Bericht zur Datenanalyse für die Leistungsprognose eines Windparks

1. Einleitung

Dieser Bericht beschreibt die Analyse eines Datensatzes, der prognostische meteorologische Parameter und die gemessene Leistungserzeugung eines Windparks enthält. Das Hauptziel war die Untersuchung der Daten zur Vorbereitung auf die Erstellung eines Leistungsprognosemodells. Die Analyse konzentrierte sich auf die Datenprüfung, -aufbereitung und die Identifizierung potenzieller Probleme, die bei der Modellierung auftreten könnten.

2. Daten laden und erste Prüfung

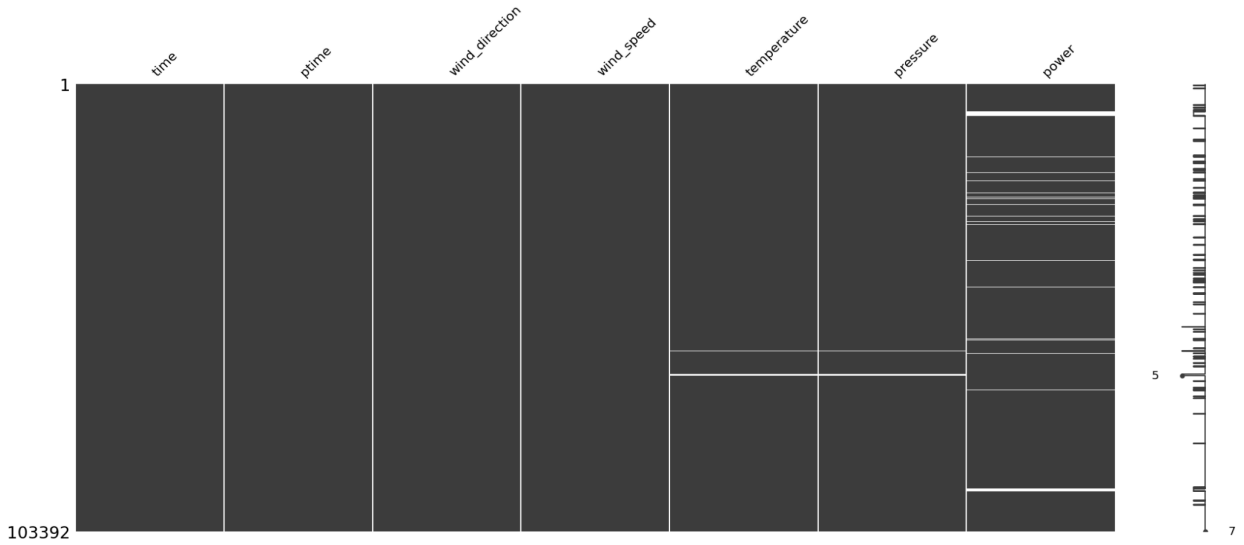
Der Datensatz, der in einer CSV-Datei ("data.csv") bereitgestellt wurde, wurde mit Python in ein Pandas DataFrame geladen. Die ersten Zeilen der Daten wurden geprüft, um ihre Struktur und ihren Inhalt zu verstehen.

3. Datenbereinigung und -aufbereitung

3.1 Fehlende Werte

Eine erste Untersuchung ergab fehlende Werte in den Spalten `temperature`, `pressure` und `power`. Insbesondere wiesen `temperature` und `pressure` das gleiche Muster fehlender Werte auf.

- `temperature`: 576 fehlende Werte
- `pressure`: 576 fehlende Werte
- `power`: 3658 fehlende Werte



Da die fehlenden Daten (Druck und Temperatur) nicht zufällig sind, habe ich beobachtet, dass die Temperatur und der Druck das gleiche Muster fehlender Werte aufweisen. Daher würde ich in diesem Fall keine Mittelwert- oder Median-Imputation bevorzugen.

