Luleå University of Technology

D7054E - Data Science programming

18 March 2023

Group 8 | Project | Berezin Ilya, Cherevichnik Stepan

Contents

Introduction

Overview of the Nordpool electricity market.

The Nordpool electricity market is one of the largest and most important electricity markets in Europe. It covers all the Nordic countries, including Denmark, Finland, Norway, Sweden, Estonia, Latvia, and Lithuania. The market was established in 1996 and is owned by the Nord-Pool Group, a pan-European exchange for trading in the wholesale electricity market.

The Nordpool operates on a spot market basis, meaning that electricity is traded for delivery on the same day. It is divided into several bidding areas, which reflect the different production and consumption patterns within the region. The prices in each bidding area are determined by the supply and demand of electricity in that area, and are influenced by factors such as weather conditions, economic activity, and the availability of renewable energy sources.

The Nordpool is known for its innovative and progressive approach to energy trading, with a focus on promoting competition and market efficiency. It was the first electricity market in the world to introduce a real-time balancing system, which helps to ensure that supply and demand are always in balance. Additionally, the Nordpool market has been a leader in the development of renewable energy, with a goal of achieving 100% renewable electricity production in Scandinavia by 2050.

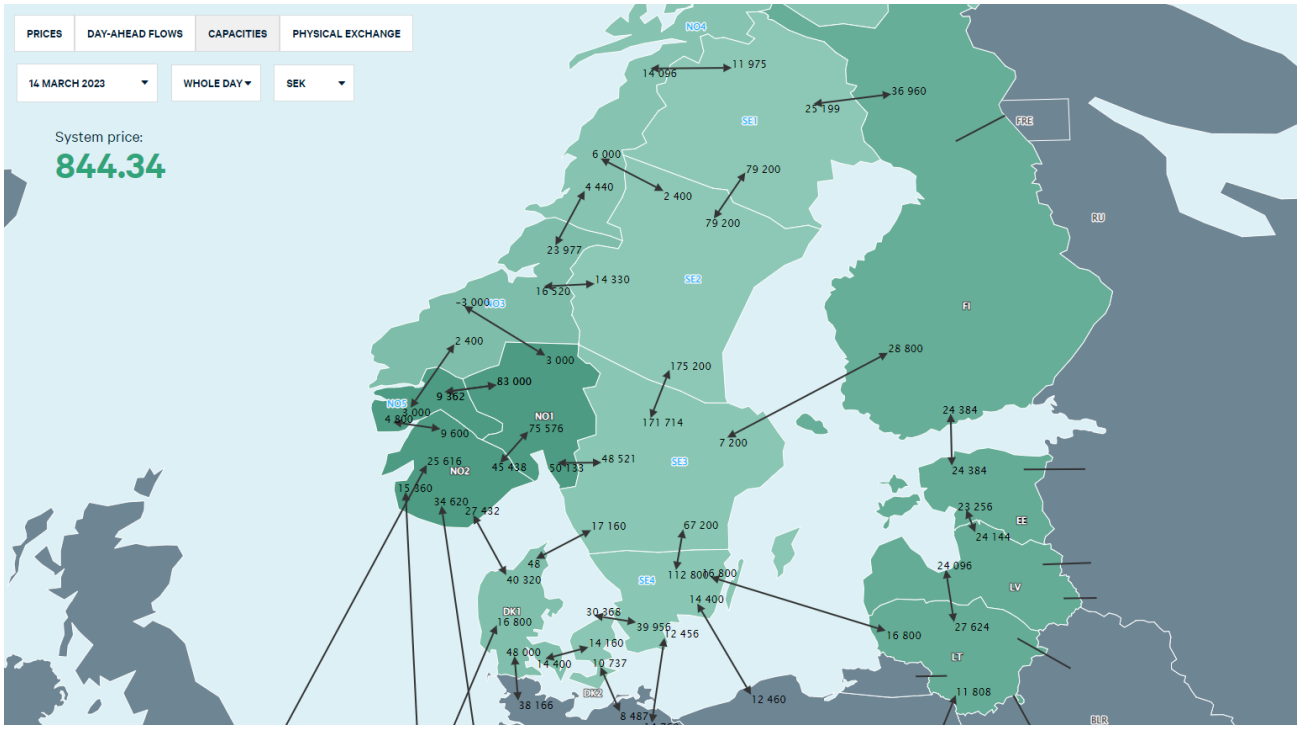


Figure 1 – Nordic Energy flow map (www.nordpoolgroup.com)

Data acquisition and Preprocessing

The source of the data, the variables included, and any missing or inconsistent data.

