university mohammed v
Faculty of Science of Rabat

Exploratory Data Analysis and Clustering of Solar Power Generation Data for Grid Integration

Fatmi Firdaous

Abstract. The integration of solar energy into the grid presents significant challenges due to its intermittent and uncontrollable nature. Accurate forecasting of solar power generation is crucial for effective grid management, ensuring a balanced supply and demand of electricity. This project focuses on the initial steps required for understanding solar power generation by conducting data cleaning, clustering, and exploratory data analysis (EDA) on a 50-year dataset of solar generation potential for European countries and NUTS 2 regions. The dataset, spanning hourly solar energy data from 1986 to 2015, is carefully processed to address missing values, outliers, and inconsistencies, ensuring a clean and reliable dataset for further analysis. Clustering techniques are applied to uncover inherent patterns and groupings in the data, providing insights into seasonal and regional variations in solar power generation. EDA is employed to visualize and better understand the data, revealing trends and anomalies in the data. This project aims to provide a comprehensive understanding of solar energy patterns, contributing to better grid integration strategies.

Key words: Solar energy, Grid integration, Data cleaning, Clustering, Exploratory data analysis, Solar power generation, Renewable energy

1. Introduction

As global energy demand rises, solar energy presents a promising solution due to its potential for widespread deployment. However, its intermittent nature poses challenges for grid integration. To address this, understanding solar generation patterns through data analysis is crucial. This project focuses on cleaning, clustering, and exploring historical solar generation data from 1986 to 2015 for European countries and NUTS 2 regions. The aim is to uncover trends and patterns in solar power generation, providing insights that support more effective grid integration and future forecasting efforts.[1]

2. Methodology

The project methodology consists of data cleaning, clustering, and exploratory data analysis (EDA) to better understand solar power generation patterns. The following steps were taken:

2.1. Data Cleaning

Both datasets (solar_ctry and solar_nuts) comprise 50 years' worth of solar generation data for European countries by country and by NUTS 2 system. The values in both datasets reflect the hourly estimates of the area's solar energy potential from 1986 to 2015. The dataset for NUTS 2 collects solar energy potential for different regions of a country, resulting in more data for a given country.

A time column was added to both datasets to track which hours of the day the area's energy potential would be higher than others. We would expect there to be a spike in the afternoon hours. Additionally, seasonality across the years was tracked with the month and week columns.

The pandas profiling report generated for the datasets showed the following:

- There are no missing values.
- Data for each country is heavily right-skewed (i.e., there are a lot more instances where no solar energy is generated).
- There is high correlation between the countries, which is expected since they are close together, and so will have similar exposure to the sun. Based on this, I could simplify the analysis by clustering the countries and perform analysis on just one country in each cluster.
- There are no negative values, so there is a remote likelihood of erroneous data collected.

Further exploration and analysis will be done after performing clustering on the countries.

To generate the time-related features, a ‘time‘ column was added to both datasets using the following function:

```
def add_time(df):
    ’’’adds time column based on start and end year’’’
    df[’time’] = pd.date_range(start=’1/1/1986’, periods=df.shape[0], freq=
    df[’hour’] = df[’time’].dt.hour
    df[’week’] = df[’time’].dt.isocalendar().week
    df[’month’] = df[’time’].dt.month
    #check
    print("-------------- Top 5 rows-----------------------------")
    display(df.head())
    print("------------------ Last 5 rows---------------------")
    display(df.tail())
```

This function generates the necessary time columns for each dataset, where ‘hour‘, ‘week‘, and ‘month‘ are extracted from the generated ‘time‘ column.

Further analysis will be conducted after clustering the countries based on their solar energy potential data.

2.2. Clustering

First, we visualize the average energy potential of each country across the years (Figure 1).

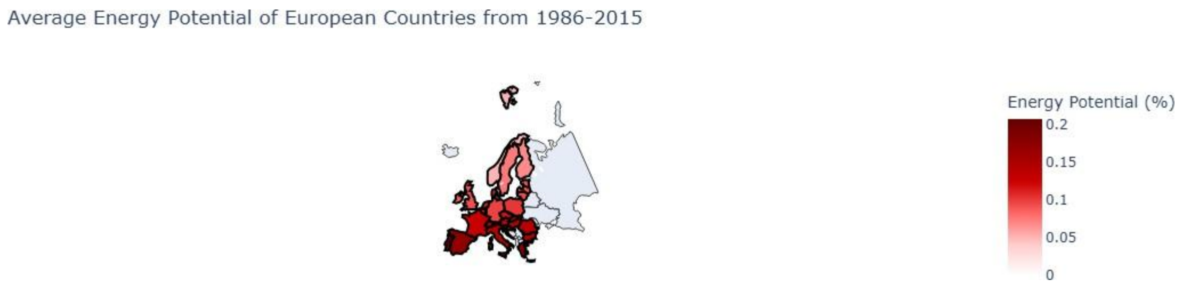

Average Energy Potential of European Countries from 1986-2015
Energy Potential (%)
0.2
0.15
0.1
0.05
0