Stattdessen würde ich mich für die lineare Interpolationstechnik entscheiden, um die fehlenden Werte aufzufüllen, da es sich bei den Daten um Zeitreihen mit einem hochfrequenten Intervall von 15 Minuten handelt. Ein wesentlicher Vorteil der Interpolation ist, dass der Trend der Daten erhalten bleibt. Darüber hinaus kann die Vorwärts- und Rückwärtsauffüllung manchmal geeignet sein, um fehlende Werte für Zeitreihendaten zu ersetzen.

Um fehlende Daten in `temperature` und `pressure` zu behandeln, wurde eine lineare Interpolation verwendet, um den Trend der Zeitreihendaten zu erhalten. Zeilen mit fehlenden `power`-Werten wurden entfernt. Dies ist wichtig, da Power die Zielvariable ist.

Anstatt das Modell jedoch nur auf der Grundlage dieser Imputationstechnik zu validieren, beabsichtige ich, seine Leistung zu bewerten, indem ich sowohl die Vorwärts- als auch die Rückwärtsfüllung zusammen mit der linearen Interpolation verwende. Dieser Ansatz wird dazu beitragen, die effektivste Methode für den Umgang mit fehlenden Werten im Datensatz zu ermitteln.

3.2 Datentypkonvertierung und Resampling

Die Spalten `time` und `ptime` wurden in Datetime-Objekte konvertiert. Das DataFrame wurde dann stündlich neu abgetastet, um die Daten zu vereinfachen und das Rauschen zu reduzieren.