Data is presented in SDV files, often referred to as Excel data files since this type is primarily created or used by Excel and similar to CSV, just with added data description at the beginning, and stored in FTP folders with sometimes ambiguous structure and description. There were discovered missing values, energy markets that are named in a different way following the market development and

electricity sub-markets split / joint, wrong variables type leading to file size increase, outliers, periods with missing data etc.

Dataset	Source	File format	Files per year	Published	Data resolution
Elbas ticker	Nord Pool	CSV	365	Continuously	Continuously
Elspot file	Nord Pool	SDV	52	Day ahead at 12:42 CET	Hourly
Regulating prices	Nord Pool	SDV	52	1–2 hours after hour of power delivery	Hourly
Elspot capacity	Nord Pool	SDV	52	Day ahead around 10:00 CET	Hourly
Elbas capacity	Nord Pool	SDV	52	Day ahead around 14:00–15:00 CET	Hourly
Operating	Nord Pool	SDV	52	Day ahead around 14:00–15:00 CET	Hourly

Detail the steps taken to clean and transform the data for analysis.

The following libraries were used for data processing:

- **os**: interacting with the operating system and managing file paths.
- **datetime**: working with dates and times.
- **pandas**: data manipulation and analysis
- **regex**: regular expression matching
- **pandarallel**: parallel processing of data using Pandas.
- **gzip**: compressing the output files in the parquet format.

Ftp-based data is split into two types: operating data and spot price data.

The operating and the spot price data are stored in SDV files and contains information about energy consumption, production, and exchange. The data is organized by country and bidding area, and we process the data for countries and bidding areas of interest for the period from 2014 to 2022.

I. We read the SDV files first and extracts the relevant information, including the country, bidding area, date, and hourly values for energy consumption, production, and exchange.
II. Extracted data is then organized into a Pandas Data Frame with columns for the country, bidding area, date, hour, and value.
III. The data is cleaned, with the 'sum' and 'file' columns being dropped and the remaining columns being melted. The hour column is also cleaned by removing the trailing A or B characters and converting it to an integer. The 'price_eur_mwh' column is also converted to a float.
IV. Data Frame is filtered to remove rows with missing data and the hour column is used to adjust the date-time stamp.

V. Countries were stored using 'object' datatype which led to a non-efficient memory utilization, transforming it to a ‘category’ datatype helps us to 3x decrease the amount of used memory.

VI. Data Frame is processed to convert data types and saved as a compressed Parquet file.

Exploratory Data Analysis

Initial exploratory analysis of the variables.

Data structure was provided to us in a PDF file, where different data descriptors are also presented.

```
1 ds.head(5)
```

	country	datatype	code	date	bidding_area	value
0	Sweden	PS	WE	2014-06-09	Sverige	637.0
1	Sweden	PS	WE	2014-06-09	Sverige2	82.0
2	Sweden	PS	WE	2014-06-09	Sverige3	388.0
3	Sweden	PS	WE	2014-06-09	Sverige1	20.0
4	Sweden	PS	WE	2014-06-09	Sverige4	146.0

```
1 ds_p.head(5)
```

	country	bidding_area	date	price_eur_mwh
0	Sweden	SE3_Stockholm	2013-12-30 01:00:00	25.55
1	Sweden	SE3_Stockholm	2013-12-31 01:00:00	26.99
2	Sweden	SE3_Stockholm	2014-01-01 01:00:00	28.67
3	Sweden	SE3_Stockholm	2014-01-02 01:00:00	27.90
4	Sweden	SE3_Stockholm	2014-01-03 01:00:00	28.08

Figure 2. First rows of datasets with operational data and prices

The DataFrame's shape and data types of the columns are printed using the info () method.

```
operating_data_df.info()
```

```
 0   country       object
 1   datatype      object
 2   code          object
 3   date          datetime64[ns]
 4   bidding_area  object
 5   hour          object
 6   value         float64
dtypes: datetime64[ns](1), float64(1), object(5)
memory usage: 1.3+ GB
```

```
operating_data_df.info()
```

```
 #   Column        Dtype
---  ------        -----
 0   date          datetime64[ns]
 1   datatype      category
 2   country       category
 3   bidding_area  category
 4   code          category
 5   value         float32
dtypes: category(4), datetime64[ns](1), float32(1)
memory usage: 354.6 MB
```

Figure 3. info () and memory usage before and after data processing

```
1 ds.describe().style.background_gradient(cmap='Blues').format('{:.2f}')
```

	value
count	23590018.00
mean	1444.66
std	3943.66
min	-10473.30
25%	3.30
50%	313.35
75%	1300.00
max	65311.10

```
1 ds_p.describe().style.background_gradient(cmap='Blues').format('{:.2f}')
```

	price_eur_mwh
count	1298304.00
mean	49.51
std	63.21
min	-60.26
25%	24.42
50%	32.95
75%	48.05
max	4000.00

Figure 4. In-built statistics on data

Visualization of the trends, seasonality, and distributions of the variables.

