Growth of the Wind Energy Industry in Germany

Tools: Python, Jupyter Notebooks, Tableau

Libraries: Pandas, Numpy, Clustering, Geopandas, SKlearn Quandl.

Posted by Graham Field on April 19, 2024 – 10 mins read

Duration of the project: 1 week

Role: Data Analyst

Credits and Data Sources:

- https://www.marktstammdatenregister.de/MaStR/Datendownload (Data Source)
- https://github.com/isellsoap/deutschlandGeoJSON/blob/main/2_bundeslaender/1_sehr_hoch.geo.json (GeoJSON)

Tools:

- Python
- Jupyter Notebooks
- Libraries: Pandas, Numpy, Clustering, Geopandas, SKlearn Quandl
- Tableau

Data:

- 34484 rows & 29 variables
- 17 Regions, 342 Districts, 115 Manufacturers
- Open-source data

Key Research Questions

- How have the dimensions of the turbines grown over time?
- What is the relationship between increase in Gross performance and physical dimensions?
- What understandings can I gain from Clustering and time series analysis?
- What is the geographical spread of the Windparks?
- Who is responsible for what kind of production?

Introduction

The renewable energy industry has seen massive growth over the years given the importance of moving away from fossil fuels. One of the most effective ways of producing renewable energy is through wind. A global leader in this transition has been Germany.

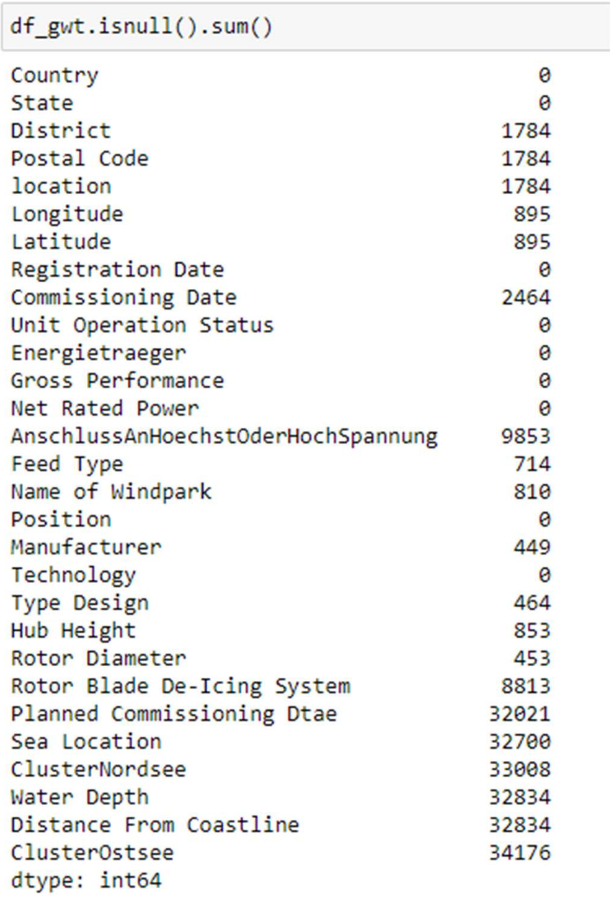
The growth of the wind industry has many aspects to it, both technologically and geographically. With limited space and limited suitable areas for building wind farms the drive to make turbines more powerful and efficient is a major driving force in the industry.

Understanding how the new technologies have improved and how this has led to large increases in production capacity and physical dimensions of the turbines makes for an interesting study. Perhaps shedding light on the future of the industry.

Step 1: Initial Data Exploration and Cleaning

This is open-source data from the German Data Register (Marktstammdatenregister).

This data is regularly updated and free to use for research purposes.



```
df_gwt.isnull().sum()
```

```
Country                                 0
State                                   0
District                             1784
Postal Code                          1784
location                             1784
Longitude                             895
Latitude                              895
Registration Date                       0
Commissioning Date                   2464
Unit Operation Status                   0
Energietraeger                          0
Gross Performance                       0
Net Rated Power                         0
AnschlussAnHoechstOderHochSpannung   9853
Feed Type                             714
Name of Windpark                      810
Position                                0
Manufacturer                          449
Technology                              0
Type Design                           464
Hub Height                            853
Rotor Diameter                        453
Rotor Blade De-Icing System          8813
Planned Commissioning Dtae          32021
Sea Location                        32700
ClusterNordsee                      33008
Water Depth                         32834
Distance From Coastline             32834
ClusterOstsee                       34176
dtype: int64
```

There were several columns with NAN values, to deal with this I dropped many of the columns that were not necessary or irrelevant or not applicable, such as water depth for onshore turbines. The remaining NANs were either deleted. Several of the columns were merged which completed the data in those columns. Ultimately the data that I worked with with Python the data was relatively easy to clean and organise.

Further I exploratory analysis confirmed that there were no duplicates, and the remaining missing values before being removed constituted approximately 2.5% these were also generally data points that represent wind parks in the planning or construction phase and have not yet been commissioned.

From the Data cleaning and exploration i have learned that the offshore industrial zone is has its own column but I can also consider this as an additional statefor the purposes of this case study. This greatly reduced the number of NANs and missing data in the State, ClusterNordsee, and ClusterOstsee Columns.

