EDA for Question 1: Solar Power Generation Forecasting (Python)

Carlos Peralta

2025-08-05

Introduction

This document performs an Exploratory Data Analysis (EDA) for Question 1 of the case study, which focuses on forecasting solar power generation in Germany. We will analyze the provided datasets to understand their structure, identify patterns, and find correlations that will be useful for building a predictive model.

Setup

Loading the necessary libraries for data manipulation, visualization, and analysis.

```
import pandas as pd
import plotly.express as px
from skimpy import skim
import warnings
warnings.filterwarnings("ignore")
```

Data Loading and Preparation

We load the two datasets relevant to Question 1: - `germany_atm_features_q1.csv`: Meteorological data. - `germany_solar_observation_q1.csv`: Solar power generation data.

These datasets are then merged into a single dataframe for easier analysis.

```
atm_q1 = pd.read_csv("../data/germany_atm_features_q1.csv", parse_dates=['DateTime'])
solar_q1 = pd.read_csv("../data/germany_solar_observation_q1.csv", parse_dates=['DateTime'])

data_q1 = pd.merge(solar_q1, atm_q1, on="DateTime")
```

Initial Data Exploration

Let's get a summary of the combined dataset.

```python
skim(data_q1)
```

```
                          skimpy summary
          Data Summary                Data Types

   Dataframe           Values    Column Type   Count

   Number of rows      29928     float64       11
   Number of columns   12        datetime64    1

                                                          number

   column                  NA    NA %    mean       sd        p0       p25     p50     p75     p1

   power                     0       0       6874     10580        0        4   192.6   10980   48
   surface_solar_radiati     0       0      131.7     195.4        0        0    7.01   214.7   85
   on_downwards
   temperature_2m            0       0      10.38     7.512   -9.185     4.61   9.795   15.93   34
   total_cloud_cover         0       0     0.6726    0.2539        0     0.51    0.73    0.88
   total_precipitation       0       0    0.09412    0.1447        0        0    0.03    0.12    1
   snowfall                  0       0   0.005298   0.02321        0        0       0       0    0
   snow_depth                0       0     0.3566     1.215        0        0       0    0.09   14
   wind_speed_10m            0       0      3.331     1.383    0.735     2.31    3.02    4.07   10
   wind_speed_100m           0       0      5.712     2.254        1     4.08   5.315   6.975    1
   apparent_temperature      0       0       9.01     8.756   -14.39     1.85    8.85   15.92   33
   relative_humidity_2m      0       0         76     14.94    20.41    67.39   80.47    87.6    9

                                                         datetime

   column            NA      NA %       first               last                        f

   DateTime              0          0      2022-01-01           2025-05-31 23:00:00        h

                               End
```

Time Series Visualization

Visualizing the solar power generation and key meteorological features over time.

Solar Power Generation

```
fig = px.line(data_q1, x='DateTime', y='power', title='Solar Power Generation over Time')
fig.show()
```

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

The plot shows a clear seasonal pattern, with higher generation during summer months and lower generation in winter. There is also a daily pattern where power generation peaks during the day.

Surface Solar Radiation

```
fig = px.line(data_q1, x='DateTime', y='surface_solar_radiation_downwards', title='Surface So
fig.show()
```

```
Unable to display output for mime type(s): text/html
```

Solar radiation follows a similar seasonal and daily pattern to power generation, which is expected.

Correlation Analysis

A correlation matrix will help us understand the relationships between the different variables.

```
numeric_vars = data_q1.select_dtypes(include='number')
cor_matrix = numeric_vars.corr()

fig = px.imshow(cor_matrix, title='Correlation Matrix of Meteorological Features and Power Ge
fig.show()
```

```
Unable to display output for mime type(s): text/html
```

The heatmap shows a strong positive correlation between `power` and `surface_solar_radiation_downwards`, as well as `temperature_2m`. This confirms that solar radiation and temperature are key drivers of solar power generation.