Electricity consumption and prices can vary significantly based on the time of the year, such as higher consumption during winter months and lower consumption during summer months. We can clearly observe seasonality as a fluctuation in energy consumption, related to outside temperature change and necessity for heating. Examining these trends could reveal interesting insights into how weather and other factors affect electricity usage and pricing.

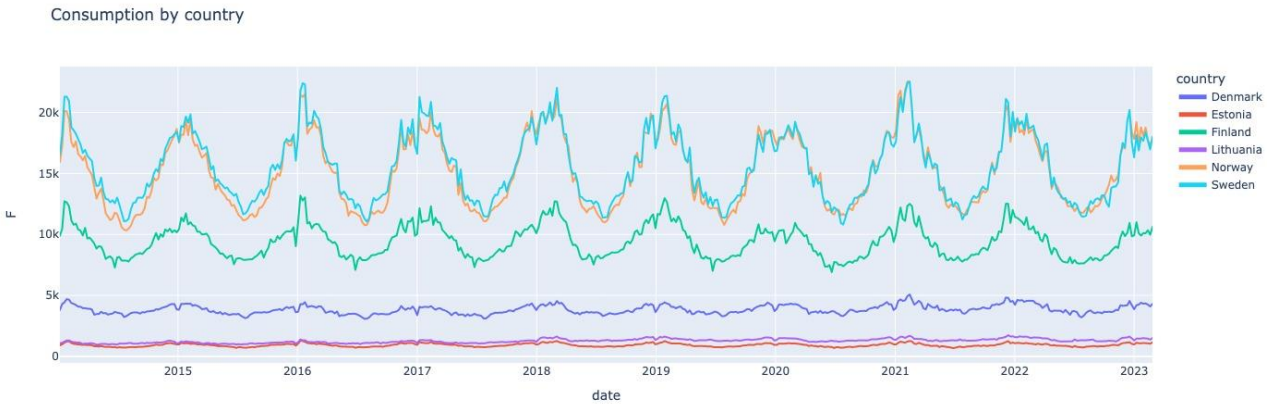


Figure 5. Consumption by country

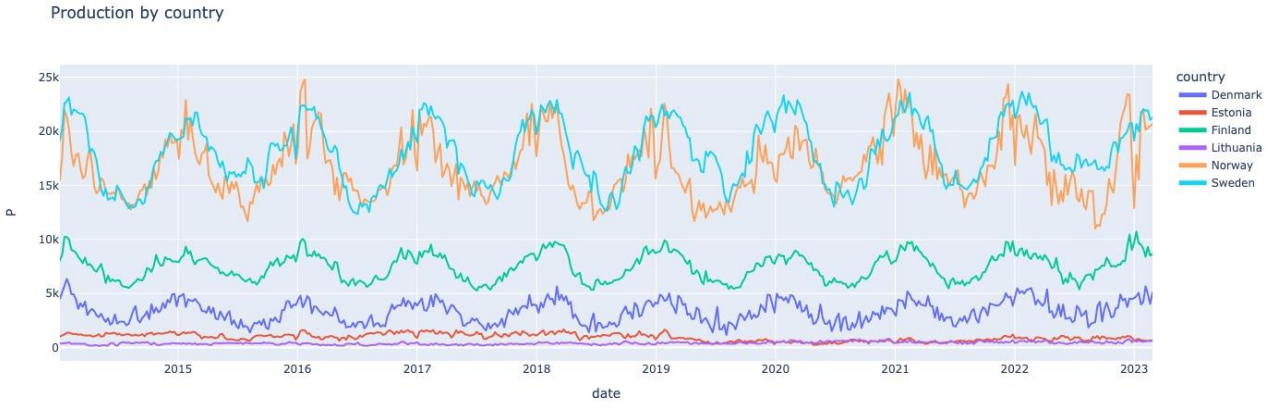


Figure 6. Production by country

Figure 7. Net Energy exchange by country

Figure 8. Wind generation by country (missing data for Sweden from 2018 to 2021)