Figure 1 Average Energy Potential by Country for 1968 to 2015

From the visualization, we observe that regions with greater solar energy potential are located closer to the Equator. This is expected as the sun’s rays strike the Earth’s surface most directly at the Equator. The color gradient is also consistent from bottom to top, showing high correlation between neighboring countries due to similar levels of exposure to the sun.

Next, I use KMeans to cluster the countries, choosing a range of 2 to 10 clusters. Using the elbow method and the distortion score, we determine the optimal number of clusters.

The distortion score is defined as the sum of squared distances from each data point to its assigned center (i.e., the sum of squared errors). The elbow method aims to identify a point where the distortion score begins to flatten as the number of clusters increases, forming an ”elbow.” This point is considered the optimal number of clusters for the data (Figure 2).

Based on this, we will go forward with 7 clusters using KMeans clustering.

2.3. Cluster Labels

I applied the cluster labels, and these are the countries in the different clusters (Figure 3):

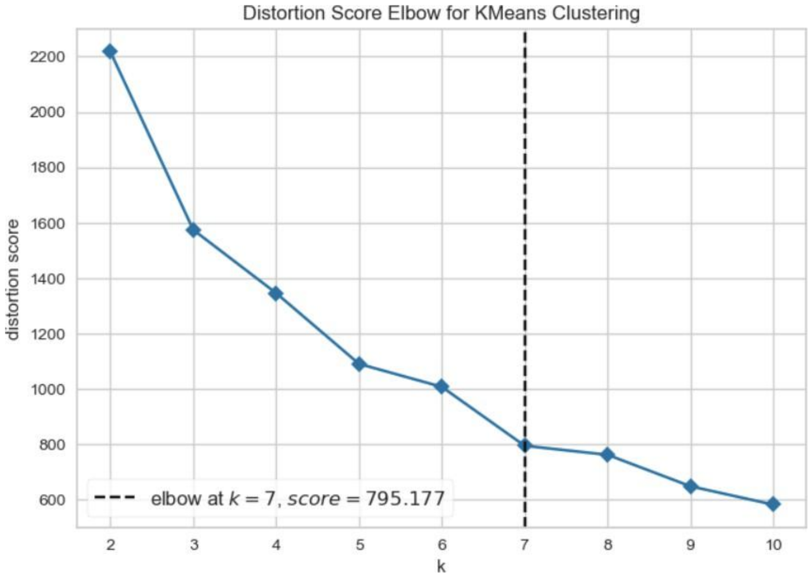
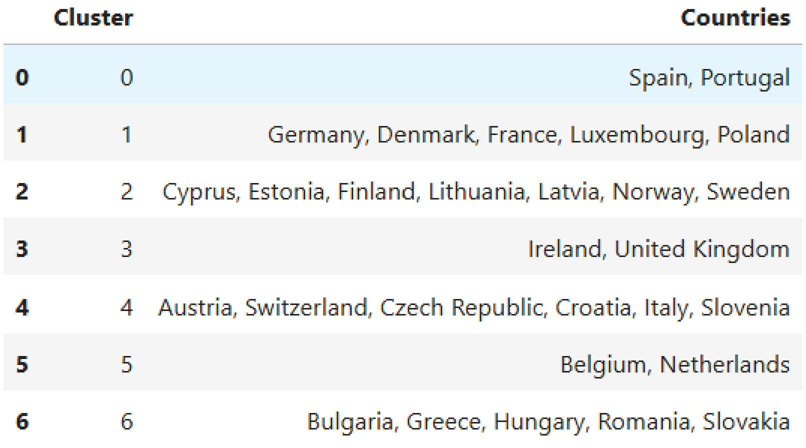
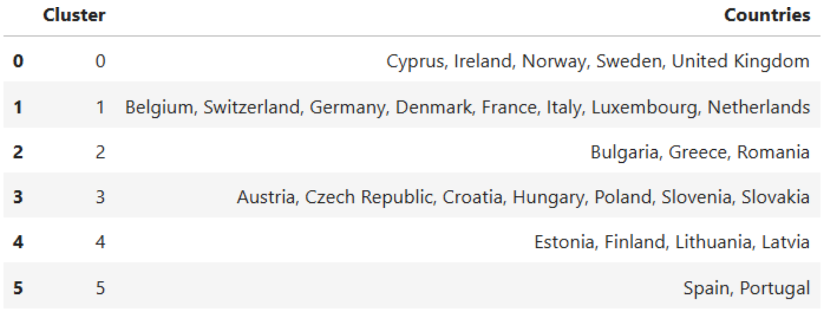


