This variable is measured in increments of 2 MW with the range of 0 to 16.

3. Remove all off-shore observations (i.e. unknown state, county, or “offshore” power curve).

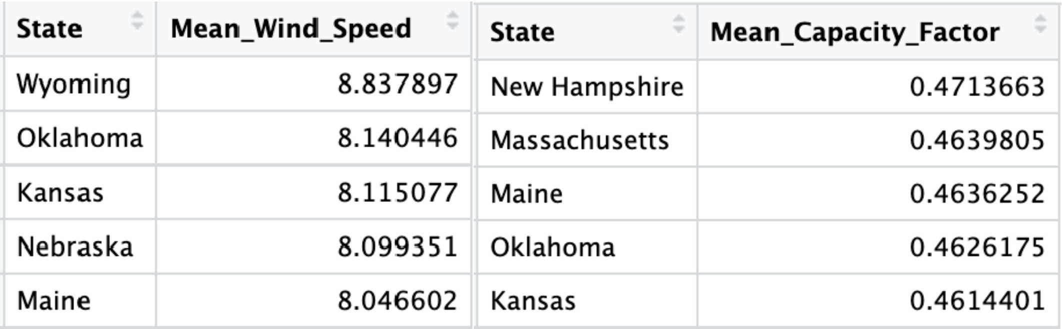
4. Remove observations in Alaska and Hawaii.

5. Create a summary table with mean, standard deviation, min, and max of wind speed (m/s), capacity factor, fraction of usable area. *



```
 $ Mean_Wind_Speed     : num 7.57
 $ SD_Wind_Speed       : num 1.02
 $ Min_Wind_Speed      : num 2.29
 $ Max_Wind_Speed      : num 12.8
 $ Mean_Capacity_Factor: num 0.409
 $ SD_Capacity_Factor  : num 0.073
 $ Min_Capacity_Factor : num 0.041
 $ Max_Capacity_Factor : num 0.655
 $ Mean_Usable_Area    : num 0.863
 $ SD_Usable_Area      : num 0.235
 $ Min_Usable_Area     : num 0
 $ Max_Usable_Area     : num 1
```

6. List the top five states with the highest wind speeds and capacity factors. Compare this with the states with the highest wind project capacity (from question 1.10). What do you notice? *



| State | Mean_Wind_Speed | State | Mean_Capacity_Factor |
|---|---|---|---|
| Wyoming | 8.837897 | New Hampshire | 0.4713663 |
| Oklahoma | 8.140446 | Massachusetts | 0.4639805 |
| Kansas | 8.115077 | Maine | 0.4636252 |
| Nebraska | 8.099351 | Oklahoma | 0.4626175 |
| Maine | 8.046602 | Kansas | 0.4614401 |

In the data provided, the top five states with the highest wind speeds do not overlap with the states that have the highest mean capacity factors, nor do they match the states with the highest wind project capacity. This discrepancy suggests that while wind speed is a crucial factor for wind power, the actual implemented capacity depends on other variables such as the quality of wind, economic policies, and the availability of technology suited to the particular characteristics of each state's wind resource. It highlights that high wind speeds alone do not directly correlate to high productivity or capacity implementation in wind energy projects.

7. Compare the top **five** states with highest wind speed to the top **five** states with most wind ordinances (from question 2.3).[2] What do you observe? *

Comparing the top five states with the highest wind speeds against those with the most wind ordinances, we observe that only one state, Nebraska, is common to both lists. Nebraska appears to be utilizing its wind resources effectively, reflected in both its wind speeds and the number of wind ordinances. The other states with high wind speeds (Wyoming, Oklahoma, Kansas, and Maine) do not feature in the top five for wind ordinances, suggesting that factors other than wind speed are influencing policy decisions. Meanwhile, states like Illinois, Iowa, Minnesota, and South Dakota are proactive in enacting wind ordinances despite not being listed among the top for wind speeds. In the end, there is not high correlation between the two.