Step 2: Exploring Correlations

Strong Correlations

Gross performance, Net rated power, Hub Height, and Rotor Diameter have very strong correlations which makes sense, as the turbine technology improves the turbines have been getting taller, their blades getting longer, and their power outputs have been increasing. with relation to Rotor Diameter, I can see this very well. it is worth noting here that net rated power and gross output are the same and that is why I have a 1-1 correlation.

with hub height and power, I see a very strong correlation, this makes sense as the larger the blades are the more wind they can catch and the more power they can generate.

with hub height and rotor diameter, I also see a strong correlation which also makes sense as the longer the blades get the higher you need to build the turbine. I don't want blades hitting the ground and I also want to keep the blades in the wind.

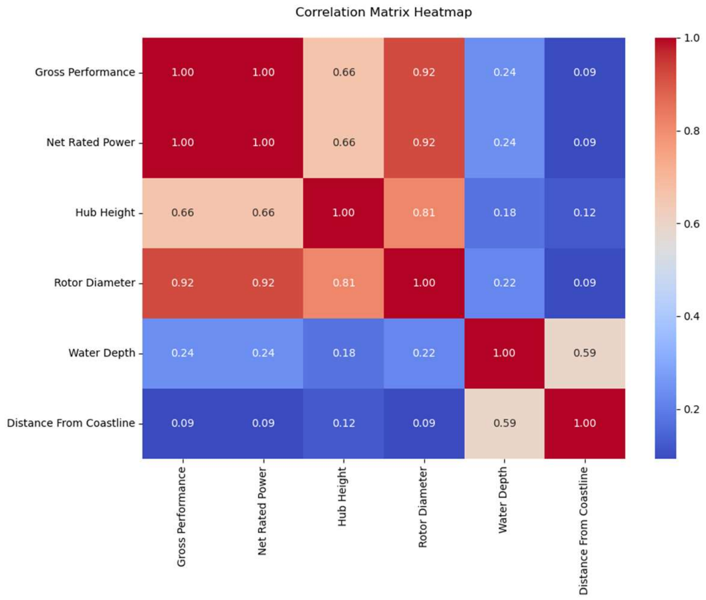
I can also see a correlation between water depth and distance from the coastline which makes sense, as the further you go from land the deeper the water gets.

Weak Correlations

when it comes to the offshore sector. I see weak correlations because this does not determine the type of turbine that is built. As the technology improves the biggest and the best will be built. given that some turbines can be built to float the water depth and distance from the coastline really should have much impact. water depth and distance from the coastline also have weak correlations and would have very little to do with the type of turbine built as the water line is the water line and the metrics I have are ultimately determined by that. I can say that the the deeper the water and the further away from the coastline does not correlate to larger more powerful turbines.

Conclusion

the improvement of technology is represented by the growth in power output and the strong correlating increase in Rotor diameter and hub height. As the turbines become more powerful the blades get longer and the towers get taller. the turbines being built offshore does not impact the improved power and size of the turbines and I can see this in the weak correlation. with water depth and distance from shore.



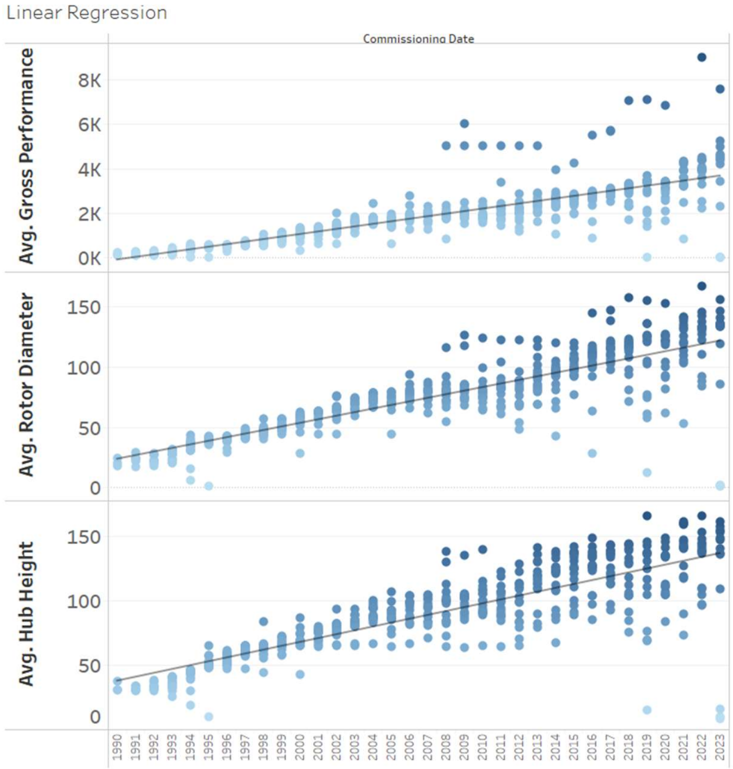
I learned that there is a very weak correlation between the physical dimensions and power rating vs the water depth and distance from the coastline. This makes sense as the type design of turbines isn't affected by whether it is onshore or offshore.

Step 3: Linear Regression

"As turbines Improve over time their dimensions and output increase."

This was the hypothesis i wanted to test, i using a simple regression method. I used test and training subsets of data to determine the relationships between the various dimensions of turbines.

I can see there is a very strong correlation between Hub Height and rotor diameter vs Gross Performance, As the power increases so does the Rotor Diameter and Hub Height. there are a few outliers however these could be between wind parks that still need to be built or prototypes, which may be listed as 0.



When comparing these 3 aspects I can see that they all follow a very similar pattern. This is going to be important to predict the size and power of future turbines.

