EDA

March 20, 2025

[18]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
from windrose import WindroseAxes
import seaborn as sns


df_on = pd.read_csv("final_onshore_data_2017_2025.csv")
df_off = pd.read_csv("final_offshore_data_2017_2025.csv")
```

[9]:
```
print("Onshore DataFrame info:\n")
df_on.info()
print("\nOffshore DataFrame info:\n")
df_off.info()

print("\nOnshore data sample:\n", df_on.head())
print("\nOffshore data sample:\n", df_off.head())

# Basic descriptive statistics
print("\nOnshore numeric stats:")
display(df_on.describe())

print("\nOffshore numeric stats:")
display(df_off.describe())
```

```
Onshore DataFrame info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70151 entries, 0 to 70150
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   year_mon_day       70151 non-null  int64
 1   hour               70151 non-null  int64
 2   wind_dir_avg_10    70151 non-null  float64
 3   wind_speed_h_avg   70151 non-null  float64
 4   wind_speed_avg_10  70151 non-null  float64
```

```
 5   air_pressure       70151 non-null  float64
 6   humidity           70151 non-null  float64
 7   full_datetime      70151 non-null  object
 8   capacity           70151 non-null  int64
 9   volume             70151 non-null  int64
 10  percentage         70151 non-null  float64
 11  emission           70151 non-null  int64
 12  emissionfactor     70151 non-null  int64
 13  correct_days       70151 non-null  object
dtypes: float64(6), int64(6), object(2)
memory usage: 7.5+ MB

Offshore DataFrame info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70151 entries, 0 to 70150
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   year_mon_day       70151 non-null  int64
 1   hour               70151 non-null  int64
 2   wind_dir_avg_10    70151 non-null  float64
 3   wind_speed_h_avg   70151 non-null  float64
 4   wind_speed_avg_10  70151 non-null  float64
 5   air_pressure       70151 non-null  float64
 6   humidity           70151 non-null  float64
 7   full_datetime      70151 non-null  object
 8   capacity           70151 non-null  int64
 9   volume             70151 non-null  int64
 10  percentage         70151 non-null  float64
 11  emission           70151 non-null  int64
 12  emissionfactor     70151 non-null  int64
 13  correct_days       70151 non-null  object
dtypes: float64(6), int64(6), object(2)
memory usage: 7.5+ MB

Onshore data sample:
   year_mon_day  hour  wind_dir_avg_10  wind_speed_h_avg  wind_speed_avg_10  \
0      20170101     1       207.708194         49.666667          49.666667
1      20170101     2       205.010321         50.000000          51.333333
2      20170101     3       202.701006         51.666667          51.000000
3      20170101     4       201.007553         52.333333          54.666667
4      20170101     5       200.325015         52.666667          53.333333

   air_pressure   humidity  full_datetime  capacity  volume  percentage  \
0  10234.526316  98.076923  2017-01-01-01    679334  679334    0.788730
1  10227.789474  98.153846  2017-01-01-02    677462  677462    0.786558
2  10219.473684  98.230769  2017-01-01-03    653746  653746    0.759025
```

```
3  10211.368421  98.038462  2017-01-01-04    705882  705882    0.819552
4  10203.526316  97.461538  2017-01-01-05    716738  716738    0.832158

   emission  emissionfactor   correct_days
0         0               0  2017-01-01-01
1         0               0  2017-01-01-02
2         0               0  2017-01-01-03
3         0               0  2017-01-01-04
4         0               0  2017-01-01-05

Offshore data sample:
    year_mon_day  hour  wind_dir_avg_10  wind_speed_h_avg  wind_speed_avg_10  \
0       20170101     1       213.586816         85.714286          86.428571
1       20170101     2       210.905296         87.142857          90.714286
2       20170101     3       208.585001         89.285714          87.857143
3       20170101     4       209.977979         90.000000          90.000000
4       20170101     5       208.541568         89.285714          87.142857

   air_pressure   humidity  full_datetime  capacity  volume  percentage  \
0      10206.75  95.714286  2017-01-01-01    873501  873501    1.014165
1      10199.75  96.142857  2017-01-01-02    883749  883749    1.026065
2      10191.50  96.000000  2017-01-01-03    872500  872500    1.013004
3      10182.25  96.142857  2017-01-01-04    889750  889750    1.033031
4      10176.25  95.571429  2017-01-01-05    893251  893251    1.037095

   emission  emissionfactor   correct_days
0         0               0  2017-01-01-01
1         0               0  2017-01-01-02
2         0               0  2017-01-01-03
3         0               0  2017-01-01-04
4         0               0  2017-01-01-05

Onshore numeric stats:
       year_mon_day          hour  wind_dir_avg_10  wind_speed_h_avg  \
count  7.015100e+04  70151.000000     70151.000000      70151.000000
mean   2.020569e+07     12.499836       189.005981         41.642699
std    2.292729e+04      6.922149        93.068223         20.822154
min    2.017010e+07      1.000000         0.004785          5.000000
25%    2.019010e+07      6.500000       115.963511         26.129032
50%    2.021010e+07     12.000000       205.672420         37.931034
75%    2.023010e+07     18.000000       253.215058         53.225806
max    2.025010e+07     24.000000       360.000000        183.666667

       wind_speed_avg_10  air_pressure      humidity      capacity  \
count       70151.000000  70151.000000  70151.000000  7.015100e+04
mean           41.779445  10153.316300     79.880228  9.111801e+05
std            20.859894    103.798390     14.360633  9.489638e+05
```

min	5.333333	9696.350000	20.592593	0.000000e+00
25%	26.333333	10092.250000	71.888889	2.246955e+05
50%	38.000000	10160.611111	83.846154	6.093480e+05
75%	53.333333	10222.700000	91.074074	1.198846e+06
max	185.000000	10481.400000	99.115385	4.229976e+06

	volume	percentage	emission	emissionfactor
count	7.015100e+04	70151.000000	70151.0	70151.0
mean	9.111801e+05	0.549307	0.0	0.0
std	9.489638e+05	0.439845	0.0	0.0
min	0.000000e+00	0.000000	0.0	0.0
25%	2.246955e+05	0.165114	0.0	0.0
50%	6.093480e+05	0.469071	0.0	0.0
75%	1.198846e+06	0.878007	0.0	0.0
max	4.229976e+06	1.925820	0.0	0.0

Offshore numeric stats:

	year_mon_day	hour	wind_dir_avg_10	wind_speed_h_avg	\
count	7.015100e+04	70151.000000	70151.000000	70151.000000	
mean	2.020569e+07	12.499836	191.095156	66.266291	
std	2.292729e+04	6.922149	95.840015	28.733757	
min	2.017010e+07	1.000000	0.018406	9.230769	
25%	2.019010e+07	6.500000	112.863702	44.285714	
50%	2.021010e+07	12.000000	207.332559	62.000000	
75%	2.023010e+07	18.000000	259.272495	83.333333	
max	2.025010e+07	24.000000	360.000000	227.333333	

	wind_speed_avg_10	air_pressure	humidity	capacity	\
count	70151.000000	70151.000000	70151.000000	7.015100e+04	
mean	66.353632	10146.876870	81.234217	8.270757e+05	
std	28.795023	108.672405	11.100016	8.439588e+05	
min	9.285714	9666.000000	26.428571	0.000000e+00	
25%	44.666667	10081.750000	74.142857	2.097495e+05	
50%	62.000000	10155.750000	82.857143	5.777500e+05	
75%	83.333333	10220.750000	90.142857	1.069500e+06	
max	228.000000	10469.500000	99.571429	4.342999e+06	

	volume	percentage	emission	emissionfactor
count	7.015100e+04	70151.000000	70151.0	70151.0
mean	8.270757e+05	0.481150	0.0	0.0
std	8.439588e+05	0.375839	0.0	0.0
min	0.000000e+00	0.000000	0.0	0.0
25%	2.097495e+05	0.143968	0.0	0.0
50%	5.777500e+05	0.412458	0.0	0.0
75%	1.069500e+06	0.790085	0.0	0.0
max	4.342999e+06	1.977283	0.0	0.0

[10]:
```
print("\nNumber of missing values (onshore):\n", df_on.isnull().sum())
print("\nNumber of missing values (offshore):\n", df_off.isnull().sum())
```

