EDA for Question 2: Detecting True Demand Growth (Python)

Carlos Peralta

2025-08-05

Introduction

This document presents an Exploratory Data Analysis (EDA) for Question 2 of the case study. The objective is to estimate the true year-over-year demand growth in Germany by accounting for the effect of behind-the-meter (BTM) solar generation.

Setup

Loading the necessary libraries for the analysis.

```
import pandas as pd
import plotly.express as px
from skimpy import skim
from statsmodels.tsa.stattools import grangercausalitytests
import numpy as np
import warnings
warnings.filterwarnings("ignore")
```

Data Loading and Preparation

We load the datasets relevant to Question 2: - `germany_electricity_demand_observation_q2.csv`: Observed electricity demand. - `germany_solar_observation_q2.csv`: Grid-scale solar power generation. - `germany_atm_features_q2.csv`: Meteorological data.

These are merged into a single dataframe.

```
demand_q2 = pd.read_csv("../data/germany_electricity_demand_observation_q2.csv", parse_dates=
solar_q2 = pd.read_csv("../data/germany_solar_observation_q2.csv", parse_dates=['DateTime'])
atm_q2 = pd.read_csv("../data/germany_atm_features_q2.csv", parse_dates=['DateTime'])

data_q2 = pd.merge(demand_q2, solar_q2, on="DateTime")
data_q2 = pd.merge(data_q2, atm_q2, on="DateTime")
```

Initial Data Exploration

Let's get a summary of the combined dataset.

```
skim(data_q2)
```

```
                              skimpy summary
          Data Summary                 Data Types

   Dataframe           Values    Column Type   Count

   Number of rows      47472     float64       12
   Number of columns   13        datetime64    1

                                                      number

   column                  NA    NA %    mean       sd         p0       p25     p50     p75     p1

   demand                   0       0      55190      9715    31280   47310   55120   62630   82
   power                    0       0       6279      9759        0       3   183.2    9892   48
   surface_solar_radiati    0       0      128.7     191.9        0       0    6.59   207.7   87
   on_downwards
   temperature_2m           0       0      10.69     7.294   -11.74   5.125   10.12   16.05   35
   total_cloud_cover        0       0     0.6746    0.2584        0   0.515    0.73    0.89
   total_precipitation      0       0     0.1002    0.1652        0       0    0.03    0.13    2
   snowfall                 0       0   0.004825   0.02432        0       0       0       0    0
   snow_depth               0       0     0.2422     1.093        0       0       0    0.01   13
   wind_speed_10m           0       0      3.381     1.408    0.715    2.33   3.065    4.13   11
   wind_speed_100m          0       0      5.805     2.295      0.9   4.125   5.425    7.13   17
   apparent_temperature     0       0      9.378     8.513   -17.02    2.43   9.315   16.05   33
   relative_humidity_2m     0       0      76.15     15.08    17.07   67.31   80.51   87.89    9

                                                     datetime
```

column	NA	NA %	first	last	f
DateTime	0	0	2020-01-01	2025-05-31 23:00:00	h

End

Time Series Visualization

Visualizing the demand and solar generation data to identify trends and patterns.

Electricity Demand

```
fig = px.line(data_q2, x='DateTime', y='demand', title='Observed Electricity Demand 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 demand shows daily and weekly cycles, as well as seasonal variations. There appears to be a dip in demand during midday, which might be caused by BTM solar generation.

Grid-Scale Solar Power

```
fig = px.line(data_q2, x='DateTime', y='power', title='Grid-Scale Solar Power Generation over
fig.show()
```

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

This plot shows the grid-scale solar generation, which does not account for rooftop solar installations.

Correlation Analysis

A correlation matrix will help us understand the relationships between the different variables in the context of demand.

```
numeric_vars = data_q2.select_dtypes(include='number')
cor_matrix = numeric_vars.corr()

fig = px.imshow(cor_matrix, title='Correlation Matrix of Demand, Solar, and Weather')
fig.show()
```

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

The correlation matrix can reveal how weather variables impact both demand and solar generation, providing clues to isolate the effect of BTM solar.

Causal Correlation Analysis

To further investigate the relationship between solar generation and demand, we can use a Granger causality test. This test helps determine if one time series is useful in forecasting another.

Granger Causality Test

We'll test if `surface_solar_radiation_downwards` Granger-causes `demand`. We'll use a lag of 6 hours.

```
# Handle missing values
data_q2_filled = data_q2[['demand', 'surface_solar_radiation_downwards']].ffill()

# Perform the Granger causality test
granger_result = grangercausalitytests(data_q2_filled, [6], verbose=True)
```

```
Granger Causality
number of lags (no zero) 6
ssr based F test:         F=1001.9321, p=0.0000  , df_denom=47453, df_num=6
ssr based chi2 test:   chi2=6013.2395, p=0.0000  , df=6
likelihood ratio test: chi2=5661.7381, p=0.0000  , df=6
parameter F test:         F=1001.9321, p=0.0000  , df_denom=47453, df_num=6
```

Interpretation

The null hypothesis of the Granger causality test is that the lagged values of the predictor variable do not add any predictive power to the model for the dependent variable. If the p-value is statistically significant (typically < 0.05), we can reject the null hypothesis and conclude that there is evidence of Granger causality.