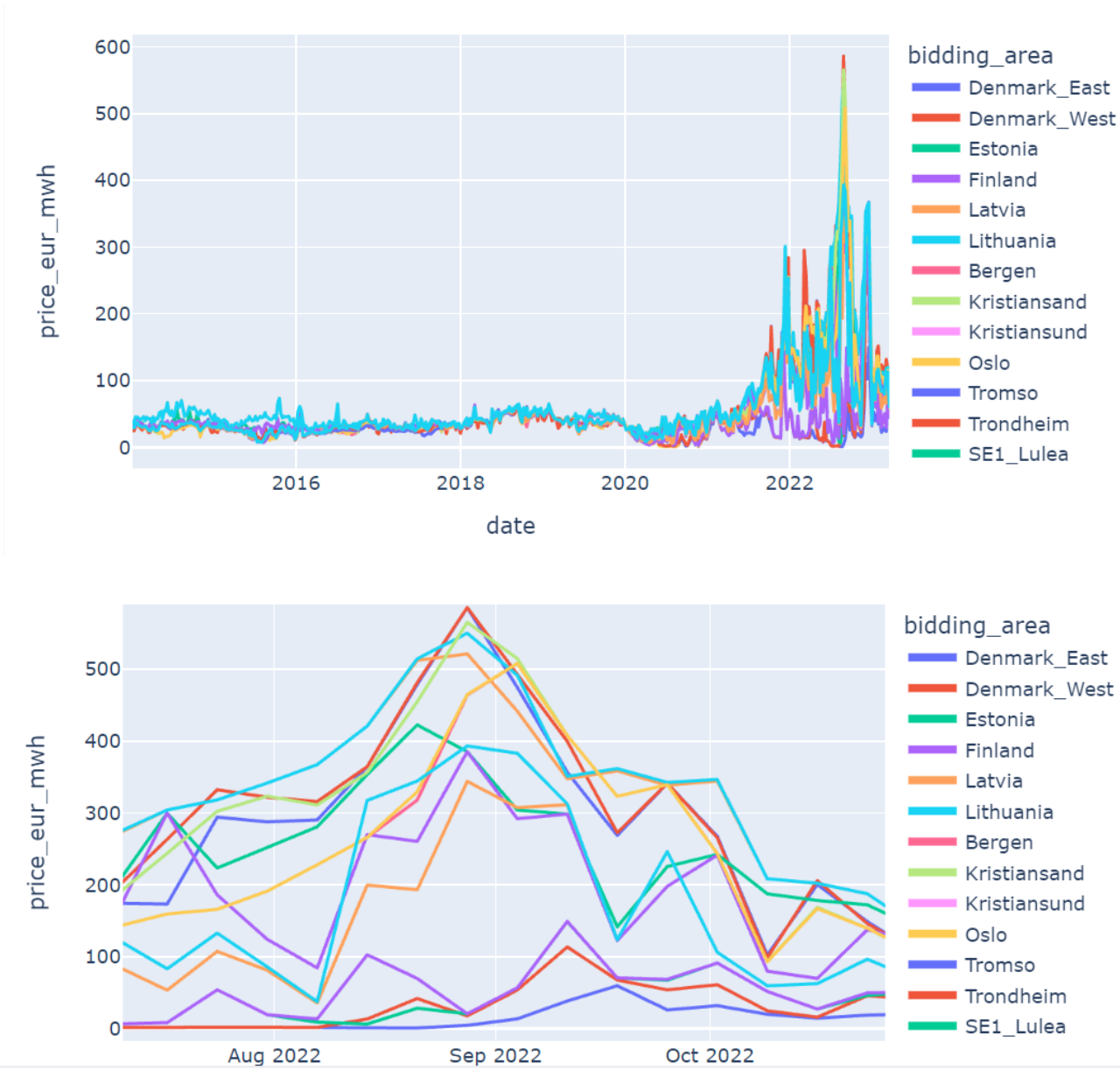


Figure 9. Spot prices by bidding area, lowest plot with zoom to observe period with higher prices.