# 4 Merging the Datasets

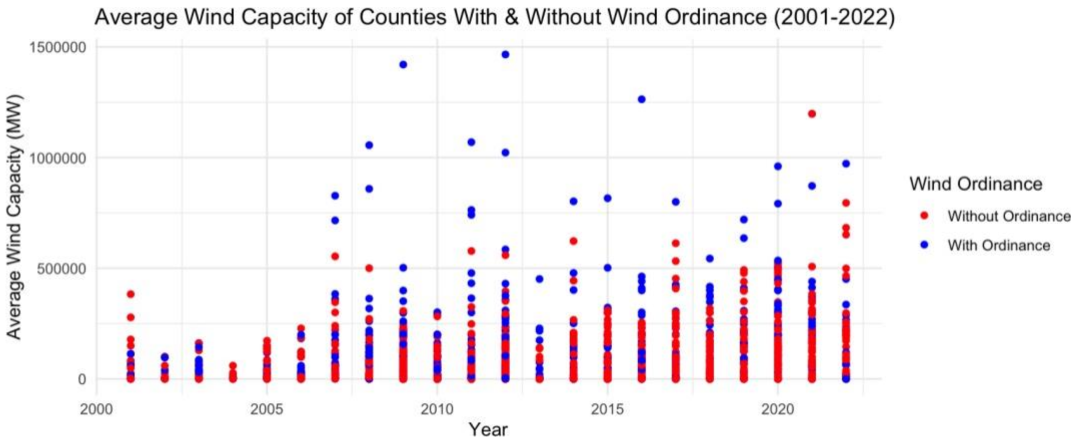
Note that you will need to merge these datasets in order to perform any meaningful analysis. The tasks below will get your started with that:

1. Merge the wind turbines and wind ordinance datasets on the basis of state and county variables.[3] Note that county names across the states may not be unique, so, keep this in mind while writing the code for merging.

2. Certain observations in the merged dataset might have NAs in the ordinance_year field. Create a binary variable that tracks which observations have an ordinance in a specific year.

3. Collapse this data to the county level. Make sure to preserve the temporal dimensions of the data.

4. For this dataset at the county level, create a graph that shows the average wind capacity (MW) of counties with and without a wind ordinance over the years (from 2001 to 2022). What do you notice? *

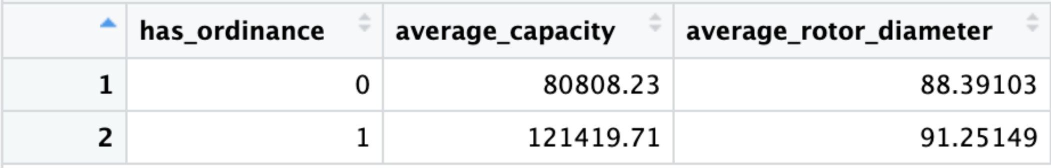


Counties with wind ordinances display higher peaks in average wind capacity, possibly indicating larger or more frequent wind energy projects. There is a clear upward trend in wind capacity over the years, reflecting the growth of the wind energy sector. Interestingly, the presence of wind ordinances doesn't consistently lead to year-over-year capacity increases, with some years showing higher capacities in counties without such ordinances. The wide spread of data points, particularly in later years, suggests significant variability in average wind capacity among different counties. These findings raise questions about the factors influencing wind energy development, such as economic incentives, geographic location, and local policies.

5. Now, collapse the merged dataset in (2). at the project level. Create a summary table that could answer the following question: *

    - Are projects in a county with an ordinance bigger in size (nameplate capac- ity), and use more technologically advanced turbines (i.e. bigger turbines measured by rotor diameter)?

    Does this information from the Table you created, provide any conclusive evidence on the effectiveness of wind ordinances? What are the confounding factors in making a claim?



| | has_ordinance | average_capacity | average_rotor_diameter |
|---|---|---|---|
| 1 | 0 | 80808.23 | 88.39103 |
| 2 | 1 | 121419.71 | 91.25149 |

    Counties with wind ordinances have an average nameplate capacity of 121,419.71 MW, significantly higher than the 80,808.23 MW in counties without ordinances, indicating larger-scale projects in the former. Additionally, the average rotor diameter in counties with ordinances is 91.25 meters, slightly larger than the 88.39 meters in counties without ordinances, potentially indicating the use of more advanced and efficient turbine technology. However, it's important to consider confounding factors, such as selection bias, technological advancements, economic and political influences, and data limitations when interpreting these findings.

6. Notice, we haven't merged the turbine data with the Wind Resource Quality data (WIND Toolkit). There are a few approaches by which you can match a wind project to the wind resource quality in that area. Briefly explain how you might do this. *

    An approach to merging turbine data with Wind Resource Quality data would involve creating predictive models that can estimate wind resource characteristics at turbine locations based on a set of features. These features might include geographic coordinates, elevation, and other relevant environmental data. By training a machine learning model on the Wind Resource Quality data, where wind characteristics are already known, the model can learn the relationship between the features and the wind resource quality. Once trained, the model could then predict wind resources at new turbine locations using the features of those locations.