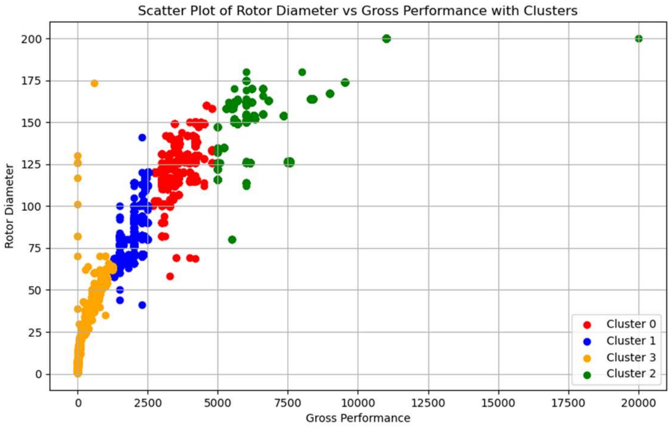
This will make it much easier to map out future potential locations for wind farms, ultimately earmarking locations for projects before they are conceptualized and helping the planning of infrastructure.

It will be quite interesting to see how the clustering performs.

When Looking at the graph I notice that I could extrapolate and predict the future dimensions of turbines as there is a nice linear pattern and relations ship between Gross performance, Rotor Diameter, and Hub Height. In future iterations, I would like to include a Predictive model and over time test it against existing theoretical and prototype Turbine Models Types.

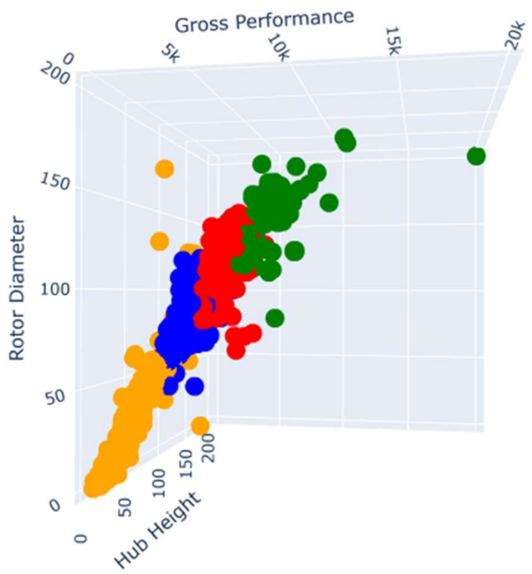
Step 4: Clustering

I used machine learning algorithms to perform k-means clustering on the data. Based on the results from the elbow method for optimal k i chose 4 clusters. Given the high correlations between the physical aspects of the turbines i did not need to standardise the information.