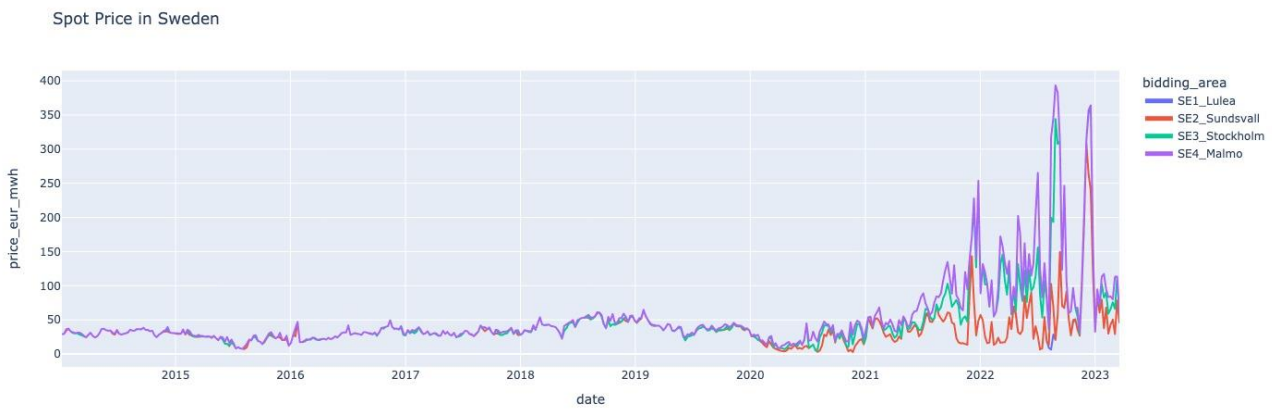


Figure 10. Spot prices in Sweden.

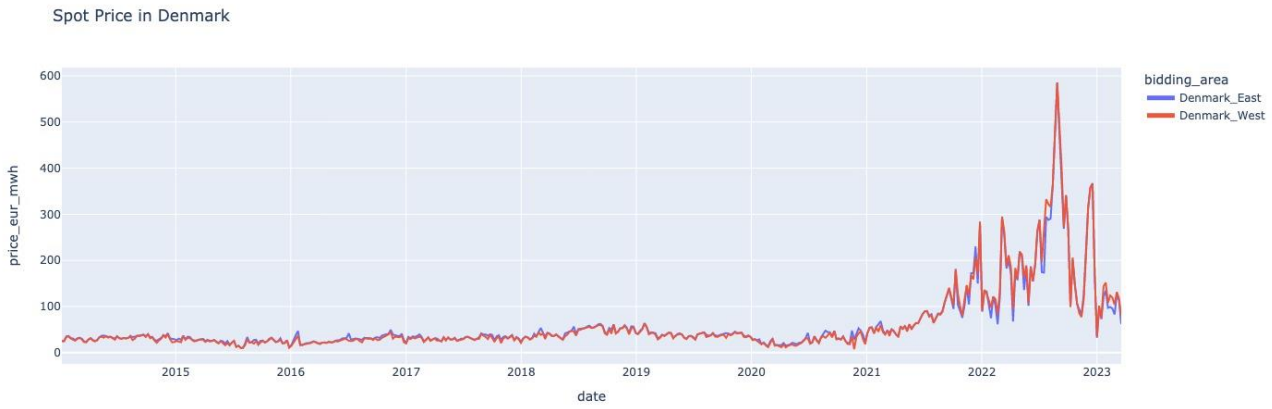


Figure 11. Spot prices in Denmark.

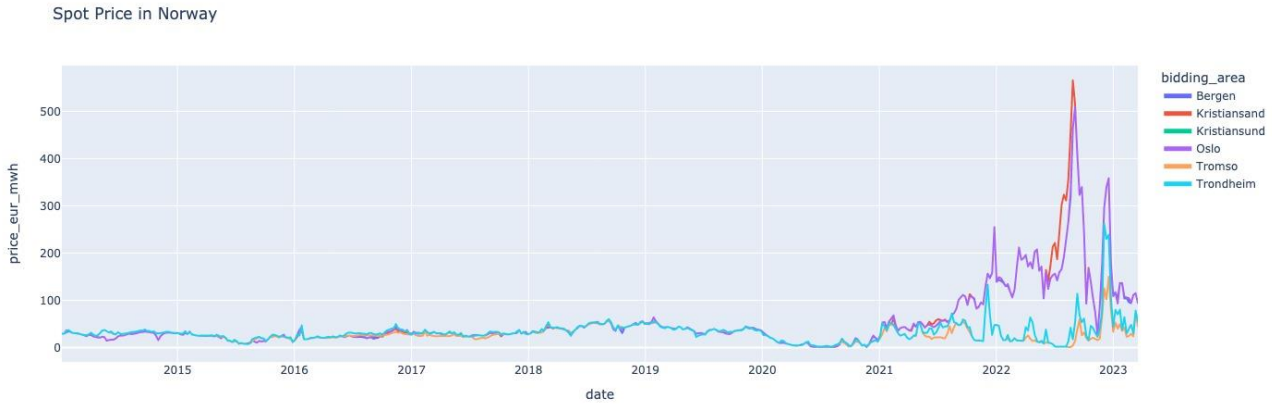


Figure 12. Spot prices in Norway.