Figure 2 Distortion Score Elbow for KMeans Clusters 2 to 10



	Cluster	Countries
0	0	Spain, Portugal
1	1	Germany, Denmark, France, Luxembourg, Poland
2	2	Cyprus, Estonia, Finland, Lithuania, Latvia, Norway, Sweden
3	3	Ireland, United Kingdom
4	4	Austria, Switzerland, Czech Republic, Croatia, Italy, Slovenia
5	5	Belgium, Netherlands
6	6	Bulgaria, Greece, Hungary, Romania, Slovakia

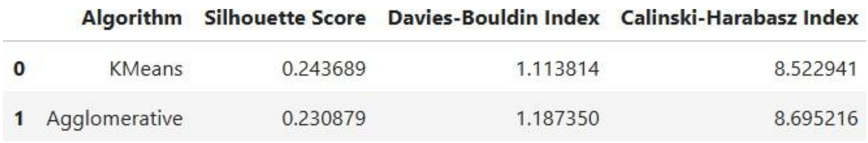
Figure 3 Countries in Different Clusters: K-means



	Cluster	Countries
0	0	Cyprus, Ireland, Norway, Sweden, United Kingdom
1	1	Belgium, Switzerland, Germany, Denmark, France, Italy, Luxembourg, Netherlands
2	2	Bulgaria, Greece, Romania
3	3	Austria, Czech Republic, Croatia, Hungary, Poland, Slovenia, Slovakia
4	4	Estonia, Finland, Lithuania, Latvia
5	5	Spain, Portugal

Figure 4 Countries in Different Clusters: Agglomerative Clustering

3. Benchmark: Comparison of Algorithms



	Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
0	KMeans	0.243689	1.113814	8.522941
1	Agglomerative	0.230879	1.187350	8.695216

Figure 5

4. Exploratory Data Analysis (EDA)

4.1. Data Selection and Sanity Check

Considering the large dataset, I decided to analyze only 10 years' worth of data. Before proceeding, I performed a sanity check to ensure:

- No negative values exist in the dataset.
- No values exceed 1 (since values represent percentages).

No anomalies were found, allowing me to proceed with data exploration.

4.2. Hourly Variation in Solar Energy Potential

First, I examine how solar energy potential changes over 24 hours (Figure 6).

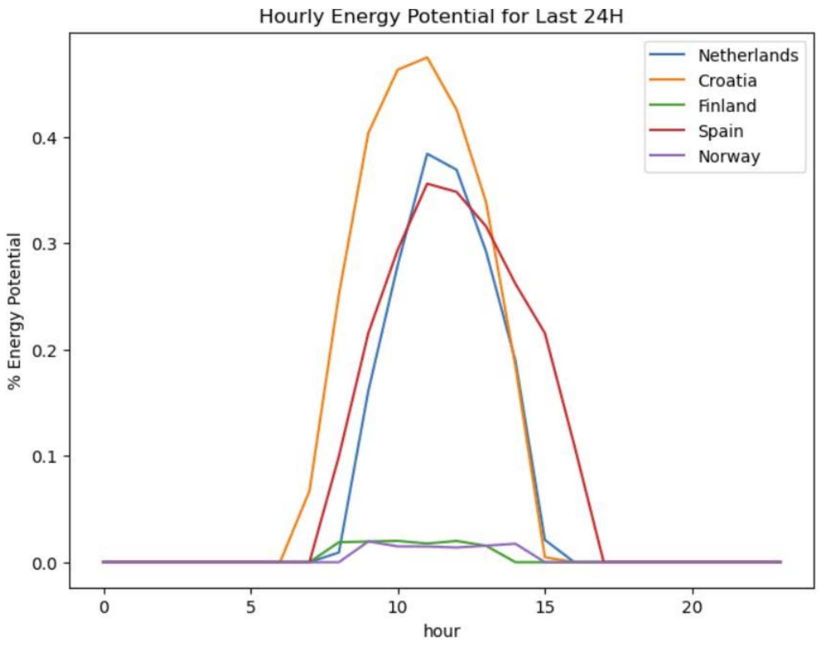


Figure 6 Hourly solar energy potential for each country

Key observations:

- Energy potential is only present between 06:00 and 17:00 hours, reflecting daylight periods.