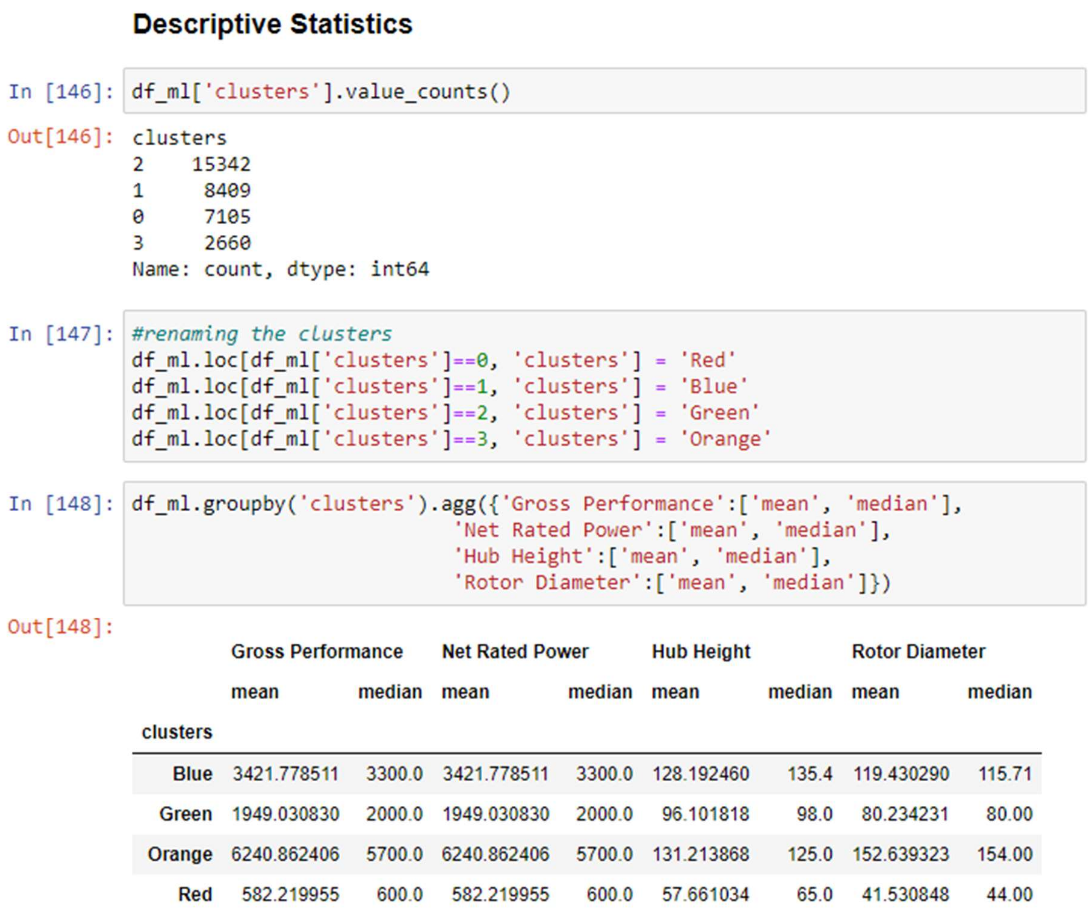


The clustering analysis between rotor diameter and gross performance has yielded insightful results, illuminating distinct patterns and relationships within the data. Through the application of clustering algorithms, I have successfully identified meaningful groupings or clusters based on the inherent characteristics of rotor diameter and gross performance. This segmentation allows us to discern clear trends and variations in the dataset, providing valuable insights into the underlying structure of the data. By uncovering these clusters, I gain a deeper understanding of how rotor diameter and gross performance interact and influence each other within the context of our analysis.

The success of the 3D clustering analysis between gross performance, rotor diameter, and hub height is evident in its ability to reveal multidimensional relationships and patterns within wind turbine data. By integrating these critical parameters into a cohesive analysis framework, clustering algorithms effectively segment the data into distinct clusters based on shared characteristics. Beyond segmentation, this analysis fosters deeper insights into the interactions between performance metrics and design attributes, informing decisions regarding turbine design, siting, and operational strategies.



Similar to the insights gained from Linear regression, I would like to create a predictive model and then compare that to upcoming Theoretical and Prototype Turbine Model Types.



Descriptive Statistics

```
In [146]: df_ml['clusters'].value_counts()
```

```
Out[146]: clusters
          2    15342
          1     8409
          0     7105
          3     2660
          Name: count, dtype: int64
```

```
In [147]: #renaming the clusters
          df_ml.loc[df_ml['clusters']==0, 'clusters'] = 'Red'
          df_ml.loc[df_ml['clusters']==1, 'clusters'] = 'Blue'
          df_ml.loc[df_ml['clusters']==2, 'clusters'] = 'Green'
          df_ml.loc[df_ml['clusters']==3, 'clusters'] = 'Orange'
```

```
In [148]: df_ml.groupby('clusters').agg({'Gross Performance':['mean', 'median'],
                                         'Net Rated Power':['mean', 'median'],
                                         'Hub Height':['mean', 'median'],
                                         'Rotor Diameter':['mean', 'median']})
```

Out[148]:

	Gross Performance		Net Rated Power		Hub Height		Rotor Diameter	
	mean	median	mean	median	mean	median	mean	median
clusters								
Blue	3421.778511	3300.0	3421.778511	3300.0	128.192460	135.4	119.430290	115.71
Green	1949.030830	2000.0	1949.030830	2000.0	96.101818	98.0	80.234231	80.00
Orange	6240.862406	5700.0	6240.862406	5700.0	131.213868	125.0	152.639323	154.00
Red	582.219955	600.0	582.219955	600.0	57.661034	65.0	41.530848	44.00

Interpretation:

- Gross Performance and Net Rated Power: These two features seem to have relatively high mean and median values in the Blue cluster compared to other clusters. The mean and median values for the Green cluster are slightly lower, followed by the Red cluster, and then the Orange cluster, which has the lowest values.
- Hub Height and Rotor Diameter: The mean and median values for Hub Height and Rotor Diameter also vary across clusters. The Blue cluster tends to have higher values for both features compared to other clusters. The Green cluster follows with slightly lower values, then the Red cluster, and finally the Orange cluster with the lowest values.

Cluster Characteristics:

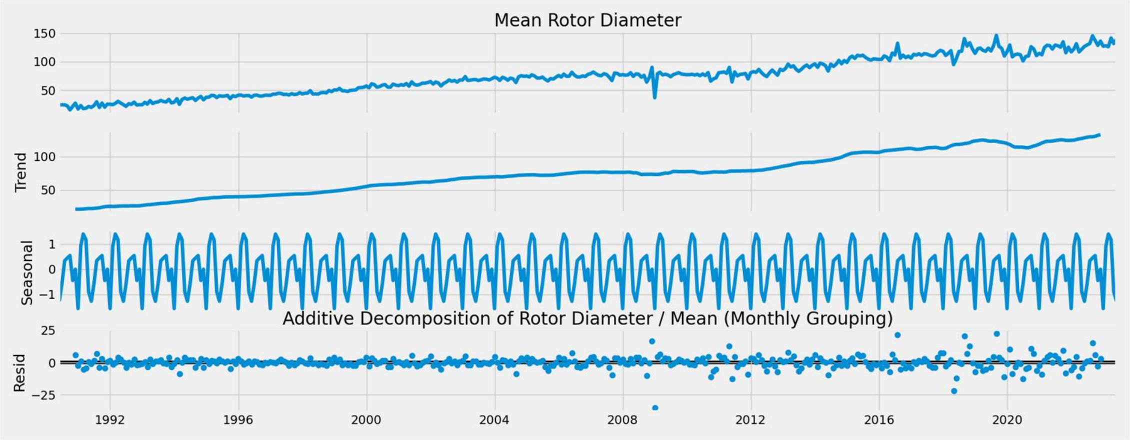
Based on the mean and median values, I can infer certain characteristics of each cluster:

- Blue Cluster: This cluster appears to have the highest values for Gross Performance, Net Rated Power, Hub Height, and Rotor Diameter among all clusters, indicating that it may represent wind turbines with higher performance and larger physical dimensions.
- Green Cluster: This cluster has slightly lower values compared to the Blue cluster but higher than the Red and Orange clusters, suggesting moderate performance and dimensions.
- Red Cluster: The Red cluster has lower values for all features compared to the Blue and Green clusters, indicating lower performance and smaller physical dimensions.
- Orange Cluster: This cluster has the lowest values for all features, suggesting the lowest performance and smallest physical dimensions among all clusters.

over all i would say this model has been succesfull at learning and clustering our data. the data is inline with all our previous analysis and has provided some valuable insights.