```
Number of missing values (onshore):
 year_mon_day        0
hour                0
wind_dir_avg_10     0
wind_speed_h_avg    0
wind_speed_avg_10   0
air_pressure        0
humidity            0
full_datetime       0
capacity            0
volume              0
percentage          0
emission            0
emissionfactor      0
correct_days        0
dtype: int64

Number of missing values (offshore):
 year_mon_day        0
hour                0
wind_dir_avg_10     0
wind_speed_h_avg    0
wind_speed_avg_10   0
air_pressure        0
humidity            0
full_datetime       0
capacity            0
volume              0
percentage          0
emission            0
emissionfactor      0
correct_days        0
dtype: int64
```

0.0.1 Ranges and Mean Values

- **Onshore:**
  - Average wind speed (`wind_speed_h_avg`) 41.6 m/s (min 5, max 183.7).
  - Average volume/power (`volume`) 9.1×10^5, with a max of 4.23×10^6.
  - Air pressure (`air_pressure`) around 10,153, ranging from 9696 to 10,481.
  - Average humidity (`humidity`) 80%.
- **Offshore:**
  - Average wind speed 66.3 m/s (min 9.2, max 227.3).
  - Average volume/power 8.27×10^5, with a max of 4.34×10^6.

- Air pressure around 10,147 (slightly lower than onshore).
- Slightly higher average humidity (81.2%).

Interpretation: - *Offshore* generally has significantly higher wind speeds (66 vs. 41). This matches the common observation that sea areas tend to have stronger winds than land. - The range of values for speed is also higher for offshore. - The maximum power/volume values are similar for both datasets, although the averages are close (9.1e5 vs. 8.27e5), indicating substantial variability.

0.0.2 Missing Values

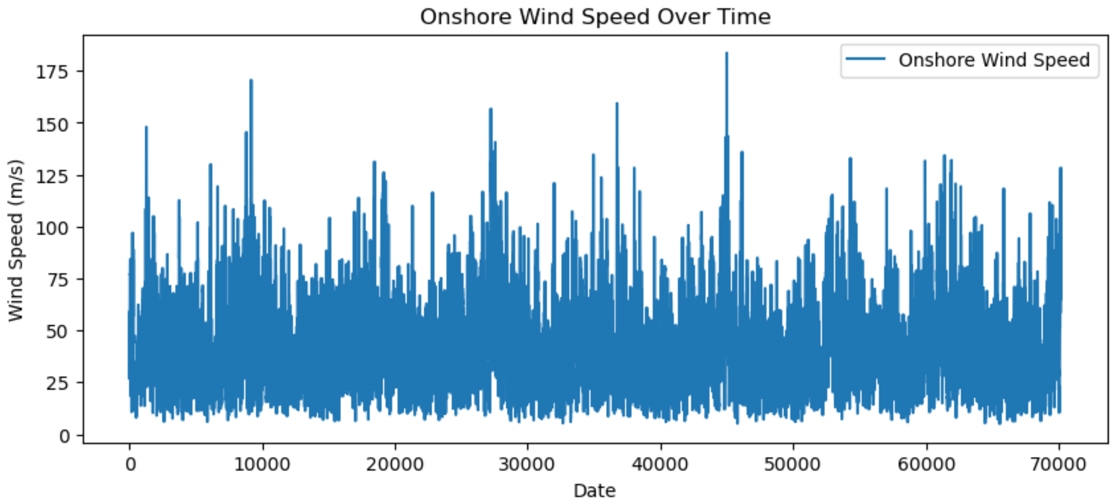
- No missing values detected (all counts match the number of rows).

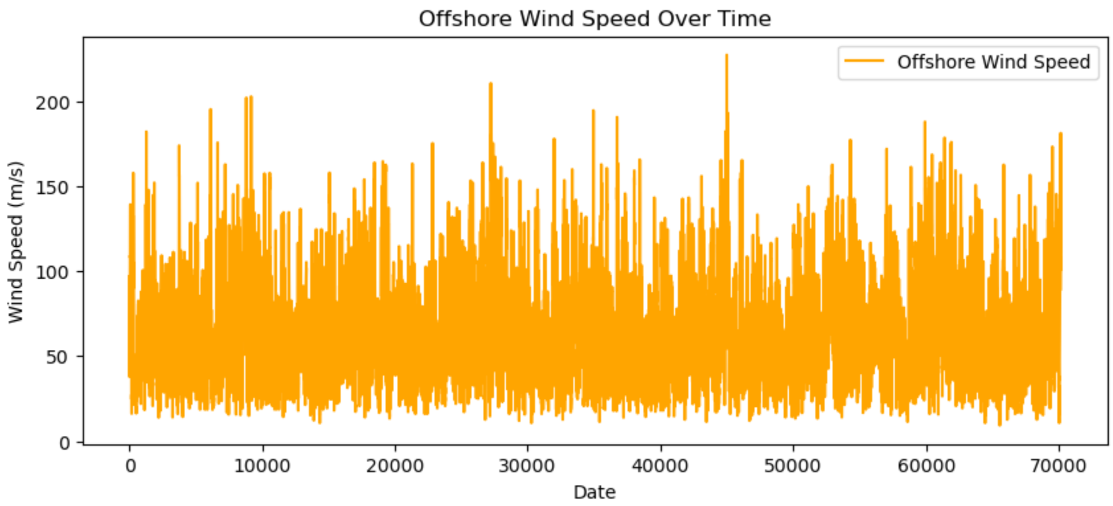
[11]:
```
df_on['full_datetime'] = pd.to_datetime(df_on['full_datetime'], errors='coerce')
df_off['full_datetime'] = pd.to_datetime(df_off['full_datetime'],␣
↪errors='coerce')
```

1 Summary Plots and Basic Visualizations

1.1 A. Time-Series Plots

[12]:
```
plt.figure(figsize=(10, 4))
plt.plot(df_on.index, df_on['wind_speed_h_avg'], label='Onshore Wind Speed')
plt.xlabel('Date')
plt.ylabel('Wind Speed (m/s)')
plt.title('Onshore Wind Speed Over Time')
plt.legend()
plt.show()