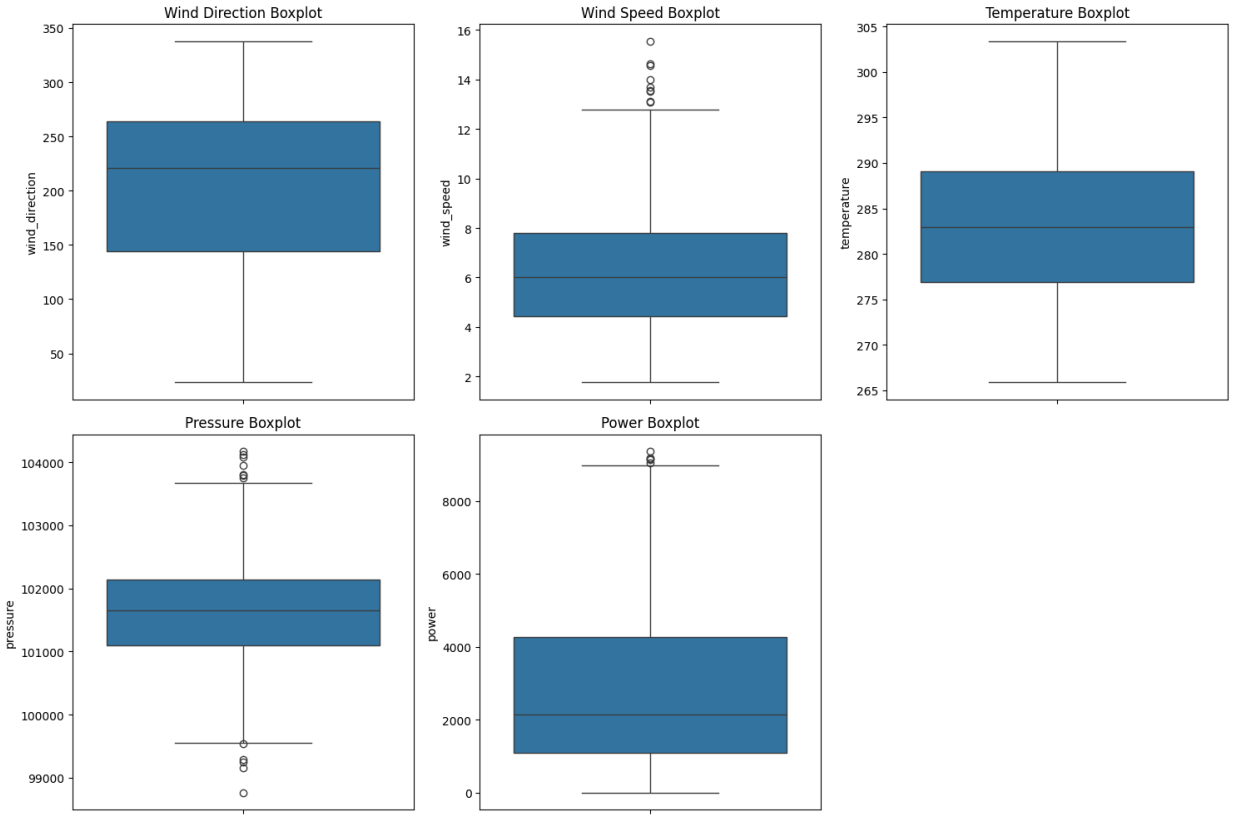
4. Explorative Datenanalyse

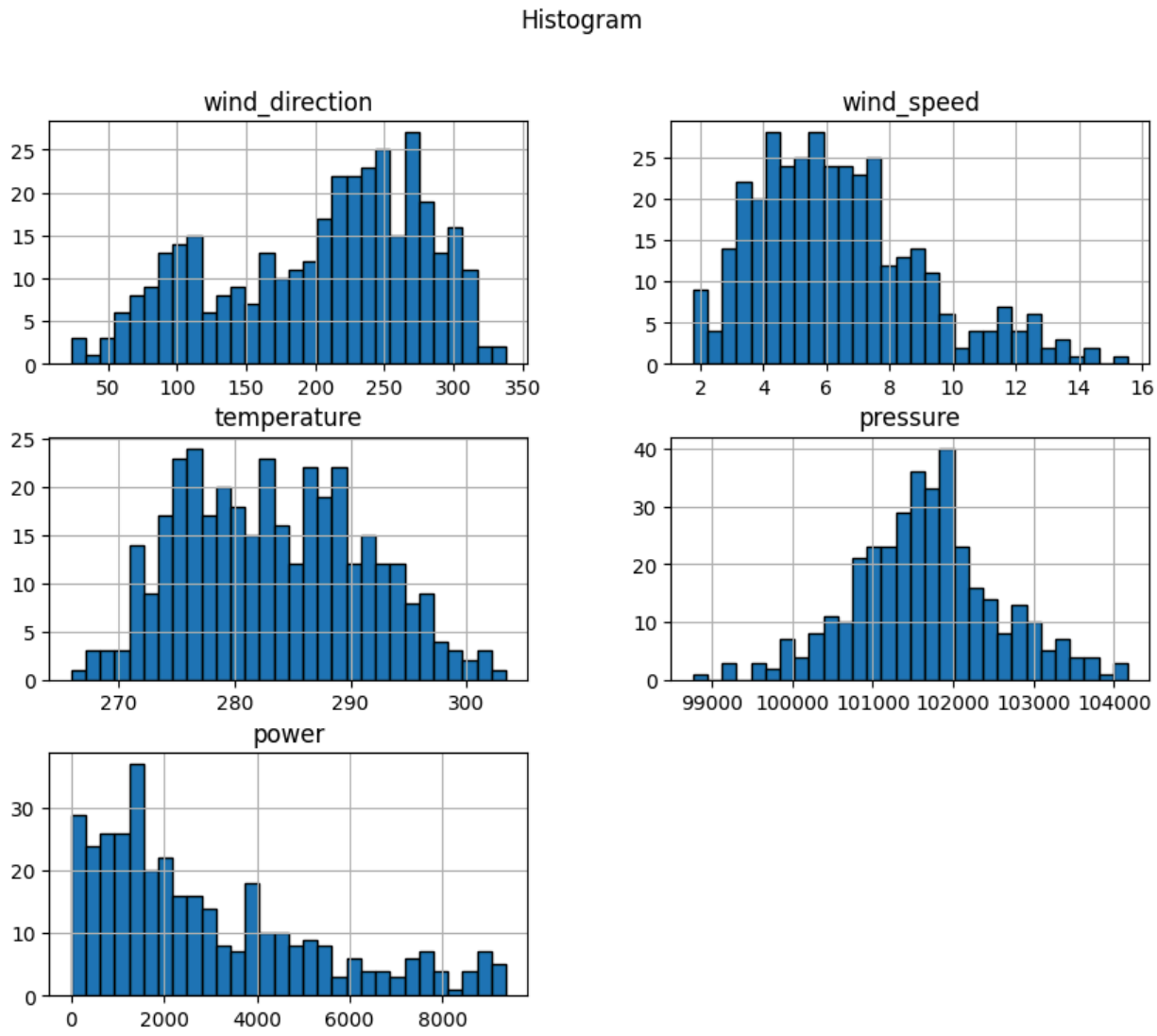
4.1 Deskriptive Statistiken und Verteilungen