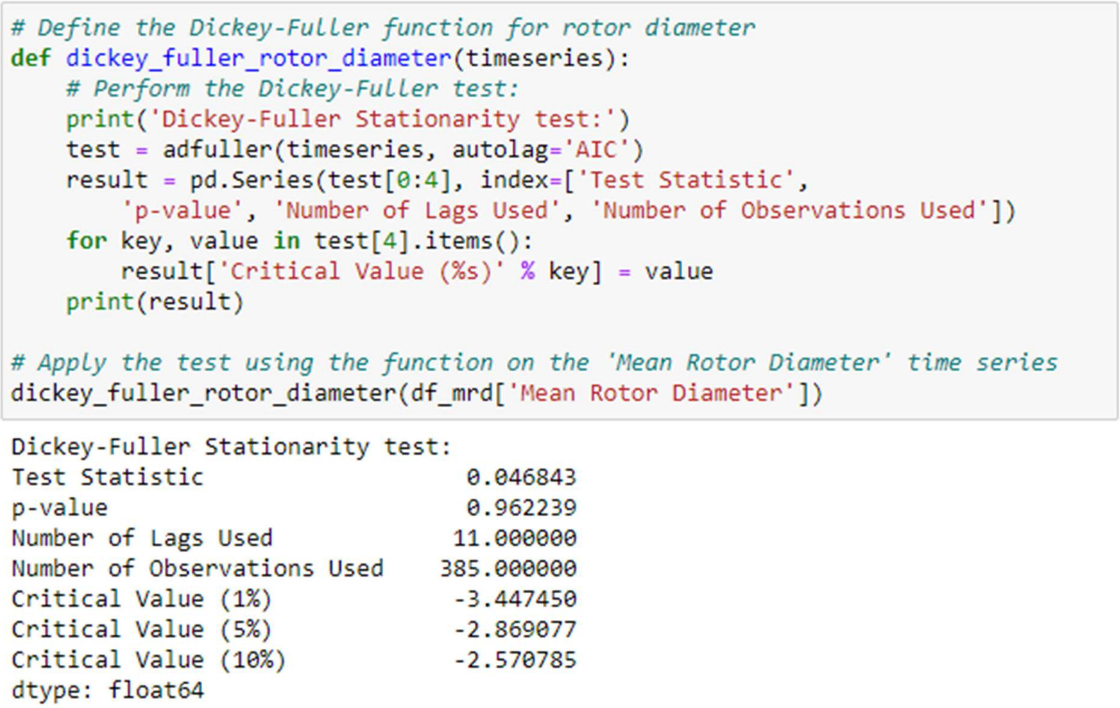
Step 5: Time Series

The time series of rotor diameter is crucial in wind turbine analysis as it directly influences turbine performance, efficiency, and maintenance requirements. Understanding its trends and variations aids in optimizing turbine design, predicting energy output, and implementing effective maintenance schedules, ultimately maximizing the operational efficiency and longevity of wind energy systems.



There is a clear upward trend for rotor diameter there is a very nice upward trend line. there are some dips here and there but generally around times where there were siginifcant changes in the economy or society such as 2008 where there was a financial crash.

Stationarizing the time series data of rotor diameter using the Dickey-Fuller method is pivotal for extracting meaningful insights and improving forecasting accuracy. By transforming the non-stationary rotor diameter time series into a stationary one, this process facilitates clearer identification of underlying patterns and relationships within the data. The success of stationarizing lies in its ability to stabilize statistical properties like mean and variance over time, achieved by removing trends and seasonality. The Dickey-Fuller test serves as a reliable diagnostic tool, confirming the stationarity of the series when the null hypothesis, indicating a unit root and non-stationarity, is rejected. Overall, stationarizing the rotor diameter time series enhances the reliability of subsequent analyses, enabling more accurate modeling and forecasting of future rotor diameter trends.



```python
# Define the Dickey-Fuller function for rotor diameter
def dickey_fuller_rotor_diameter(timeseries):
    # Perform the Dickey-Fuller test:
    print('Dickey-Fuller Stationarity test:')
    test = adfuller(timeseries, autolag='AIC')
    result = pd.Series(test[0:4], index=['Test Statistic',
        'p-value', 'Number of Lags Used', 'Number of Observations Used'])
    for key, value in test[4].items():
        result['Critical Value (%s)' % key] = value
    print(result)

# Apply the test using the function on the 'Mean Rotor Diameter' time series
dickey_fuller_rotor_diameter(df_mrd['Mean Rotor Diameter'])
```