plt.figure(figsize=(10, 4))
plt.plot(df_off.index, df_off['wind_speed_h_avg'], color='orange',␣
↪label='Offshore Wind Speed')
plt.xlabel('Date')
plt.ylabel('Wind Speed (m/s)')
plt.title('Offshore Wind Speed Over Time')
plt.legend()
plt.show()
```



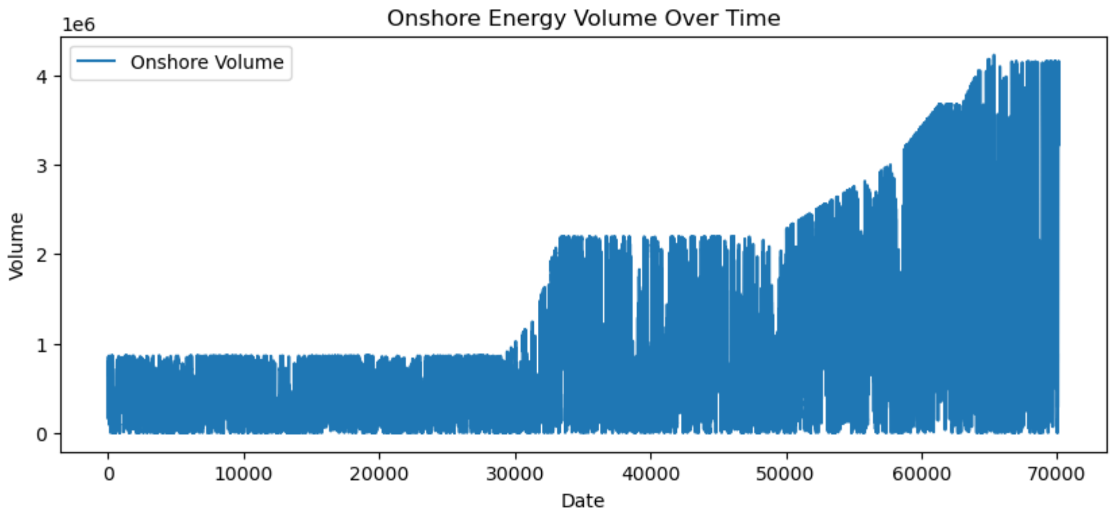


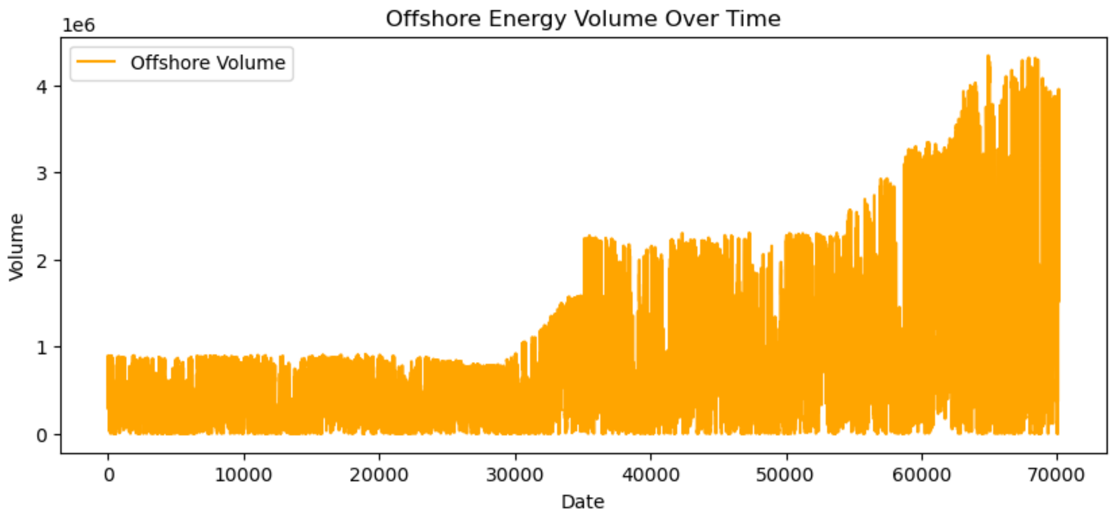
- Wind speed plots (onshore/offshore) over time show clear fluctuations throughout the period.
- The *Offshore* plot appears to be higher on the scale (i.e., higher wind speeds).
- For volume/power (onshore), significant fluctuations are also visible, sometimes reaching peaks.

Interpretation: - Wind speed is highly volatile and forms 'noise' — a typical situation for wind generation. - High volatility implies that forecasting should carefully account for seasonality, time lags, etc.

[37]:
```
plt.figure(figsize=(10, 4))
plt.plot(df_on.index, df_on['volume'], label='Onshore Volume')
plt.xlabel('Date')
plt.ylabel('Volume')
```

```
plt.title('Onshore Energy Volume Over Time')
plt.legend()
plt.show()