Boxplots wurden erstellt, um die Verteilung von `wind_direction`, `wind_speed`, `temperature`, `pressure` und `power` zu visualisieren. Histogramme wurden ebenfalls geplottet, um die Verteilung jeder Variablen weiter zu verstehen.

Beobachtungen zu Verteilungen:

- `Wind Direction`: Scheint eine multimodale Verteilung zu haben, nicht normal.
- `Wind Speed`: Rechtsschief (langer Schwanz nach rechts).
- `Temperature`: Annähernd symmetrisch, aber keine perfekte Normalverteilung.
- `Pressure`: Fast normal, aber leicht schief.
- `Power`: Stark rechtsschief (nicht normal).





4.2 Ausreißeranalyse

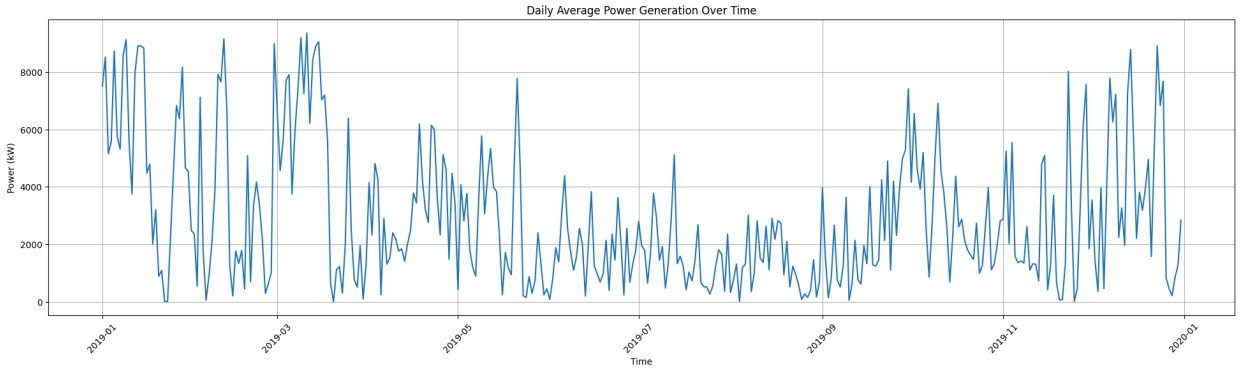
Ausreißer wurden mit der Interquartilsbereichs-Methode (IQR) für `wind_speed`, `pressure` und `power` erkannt.

- `Wind Speed`: 9 Ausreißer (2,47 %)
- `Pressure`: 12 Ausreißer (3,29 %)
- `Power`: 5 Ausreißer (1,37 %)

Es gab 2 Fälle, in denen Ausreißer von Power und Windgeschwindigkeit den gleichen Index hatten. Alle Ausreißer wurden aus dem Datensatz entfernt.

4.3 Zeitreihenanalyse

Ein Liniendiagramm von `power` über die Zeit wurde erstellt, um das zeitliche Muster der Leistungserzeugung zu visualisieren. Dieses Diagramm zeigte eine saisonale Abhängigkeit der Leistungserzeugung.