Outliers, anomalies, or interesting patterns.

Figure 9 shows that despite higher prices in general all over the Scandinavia, the cost of the electricity in the northern regions was lower, which could be considered strange, assuming higher heating expenses during

the winter. However, northern regions energy production is equal higher at the North, while energy flow capacity (export power lines) is limited. All that leads to the higher export to EU and UK from the south and therefore higher local prices there.

Another anomaly could be observed where Norway energy production at the south significantly decreased in 2022 while the energy prices were extremely high. It can be explained by low rate of water reservoirs refill due to the lack of precipitation. Energy companies sold 'water' in form of generated electricity prior to that season, thus emptied reservoirs hoping for proper precipitation and unbalanced the electricity market pushing prices even higher.

Modeling and Analysis

Statistical and machine learning models to investigate the relationship between the variables.

The following libraries were used for further data processing and visualization:

- **pandas**: data manipulation and analysis.
- **plotly**: interactive data visualization.
- **orion**: hyperparameter optimization and configuration management.
- **tqdm**: progress bars.
- **pandarallel**: parallel processing of pandas dataframes, based on number of cores.
- **seaborn**: statistical data visualization.
- **numpy**: numerical operations and mathematical functions.
- **matplotlib**: creating visualizations in Python.

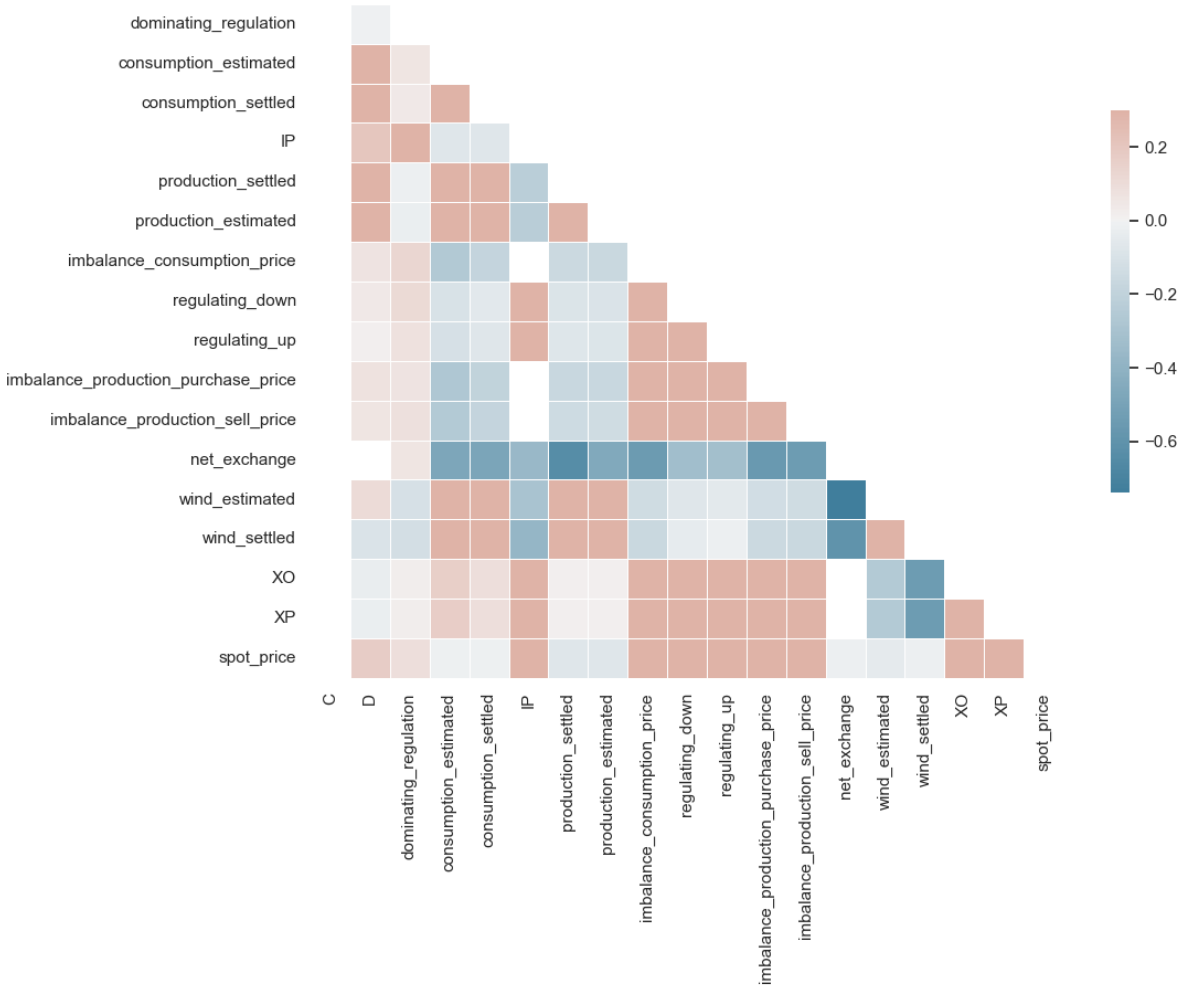


Figure 13. Correlation analysis

The highest correlation is between estimated wind power and a net exchange, which can be described by predictive energy exchange requests based on expected production. Net exchange has also high correlation with settled production, which is also explainable.

DBSCAN clustering was applied to cluster the spot prices for Bergen, Norway to identify patterns in the data. Resulted scatter plot provided below; however, it didn't provide us with any meaningful information.