plt.figure(figsize=(10, 4))
plt.plot(df_off.index, df_off['volume'], color='orange', label='Offshore␣
↪Volume')
plt.xlabel('Date')
plt.ylabel('Volume')
plt.title('Offshore Energy Volume Over Time')
plt.legend()
plt.show()
```

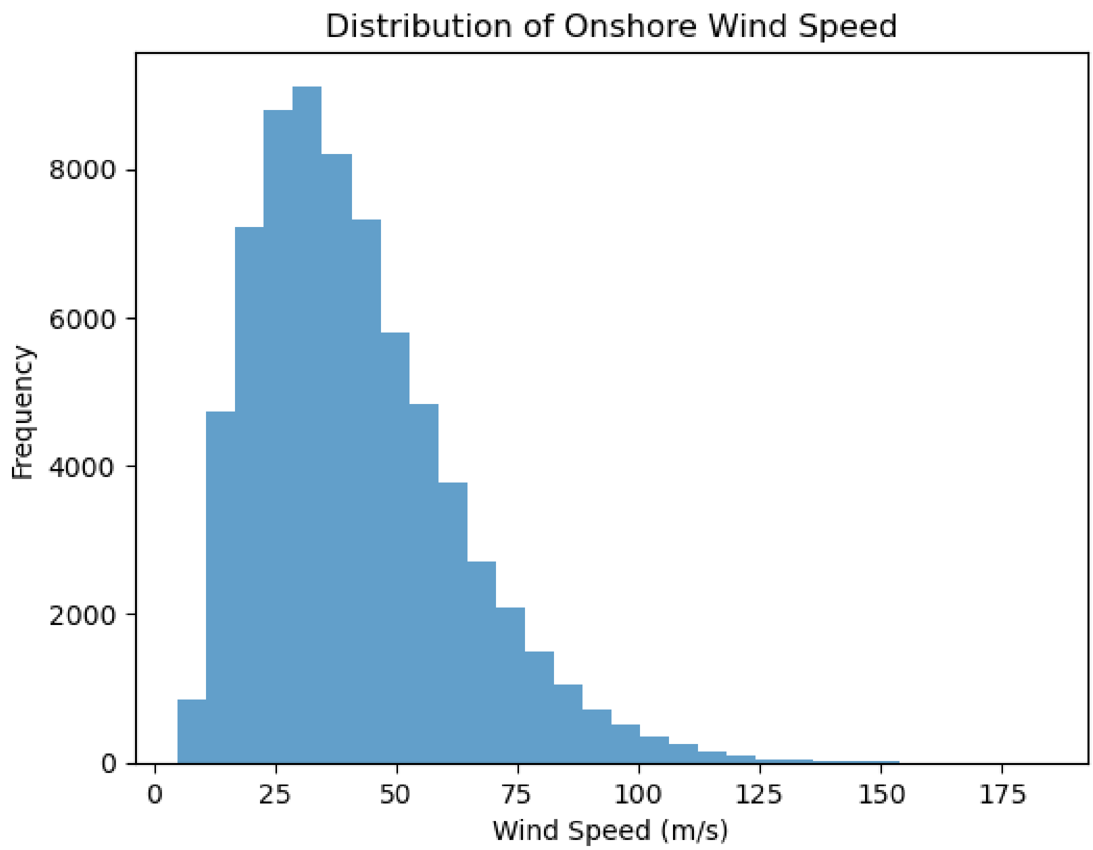


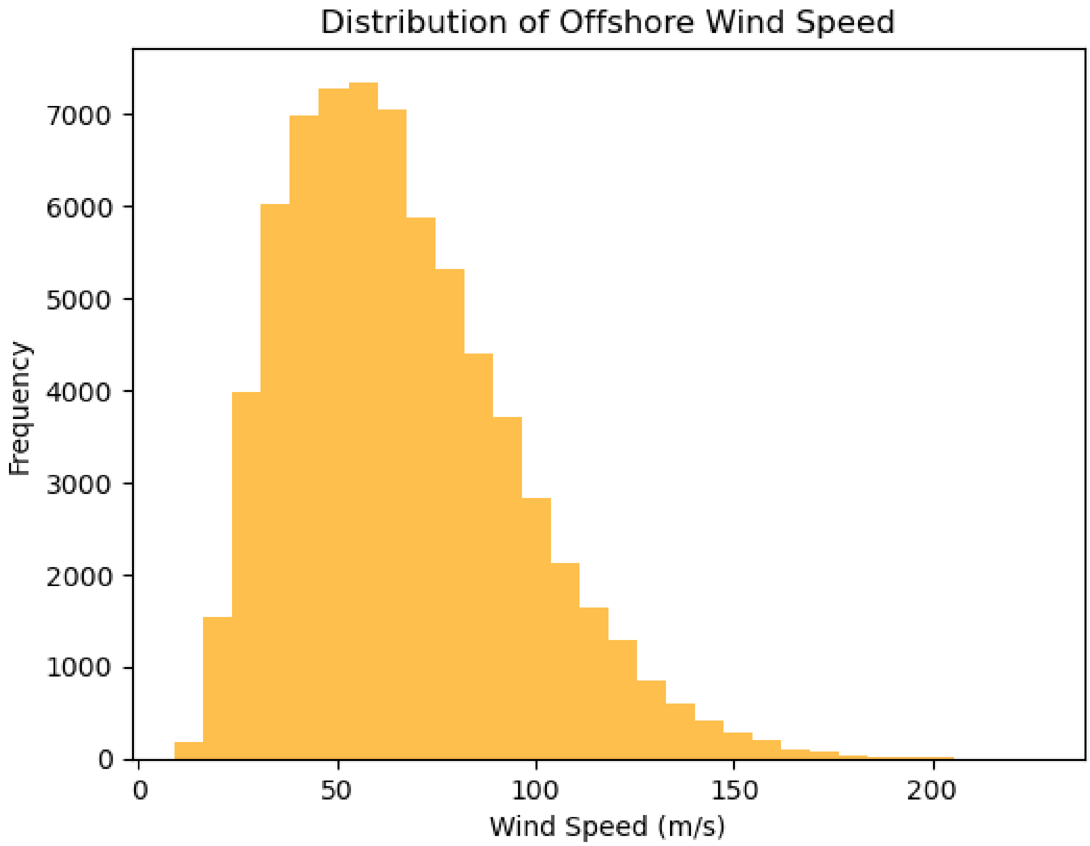


Summary: in both cases, there is a clear "stepwise" increase in the average `volume` values at later stages, accompanied by fluctuations (downward spikes). Apparently, both on land and at sea, maximum production values are reached by the end of the time series.

1.2 B. Histograms

[38]:
```
plt.hist(df_on['wind_speed_h_avg'].dropna(), bins=30, alpha=0.7)
plt.title('Distribution of Onshore Wind Speed')
plt.xlabel('Wind Speed (m/s)')
plt.ylabel('Frequency')
plt.show()

plt.hist(df_off['wind_speed_h_avg'].dropna(), bins=30, alpha=0.7,␣
 ↪color='orange')
plt.title('Distribution of Offshore Wind Speed')
plt.xlabel('Wind Speed (m/s)')
plt.ylabel('Frequency')
plt.show()
```





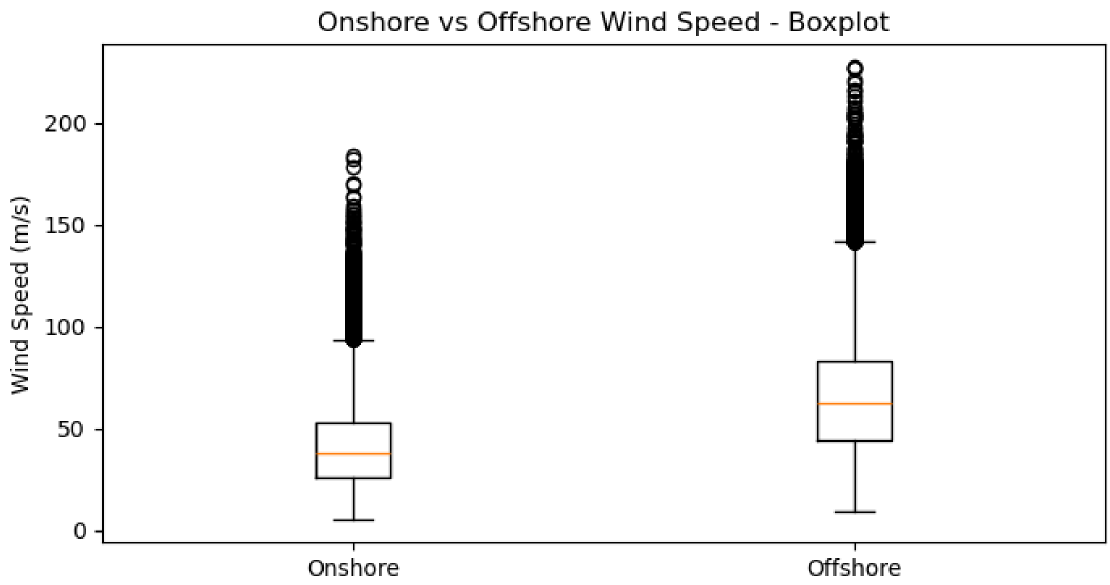
- **Onshore Wind Speed:** The distribution is concentrated around ~10–80 m/s.
- **Offshore Wind Speed:** The distribution is concentrated around ~40–100 m/s.

Interpretation: - *Offshore* wind speeds are higher

1.3 C. Boxplots

[15]:
```
plt.figure(figsize=(8,4))
plt.boxplot([df_on['wind_speed_h_avg'].dropna(), df_off['wind_speed_h_avg'].
 ↪dropna()],
            labels=['Onshore','Offshore'])
plt.ylabel('Wind Speed (m/s)')
plt.title('Onshore vs Offshore Wind Speed - Boxplot')
plt.show()
```