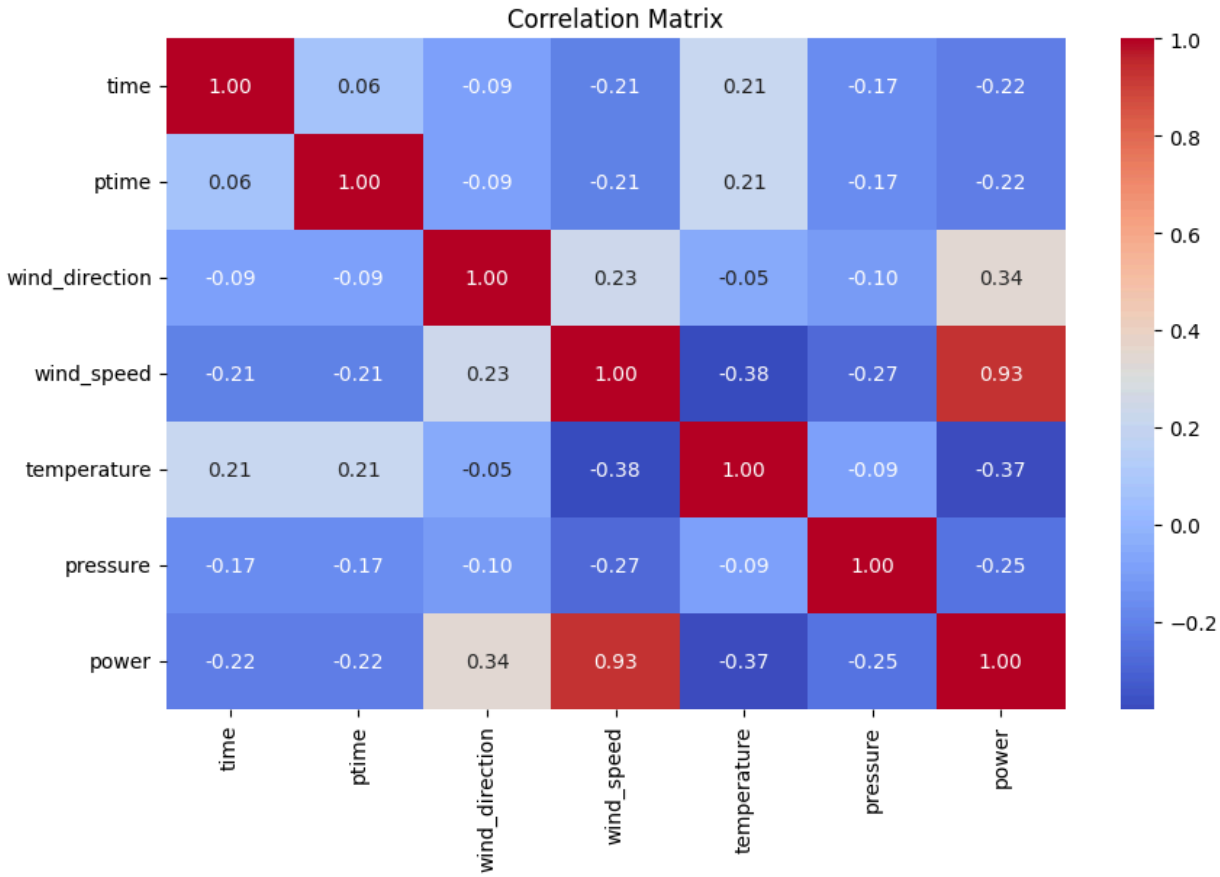


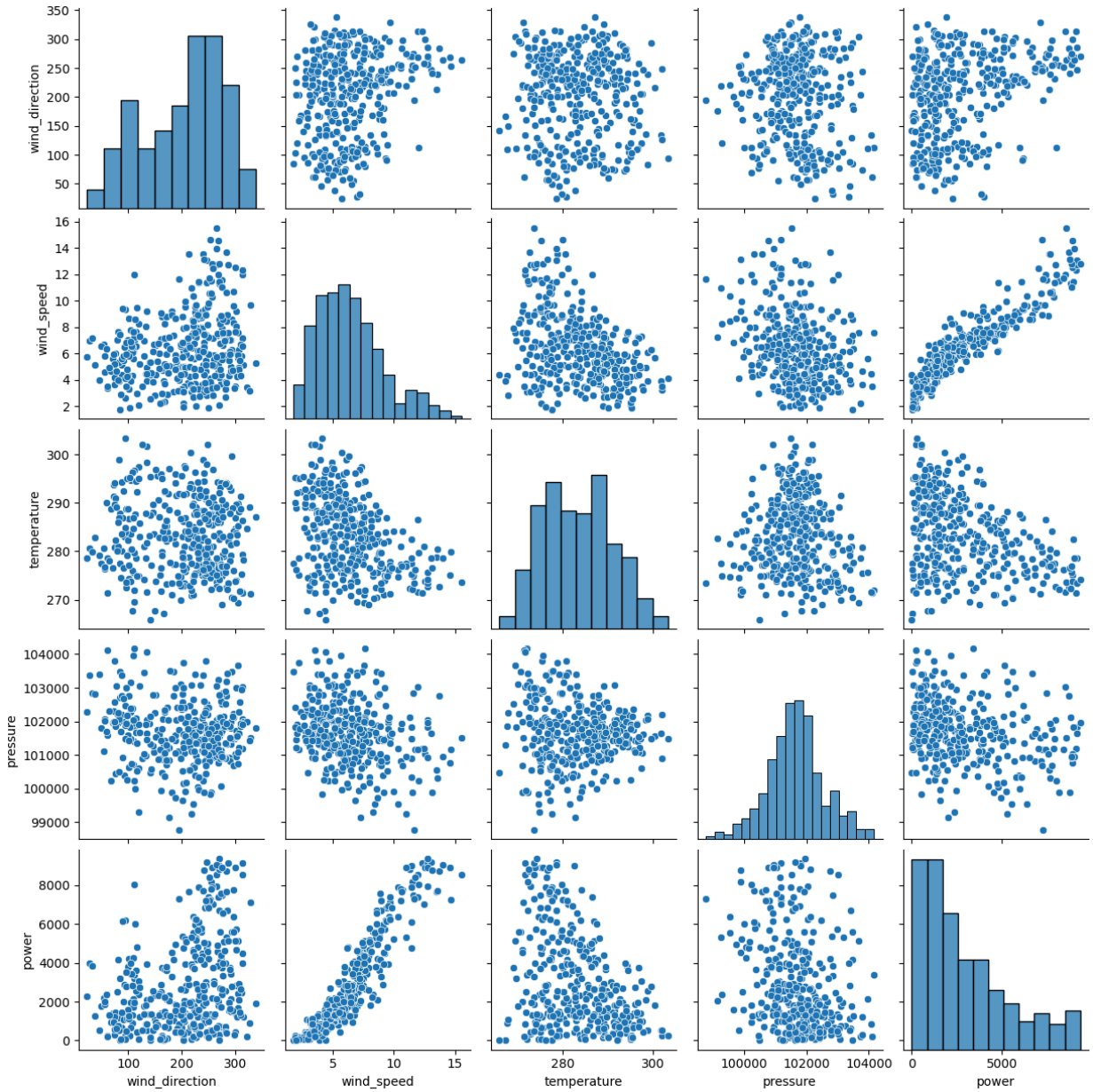
4.4 Korrelationsanalyse

Ein Pairplot wurde erstellt, um die Beziehungen zwischen allen Variablen zu visualisieren. Eine Korrelations-Heatmap wurde ebenfalls erstellt, um die Korrelationskoeffizienten zwischen Variablen anzuzeigen. Die Korrelationsmatrix wurde gedruckt, um Korrelationswerte anzuzeigen.

Wichtige Korrelationen:

- `Power` und `Wind Speed` zeigen eine starke positive Korrelation (0,93).
- `Temperature` und `Wind Speed` zeigen eine negative Korrelation (-0,38).