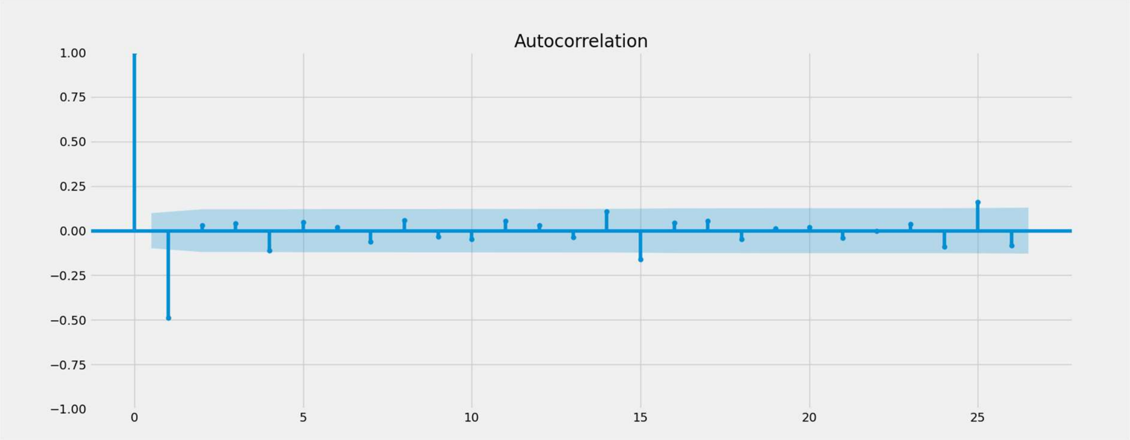
```
Dickey-Fuller Stationarity test:
Test Statistic                   0.046843
p-value                          0.962239
Number of Lags Used             11.000000
Number of Observations Used    385.000000
Critical Value (1%)             -3.447450
Critical Value (5%)             -2.869077
Critical Value (10%)            -2.570785
dtype: float64
```

Typically, if the p-value is less than a chosen significance level (such as 0.05), you reject the null hypothesis and conclude that the time series is stationary. In this case, the p-value is approximately 0.962239, which is greater than 0.05. Therefore, I fail to reject the null hypothesis, indicating that there is insufficient evidence to conclude that the time series is stationary.

Here I can see the success of the dicky fuller test in stationorising the time series.

Autocorrelation

- First test, The vertical lines represent the lags in the series, while the blue area represents the confidence interval. When lines go above the blue edge of the confidence interval, this means the lags are significantly correlated with each other. I have many lags beyond this interval and can deduce that this data is non-stationary.
- Second test, things have improved a lot here, but it's still not perfect. This means that, despite the Dickey-Fuller test saying that data is stationary, that stationarity is more acceptable but still weak. there are only 4 lags that are not within the confidence area.



Conclusion

For all the time series that I tested I can reject the null hypothese and consider the time series as non stationary. Based on the interpretation of the Dickey-Fuller test results, where the test statistic is not significantly low and the p-value is greater than common significance levels like 0.05 or 0.01, I cannot reject the null hypothesis. Therefore, I conclude that the data is likely non-stationary, indicating the presence of a trend or seasonality. The autocorrelation confirms this.

After stationarizing a time series and observing less than 5 nodes outside the spread, it suggests a strong autocorrelation pattern within the data. This indicates that the values in the time series are highly correlated with their past values, which is a common characteristic of stationary time series data. With fewer than 5 nodes outside the spread, it implies that the majority of the data points closely adhere to the autocorrelation structure, indicating a stable and predictable behaviour over time. This could be advantageous for forecasting and modeling purposes, as it suggests that past values can reliably inform future trends within the time series.

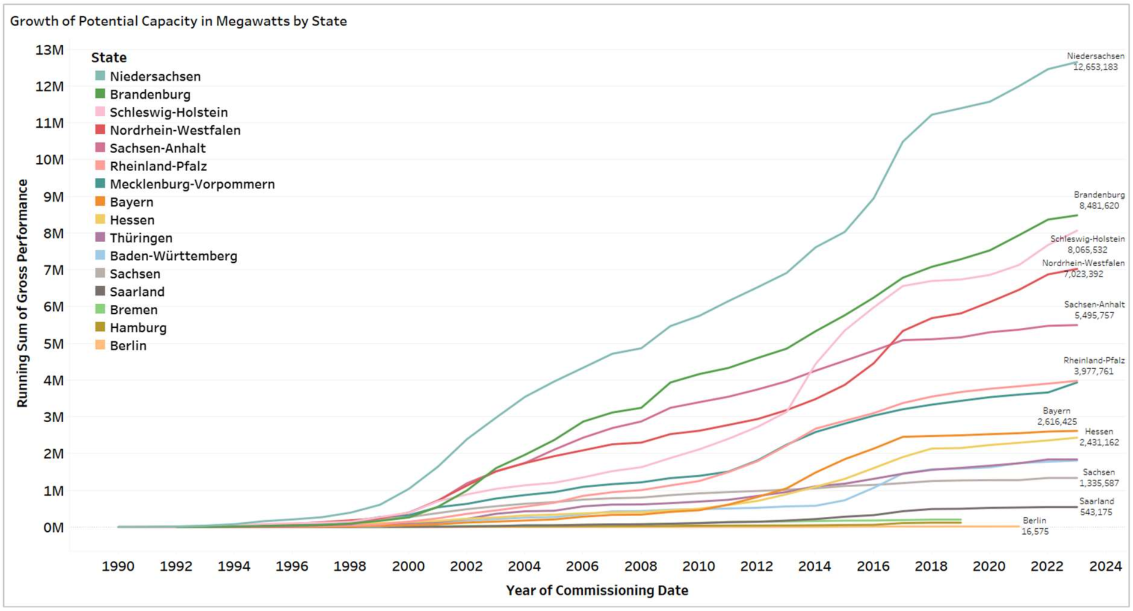
I was not Surprised by the results of the time series analysis, in relation to previous steps in this analysis I can see a clear pattern of increase in all dimensions over time. The Rotor Diameter has the Most linear and predictable trend of all the dimensions. This would be a great start for predictive modeling and forecasting.