```
/var/folders/tq/jkwj8f8n5pq9f2q4g1mjl4r40000gn/T/ipykernel_18691/3675200272.py:2
: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been
renamed 'tick_labels' since Matplotlib 3.9; support for the old name will be
dropped in 3.11.
  plt.boxplot([df_on['wind_speed_h_avg'].dropna(),
df_off['wind_speed_h_avg'].dropna()],
```

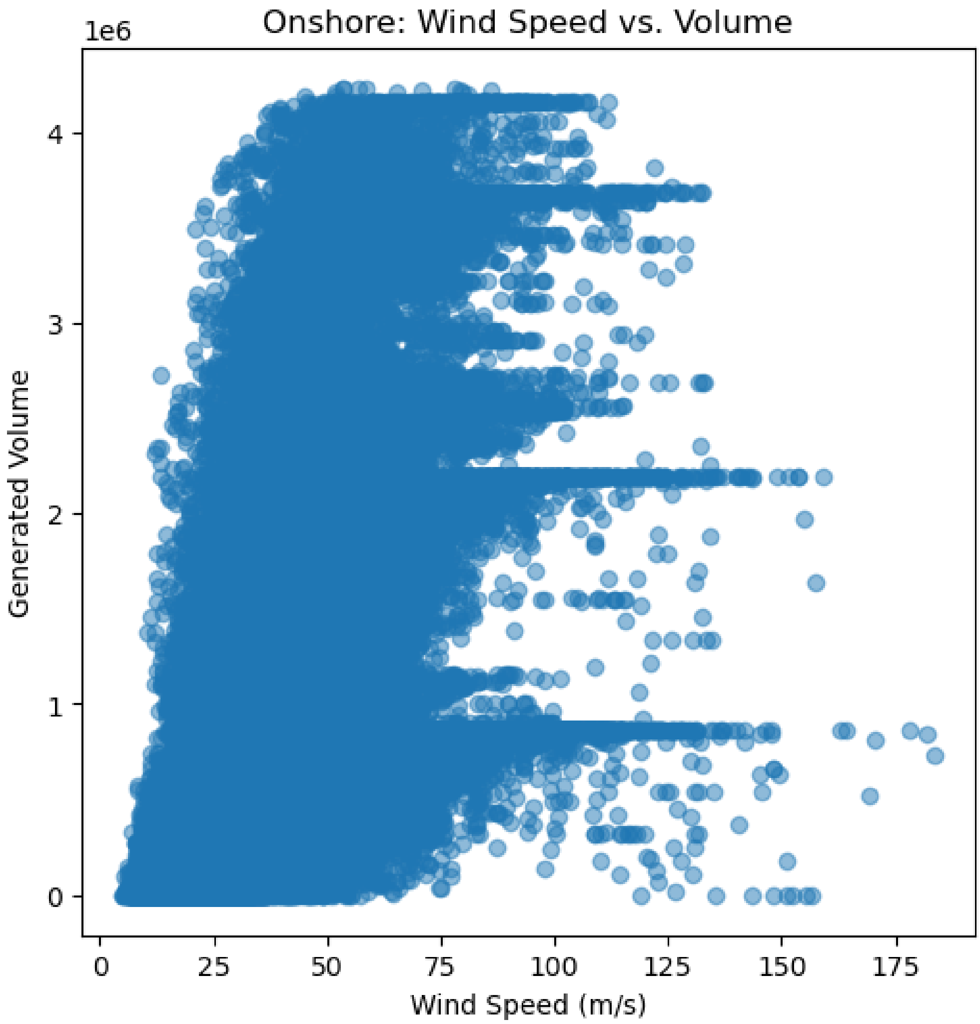


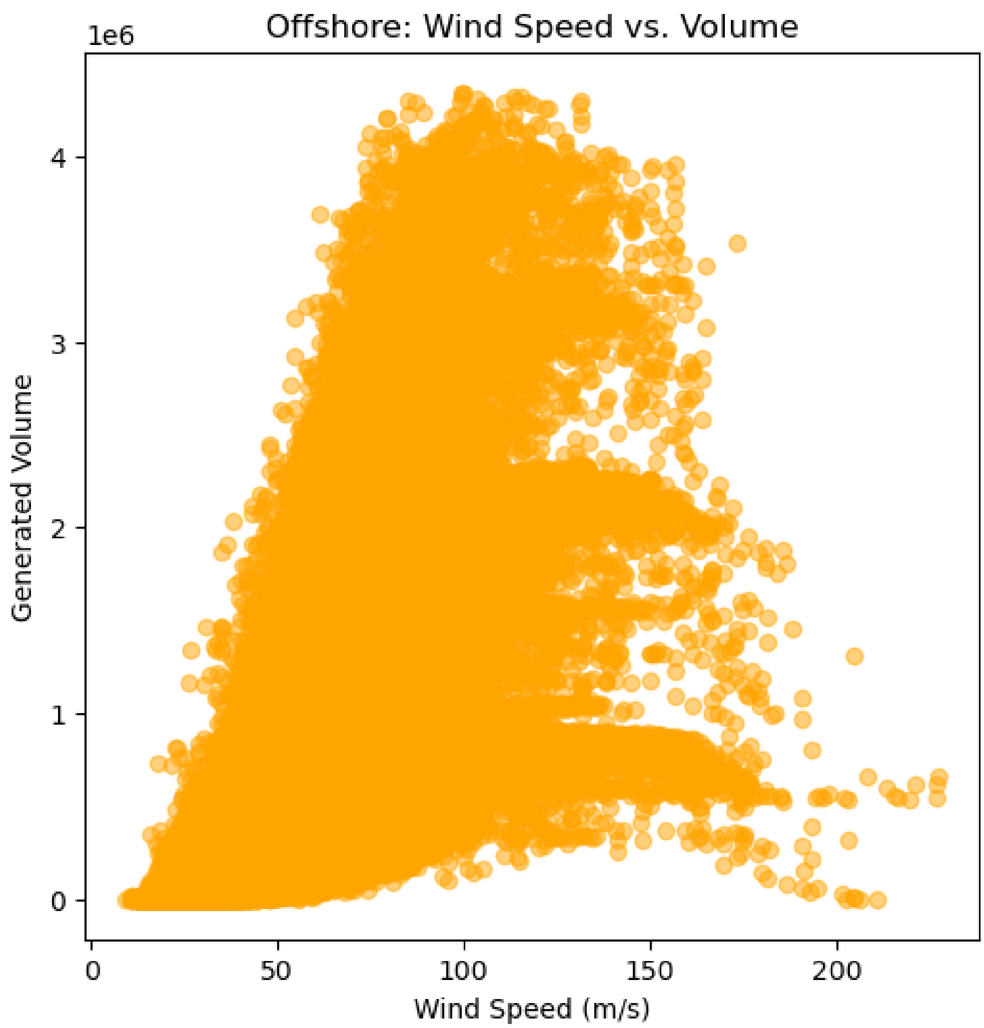
- The offshore median is clearly higher than the onshore median.
- The range (IQR) and 'whiskers' for offshore are also larger.
- There are outliers (very high speeds).

Interpretation: - Confirms the observation that *offshore* wind speeds are not only higher on average but can fluctuate over a broader range.

1.4 D. Scatter Plots

[25]:
```
plt.figure(figsize=(6,6))
plt.scatter(df_on['wind_speed_h_avg'], df_on['volume'], alpha=0.5)
plt.title('Onshore: Wind Speed vs. Volume')
plt.xlabel('Wind Speed (m/s)')
plt.ylabel('Generated Volume')
plt.show()