Correlation Matrix
time
ptime
wind_direction
wind_speed
temperature
pressure
power
1.00 0.06 -0.09 -0.21 0.21 -0.17 -0.22
0.06 1.00 -0.09 -0.21 0.21 -0.17 -0.22
-0.09 -0.09 1.00 0.23 -0.05 -0.10 0.34
-0.21 -0.21 0.23 1.00 -0.38 -0.27 0.93
0.21 0.21 -0.05 -0.38 1.00 -0.09 -0.37
-0.17 -0.17 -0.10 -0.27 -0.09 1.00 -0.25
-0.22 -0.22 0.34 0.93 -0.37 -0.25 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



5. Merkmalstechnik

- **Zeitabhängige Merkmale:** Es wurden zeitbasierte Merkmale erstellt, einschließlich Wochentag, Monat und Stunde, um saisonale Trends zu erfassen.
- **Zyklische Kodierung:** Konvertierung von wind_direction in sin- und cos-Komponenten zur Erfassung der periodischen Richtungsabhängigkeit.

6. Potenzielle Probleme und Überlegungen für die Modellierung

- **Nicht-normale Verteilungen:** Viele der Variablen, einschließlich der Zielvariable `power`, sind nicht normalverteilt. Dies kann die Verwendung nichtparametrischer Modelle oder Datentransformationen erforderlich machen.
- **Ausreißer:** Das Vorhandensein von Ausreißern kann die Modellleistung beeinträchtigen. Es ist wichtig, Ausreißer angemessen zu behandeln, wie in dieser Analyse geschehen. Es ist auch sehr wichtig festzustellen, ob die Ausreißer auf Fehlfunktionen des Sensors oder auf tatsächliche seltene Ereignisse zurückzuführen sind.
- **Saisonalität:** Das saisonale Muster der Leistungserzeugung muss im Modell berücksichtigt werden. Zeitabhängige Merkmale oder saisonale Zerlegungstechniken können nützlich sein.
- **Fehlende Daten:** Obwohl fehlende Werte durch Interpolation behandelt wurden, ist es wichtig, die Ursache für das Fehlen von Daten und die Frage zu verstehen, ob dadurch Verzerrungen eingeführt werden.
- **Korrelation:** Die starke Korrelation zwischen `power` und `wind_speed` deutet darauf hin, dass `wind_speed` ein wichtiger Prädiktor sein wird. Andere Variablen können jedoch auch zur Genauigkeit des Modells beitragen.

7. Zusätzliche Daten für eine bessere Prognosemodellierung

Um die Genauigkeit des Leistungsprognosemodells zu verbessern, wären die folgenden zusätzlichen Daten von Vorteil:

- **Wartungsprotokolle der Windkraftanlage:** Daten zu Anlagenwartung, Störungen und Reparaturen können helfen, Anomalien in der Leistungserzeugung zu erklären und die Modellgenauigkeit zu verbessern, indem Ausfallzeiten oder eine verringerte Effizienz berücksichtigt werden.
- **Zusätzliche meteorologische Parameter:** Die Einbeziehung anderer relevanter meteorologischer Variablen wie Luftfeuchtigkeit, Luftdichte und Turbulenzintensität könnte einen umfassenderen Überblick über die Bedingungen ermöglichen, die die Leistungserzeugung beeinflussen.

- **Räumliche Winddaten:** Wenn der Windpark ein großes Gebiet abdeckt, könnten Windgeschwindigkeits- und -richtungsdaten an mehreren Standorten innerhalb des Parks helfen, die räumliche Variabilität zu erfassen.

- **Hochauflösende Zeitdaten:** Noch höher aufgelöste Daten (z. B. in Sekunden) könnten kurzfristige Schwankungen von Wind und Leistung erfassen und zu genaueren Kurzfristprognosen führen.

Die Behebung der in diesem Bericht hervorgehobenen Probleme und die Einbeziehung zusätzlicher relevanter Daten werden dazu beitragen, ein genaueres und zuverlässigeres Modell zu erstellen.