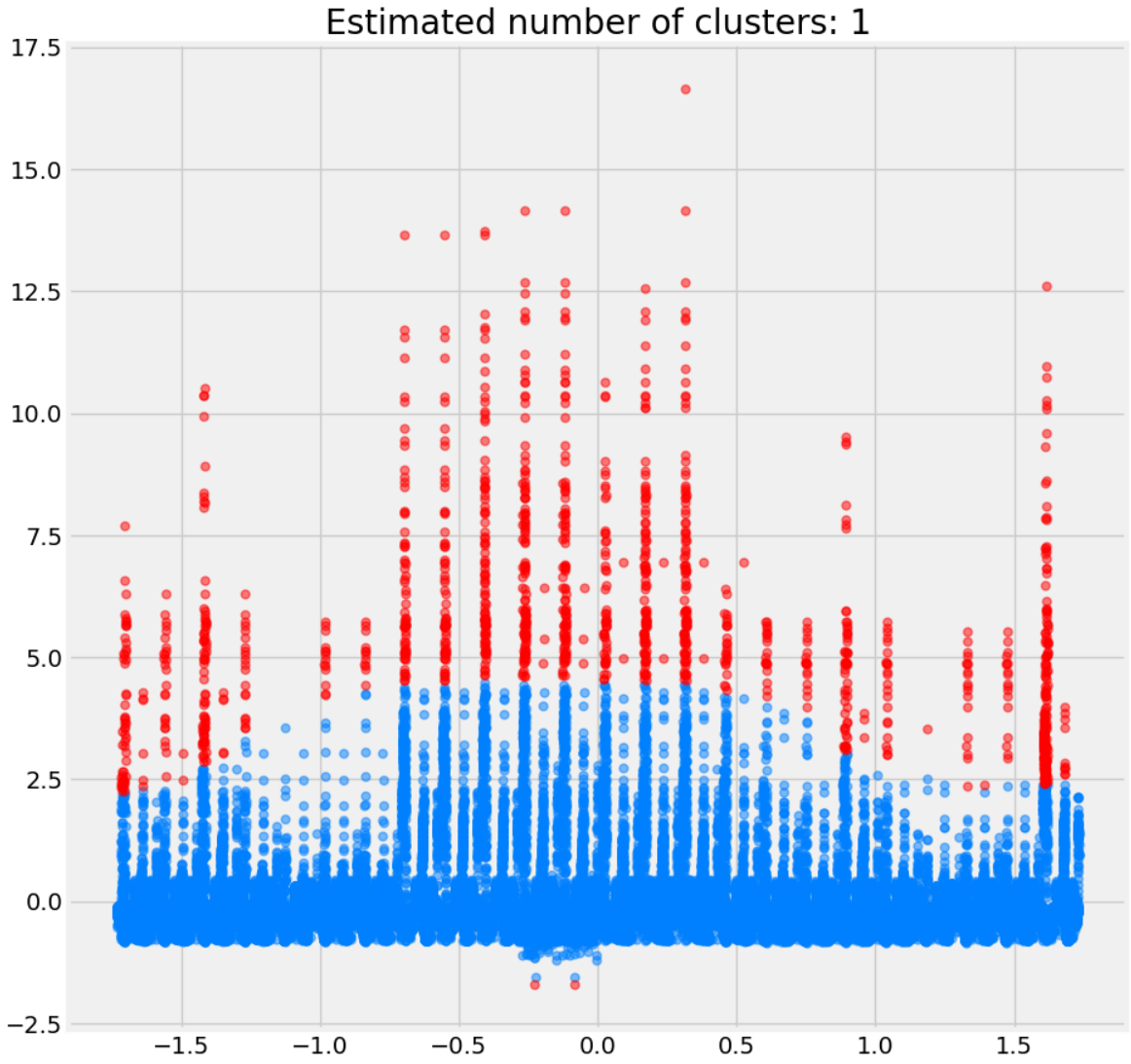
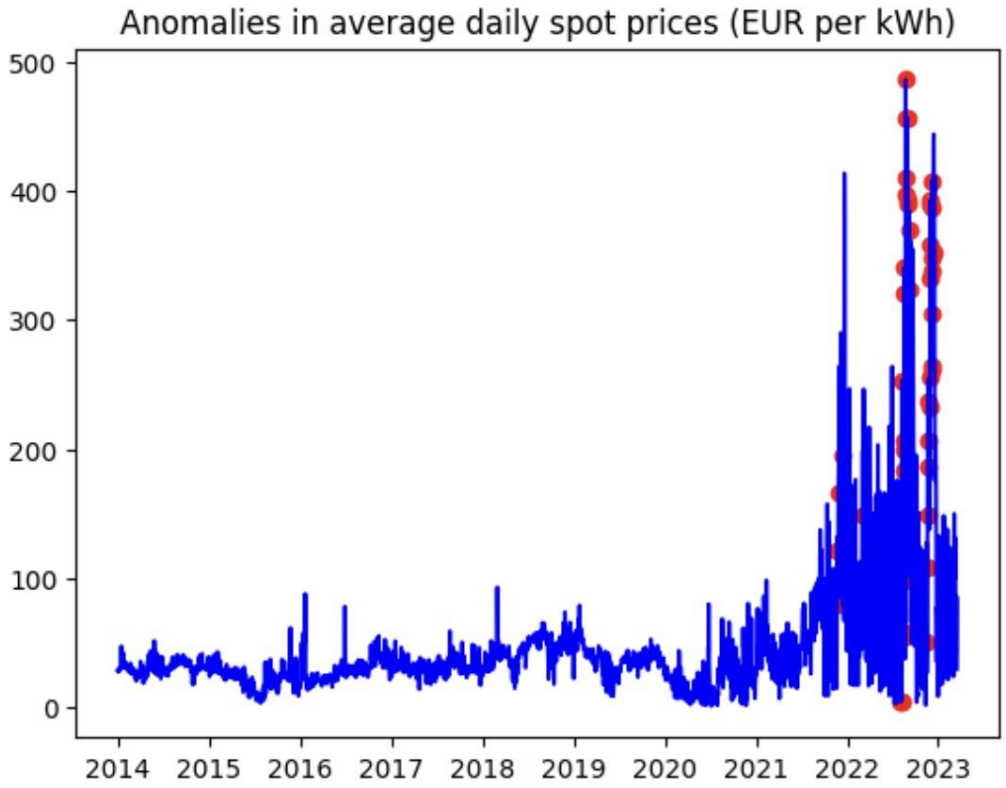


Figure 14. DBS Scan clustering

Anomaly detection

In the project we attempted to use LSTM model to implement unsupervised anomaly detection. We used Stockholm area as a target for the analyses. In a separate notebook (AnomalyDetection.ipynb) we prepare the data (extract, clean, engineer new features), then we build a TF LSTM model for sequential analyses. Then we consider the data that diverge from the predicted values as outliers. We take 5% of the outliers as anomalies. Using visualization we can conclude that the last 2 years are unusual for the energy market. That can be explained by political and epidemical turbulences.



Data Ethics

It is important to recognize that data ethics play a crucial role in any data analysis project. In our case, it is essential to acknowledge that the RAW data used in this project was obtained with permission from Nordpool on the student request and to be used only for research purposes. Therefore, it is important to handle data responsibly, ensuring that it is obtained legally and with proper consent.

. gitignore was adjusted to exclude to avoid data from uploading onto GitHub. Data analysis results to be sent by us to Nordpool as a part of agreement.

References

www.nordpoolgroup.com