plt.figure(figsize=(6,6))
plt.scatter(df_off['wind_speed_h_avg'], df_off['volume'], alpha=0.5,
 ↪color='orange')
plt.title('Offshore: Wind Speed vs. Volume')
plt.xlabel('Wind Speed (m/s)')
plt.ylabel('Generated Volume')
plt.show()
```


Onshore: Wind Speed vs. Volume
1e6
Generated Volume
Wind Speed (m/s)
0
1
2
3
4
0
25
50
75
100
125
150
175



- **Onshore:** Shows a trend of increasing power with higher speed (correlation 0.52). At low speeds (up to ~20–25 m/s), the volume is low, but at speeds ranging from 50–100 m/s, substantial `volume` values are observed.
- **Offshore:** The relationship is more pronounced (correlation 0.61).

Interpretation: - The stronger the wind, the higher the `volume` (as expected), with a stronger correlation offshore. - However, at the highest speeds (above the 'nominal' threshold), power may plateau or be 'cut-off' due to turbine limitations.

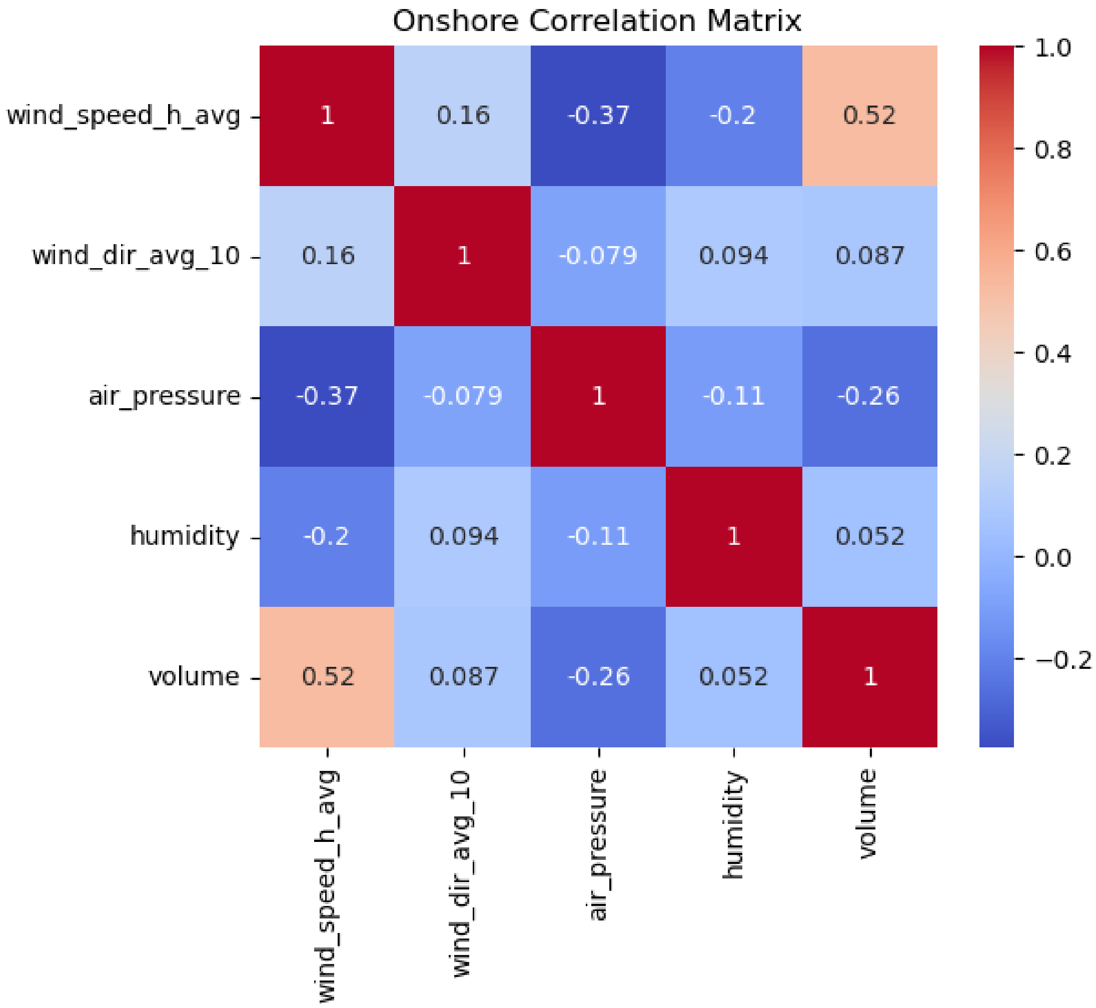
1.5 E. Correlation Analysis

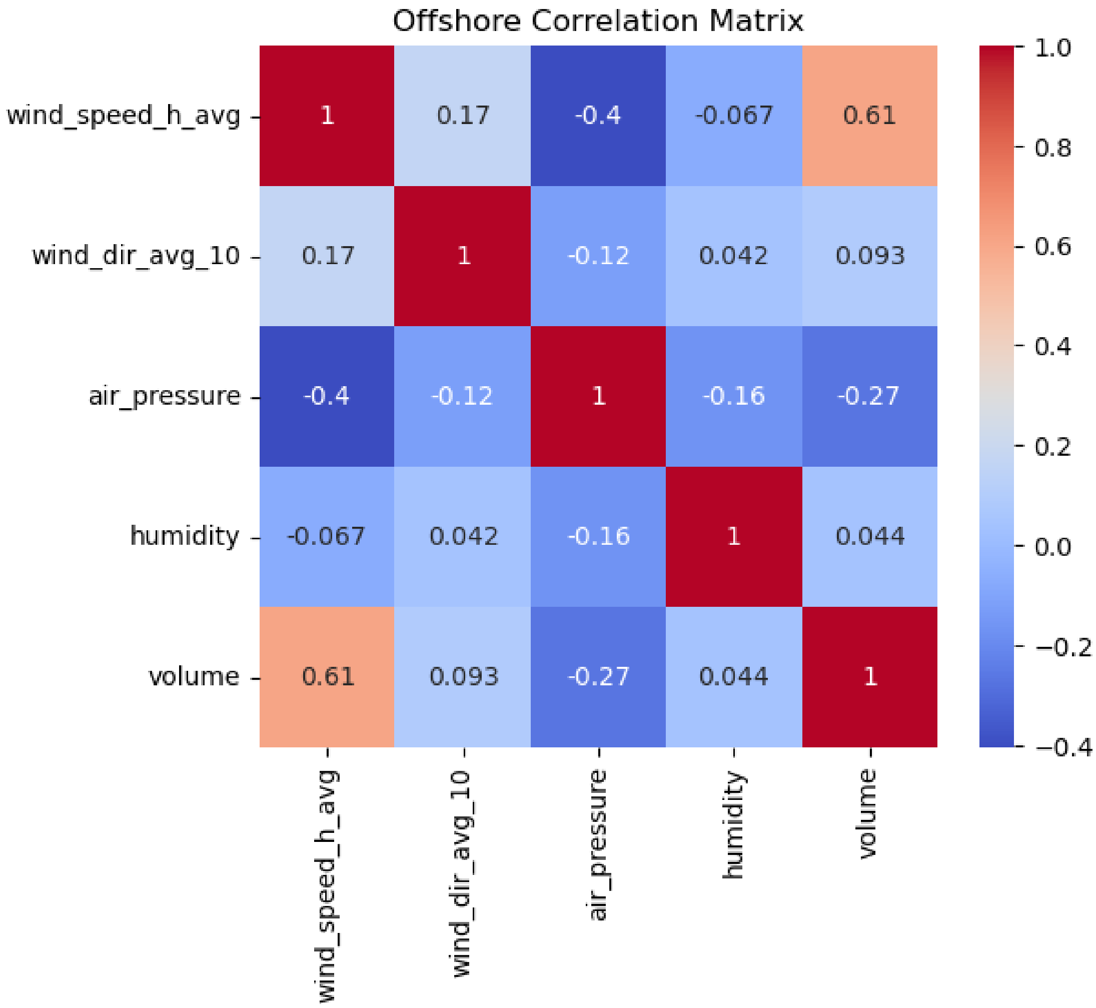
[26]:
```
corr_on = df_on[['wind_speed_h_avg', 'wind_dir_avg_10', 'air_pressure',
                 'humidity', 'volume']].corr()
print("Onshore correlation matrix:\n", corr_on)

corr_off = df_off[['wind_speed_h_avg', 'wind_dir_avg_10', 'air_pressure',
```

```
                'humidity', 'volume']].corr()
print("Offshore correlation matrix:\n", corr_off)
```