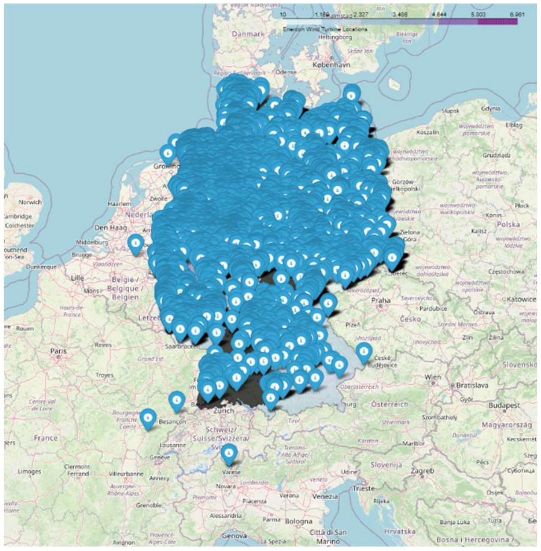
Step 6: Geospatial

Given that wind farms require quite alot of space, it is important that I consider where and who is building the turbines. In this section i wanted to look at the areas with the highest production and the top manufacturers responsible for that production. In this section I will look at the top 3 manufacturesres as they all useslighlty different strategies.

Top 10 Manufacturers by State

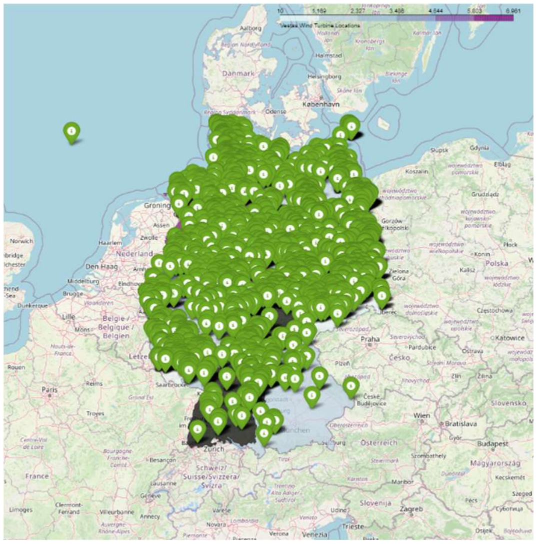
From the previous analytic steps I can see and confirm that the top 3 states are Niedersachsen, Schleswig-Holstein, and Brandenburg. I also have the offshore sector which actually comes in second if I consider it as a state





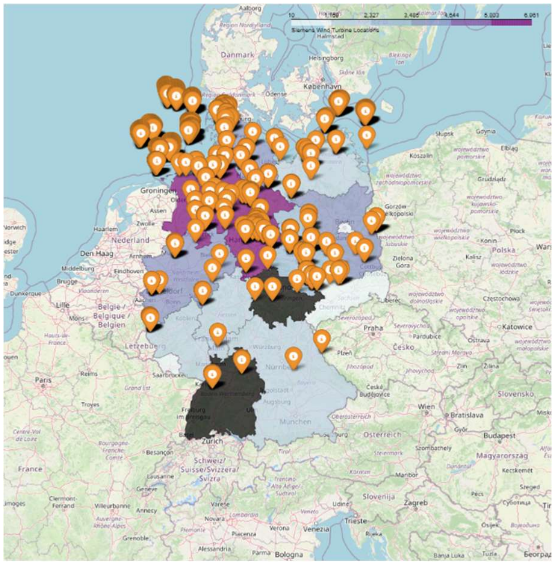
Enercon Locations

Given that Enercon has so many smaller installations it's no surprise that I see such a densely populated Choropleth. I notice that Enercon does not have any offshore windparks which is interesting.



Vestas Locations

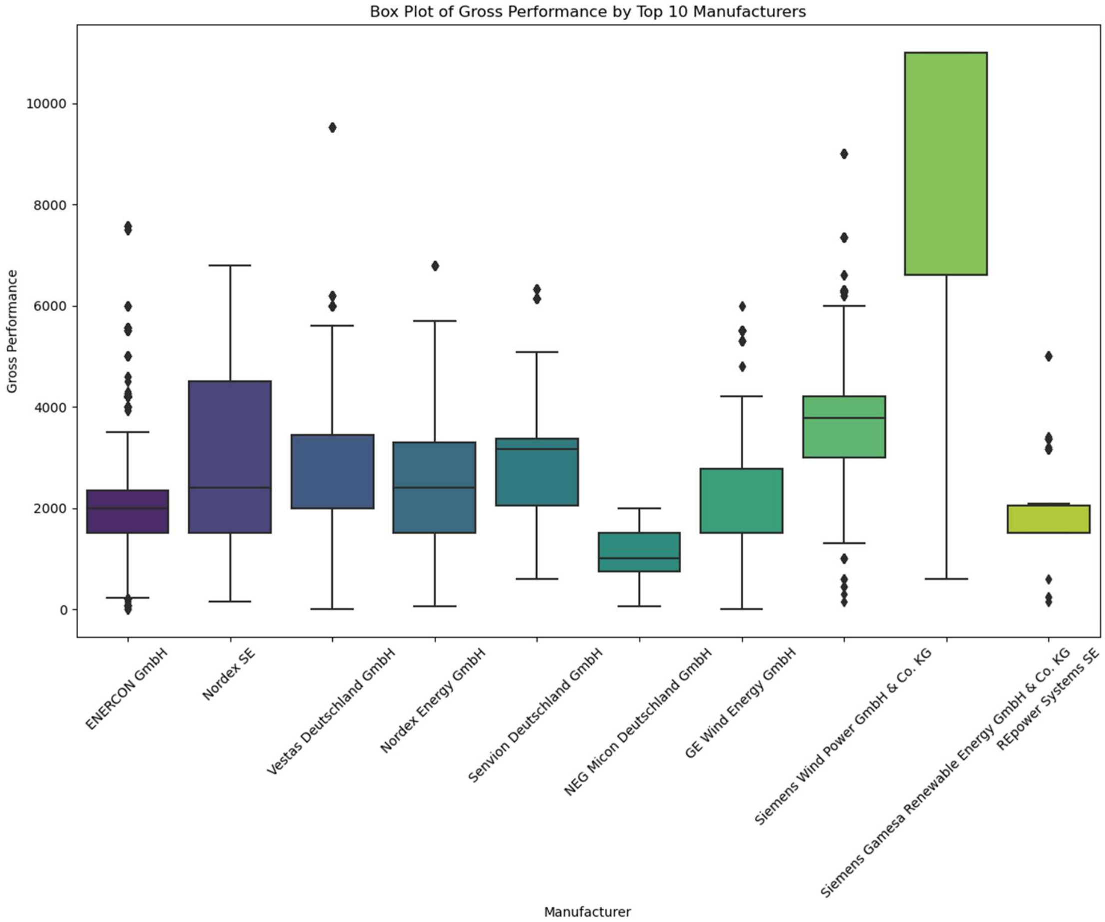
Vestas has a mix of sizes at their locations. I see that they have very good land coverage but they also have offshore installations. They are very Ill spread across the country and surrounding sea.



Siemens Locations

Siemens has by far the fewest locations which corresponds with our previous analytical steps. They have a few massive installations however it seems their main business is building these windparks offshore.

When looking at this box and whisker plot I can see that Enercon has mainly smaller installation with a few outliers of large wind parks. Vestas has slightly more spread with most of its windparks in the medium sized category. When I look at siemens I can see that they generally have massive installation which makes sense as most off shore wind park are huge. Given that siemens is mostly responsible for offshore wind parks this would make sense.



I found the results to be Surprising, Smaller installations at a higher frequency are more effective at producing a greater cumulative power capacity, this would also have major benefits for transmission of power as the electricity doesn’t need to travel large distances to where it is needed. Building the production close to where it is needed is clearly a preferable method of producing power.

Further Research

It would be nice if I could include meteorological data so I could see where the highest and most consistent sites are for wind turbines and compare that to the current infrastructure.

Combining the two and then conducting a time series analysis to see how or if the weather patterns change would be an interesting theory to test.

Retrospective

There are some skills I will definitely need to develop for the Further research of this project. Some of the main challenges I have are linking the dynamic meteorological data to the cumulative geospatial aspects. It will be very interesting trying to map the growth in height in turbines, the rotor diameter, and the gross output (the power being extracted from the wind) against the Meteorological conditions to see if there are any changes in local and regional climatic conditions will be challenging, I am quite excited to develop these skills and share the results in the near future. Perhaps I will notice a change in weather patterns one the physical dimensions or physical dimensions reach a certain point or a combination of these aspects. Perhaps I will see a change when a certain density of these aspects is reached.

Conclusions

The data I've explored paints a compelling picture of the ongoing evolution in wind turbine technology, with a clear trend towards larger physical dimensions and increased power output. The analysis of rotor diameter, hub height, and net rated power reveals a notable upward trajectory over time, suggesting a concerted effort within the industry to enhance turbine efficiency and performance. This trend underscores a broader commitment to harnessing renewable energy resources more effectively, as larger turbines with higher power capacities have the potential to generate greater electricity yields while maximizing land and resource utilization. As technology continues to advance and economies of scale drive further innovation, I anticipate this trend towards larger and more powerful turbines to persist, signaling a promising future for the continued growth and sustainability of wind energy.

Reflections

During this analysis, I encountered several challenges that required careful navigation and problem-solving. One notable issue was the handling of non-stationary time series data, particularly in relation to rotor diameter trends. Stationarizing the data using the Dickey-Fuller method proved crucial but required attention to detail and an understanding of the underlying statistical concepts. Additionally, managing the visualization and interpretation of multidimensional data, such as clustering between rotor diameter, hub height, and gross performance, presented complexities in conveying insights effectively. Despite these challenges, each hurdle served as an opportunity for growth and learning, ultimately enhancing my analytical skills and proficiency in wind turbine analysis.

The Future of the Industry

Looking ahead, the future of the wind energy industry holds immense promise, propelled by advancements in technology and a growing commitment to sustainability. Integrating meteorological data, especially wind speed information, into our analyses provides invaluable insights into turbine performance and energy production. By considering environmental factors alongside turbine specifications, I can optimize siting decisions, improve operational efficiency, and enhance forecasting accuracy. This holistic approach not only fosters a deeper understanding of wind energy dynamics but also paves the way for more robust and resilient energy systems. As the industry continues to innovate and adapt to evolving challenges, the integration of meteorological data stands poised to play a pivotal role in shaping the future trajectory of wind energy, driving progress towards a more sustainable and renewable energy landscape.