```
Onshore correlation matrix:
                  wind_speed_h_avg  wind_dir_avg_10  air_pressure  humidity  \
wind_speed_h_avg          1.000000         0.156547     -0.373753 -0.204465
wind_dir_avg_10           0.156547         1.000000     -0.078938  0.093677
air_pressure             -0.373753        -0.078938      1.000000 -0.107381
humidity                 -0.204465         0.093677     -0.107381  1.000000
volume                    0.523625         0.087356     -0.255845  0.052193

                    volume
wind_speed_h_avg  0.523625
wind_dir_avg_10   0.087356
air_pressure     -0.255845
humidity          0.052193
volume            1.000000
Offshore correlation matrix:
                  wind_speed_h_avg  wind_dir_avg_10  air_pressure  humidity  \
wind_speed_h_avg          1.000000         0.167422     -0.404537 -0.067059
wind_dir_avg_10           0.167422         1.000000     -0.121687  0.041510
air_pressure             -0.404537        -0.121687      1.000000 -0.164703
humidity                 -0.067059         0.041510     -0.164703  1.000000
volume                    0.610184         0.092795     -0.272209  0.044191

                    volume
wind_speed_h_avg  0.610184
wind_dir_avg_10   0.092795
air_pressure     -0.272209
humidity          0.044191
volume            1.000000
```

[27]:
```
plt.figure(figsize=(6,5))
sns.heatmap(corr_on, annot=True, cmap='coolwarm')
plt.title('Onshore Correlation Matrix')
plt.show()

plt.figure(figsize=(6,5))
sns.heatmap(corr_off, annot=True, cmap='coolwarm')
plt.title('Offshore Correlation Matrix')
plt.show()
```


Onshore Correlation Matrix
wind_speed_h_avg
wind_dir_avg_10
air_pressure
humidity
volume
1
0.16
-0.37
-0.2
0.52
0.16
1
-0.079
0.094
0.087
-0.37
-0.079
1
-0.11
-0.26
-0.2
0.094
-0.11
1
0.052
0.52
0.087
-0.26
0.052
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



1.5.1 Onshore:

- `wind_speed_h_avg` `volume`: r 0.52 — Moderate positive correlation (wind speed affects volume).
- `wind_dir_avg_10` `volume`: r 0.087 — Very weak correlation (wind direction has little overall impact).
- `air_pressure` `volume`: r -0.256 — Weak/Moderate negative correlation (higher pressure, lower wind speed).
- `humidity` `volume`: r 0.052 — Very weak correlation.

1.5.2 Offshore:

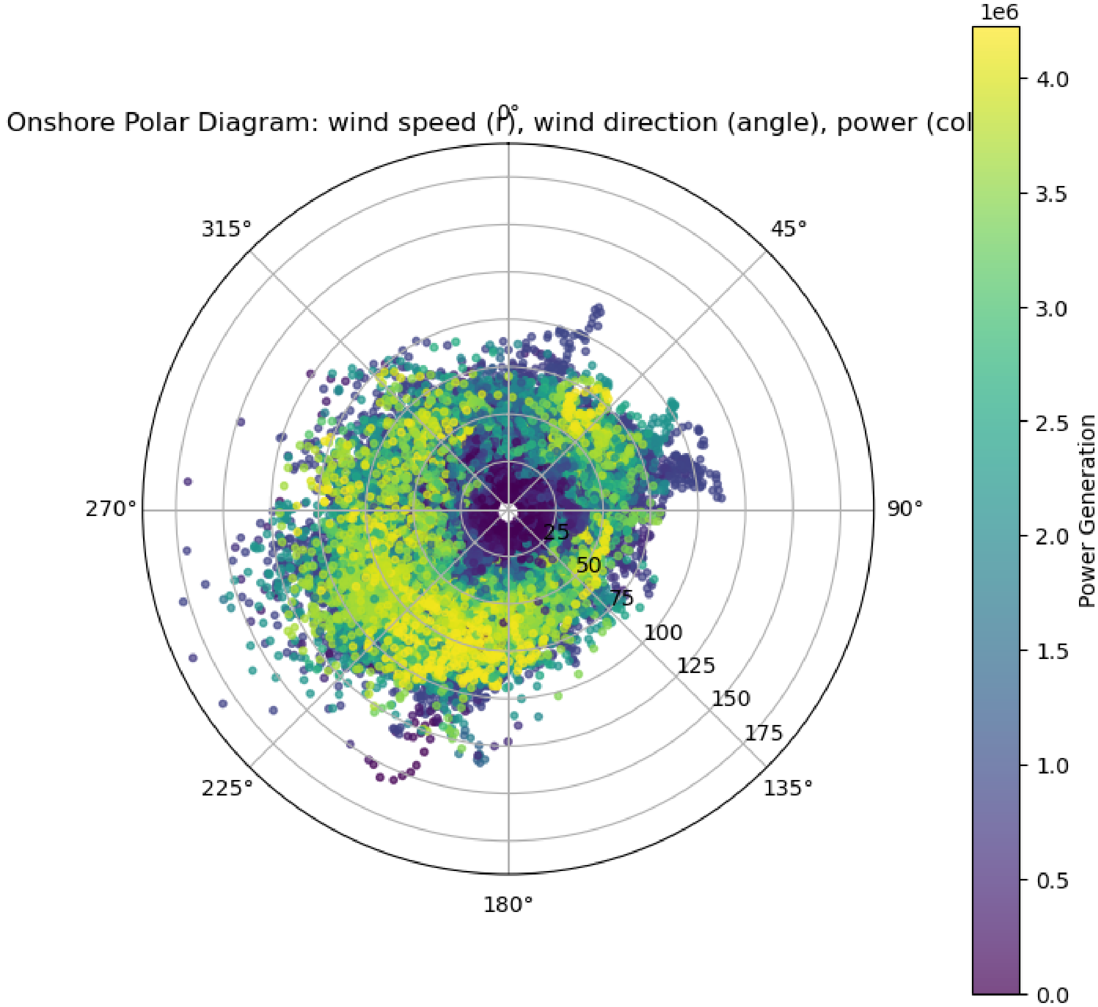
- `wind_speed_h_avg` `volume`: r 0.61 — Stronger correlation.
- `wind_dir_avg_10` `volume`: r 0.093 — Again, weak correlation.
- `air_pressure` `volume`: r -0.27 — Similar to onshore.
- `humidity` `volume`: r 0.044 — Almost zero correlation.

Main takeaway: Wind speed is the key predictor of power, with a stronger relationship offshore (0.61 vs. 0.52). Wind direction and humidity have minimal impact.

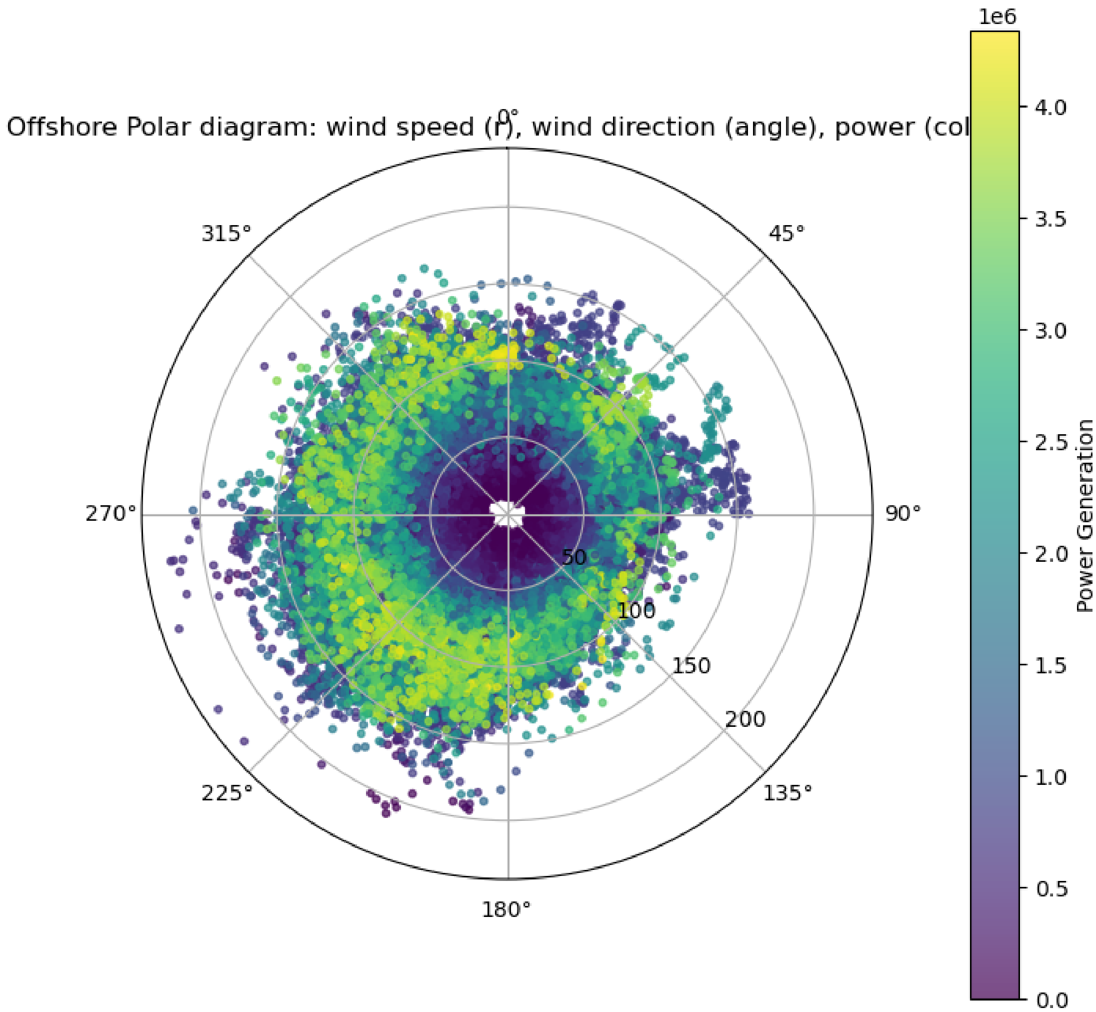
1.6 F. Polar or Windrose Plots

[40]:
```
direction_col = 'wind_dir_avg_10'
speed_col     = 'wind_speed_h_avg'
power_col     = 'volume'

# Convert wind direction from degrees to radians
theta_on = np.deg2rad(df_on[direction_col].values)

# Radius = wind speed
r_on = df_on[speed_col].values

# Color = power generation
c_on = df_on[power_col].values

plt.figure(figsize=(8,8))
ax_on = plt.subplot(111, projection='polar')

sc_on = ax_on.scatter(
    theta_on,          # angle
    r_on,              # radius
    c=c_on,            # point color representing power generation
    s=10,              # marker size (adjust if needed)
    cmap='viridis',    # color palette (viridis, plasma, jet, etc.)
    alpha=0.7          # transparency
)

# Color legend
cbar_on = plt.colorbar(sc_on, pad=0.1)
cbar_on.set_label("Power Generation")

# Set 0° at the top and clockwise angle counting
ax_on.set_theta_zero_location('N')
ax_on.set_theta_direction(-1)

# Adjust radius labels to avoid overlapping
ax_on.set_rlabel_position(135)

plt.title("Onshore Polar Diagram: wind speed (r), wind direction (angle), power␣
 ↪(color)")
plt.show()

direction_col = 'wind_dir_avg_10'
speed_col     = 'wind_speed_h_avg'
```

```python
power_col      = 'volume'

# Angle in radians: wind_dir_avg_10 usually ranges from [0..360]
theta = np.deg2rad(df_off[direction_col].values)

# Radius = wind speed:
r = df_off[speed_col].values

# Color = power (or another indicator)
c = df_off[power_col].values

plt.figure(figsize=(8,8))
ax = plt.subplot(111, projection='polar')

# Create a scatter plot on a polar projection
sc = ax.scatter(
    theta,           # angle (theta)
    r,               # radius (speed)
    c=c,             # color coding based on power
    s=10,            # marker size (adjust as needed)
    cmap='viridis', # color palette
    alpha=0.7        # transparency
)

# Add a color bar as a legend
cbar = plt.colorbar(sc, pad=0.1)
cbar.set_label("Power Generation")

# Set "north" upwards (0° = N) and angle counting clockwise
ax.set_theta_zero_location('N')
ax.set_theta_direction(-1)

# Adjust radius labels to avoid overlapping with the plot
ax.set_rlabel_position(135)

plt.title("Offshore Polar diagram: wind speed (r), wind direction (angle),␣
 ↪power (color)")
plt.show()
```



Onshore Polar Diagram: wind speed (r), wind direction (angle), power (col
0°
45°
90°
135°
180°
225°
270°
315°
25
50
75
100
125
150
175
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Power Generation



If the *onshore* polar plot shows more "yellow" points, this may indicate that the dataset more frequently contains combinations of wind speed and direction that result in power levels at the upper end of the scale (closer to 3–4 × 10ˆ6). That is, *onshore* turbines, according to the data, often reach high levels of power generation.

1.6.1 Onshore Polar Scatter

- **Many yellow points** indicate areas of high power values.

- Specifically, at various angles (180°–250° and other sectors), if the speed (radius) is high, the power reaches 3–4 × 10ˆ6.

1.6.2 Offshore Polar Scatter

- The colors are more frequently "green," which suggests either more consistent speeds.
- The diagram shows that even at different directions (not only within one sector), high speeds and high power are achieved.

Interpretation
- The diagrams allow us to see which angles and wind speeds result in the highest power generation.
- Strong winds come from various directions, and the color scale shows that the further from the center (higher speed), the higher the power.
- If there is no specific "main" direction, points are relatively evenly distributed around the circle, and the primary driver of power generation is wind speed.

1.7 Key Conclusions

1. **Offshore** has higher mean wind speeds (66 vs. 41 m/s) and a stronger link to power output (r ~ 0.61 vs. 0.52 onshore).
2. **Wind speed distribution:** broader and higher offshore, meaning more powerful and volatile conditions.
3. **Polar scatter plots** confirm that higher speeds (farther radius) yield higher power (brighter colors), regardless of wind direction.
4. **Wind direction** has almost negligible impact on total volume (r ~ 0.09), **Air pressure** inversely correlates with power generation (weak/moderate), **Humidity** barely affects power output (r ~ 0.05).

1.8 Next Step: build predictive models with wind speed as the main predictor.