- Croatia has the highest energy potential, while Norway and Finland have the lowest.
- Spain has the ”fattest” curve, indicating the longest sun exposure.
- The peak occurs at noon across all countries.

4.3. Distribution of Solar Energy Potential During Daylight Hours

To further analyze energy potential patterns, I examine the distribution of solar energy potential during daylight hours (Figure 7).

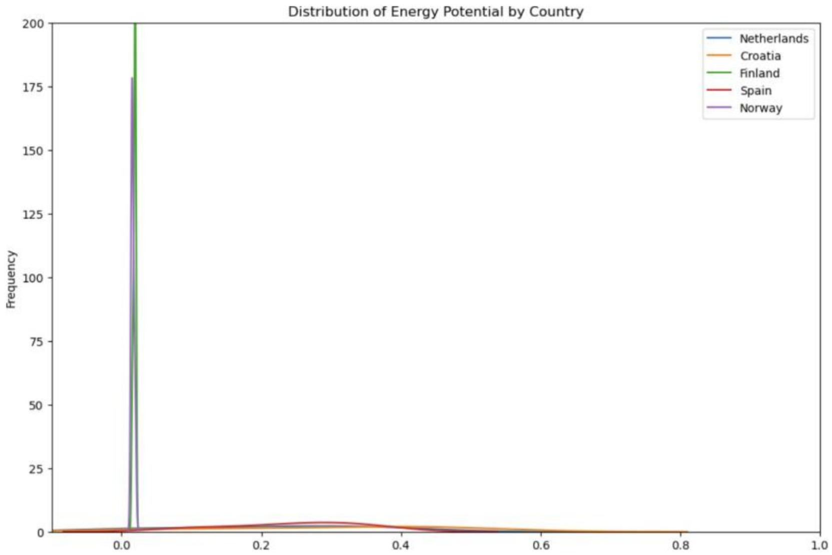


Figure 7 Distribution of hourly solar energy potential for each country

Key insights:

- Norway and Finland’s peak energy potential is close to 0%, suggesting they are not ideal for heavy reliance on solar energy.
- The remaining countries have energy distributions ranging from 0% to about 0.6%, with peaks between 0.3% and 0.5%.
- Spain has the highest solar exposure, supporting earlier findings.

4.4. Seasonal Trends in Solar Energy Potential

To analyze seasonal patterns, I plot Spain’s 10-year energy potential data (Figure 8).

Observations:

- Clear seasonal trends emerge, with consistent peaks and troughs each year.
- The regularity of these patterns makes predictive modeling more reliable.

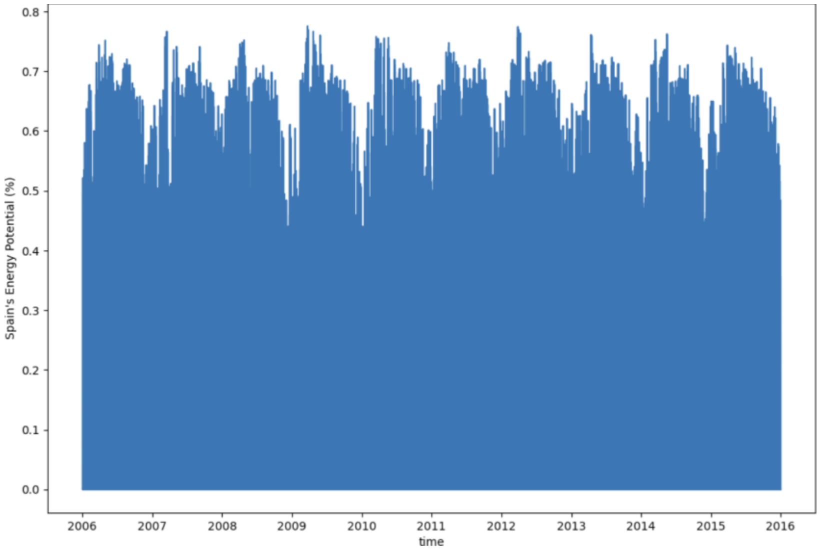


Figure 8 Spain's yearly seasonality of energy potential

4.5. Monthly Seasonality Analysis

Next, I analyze how solar energy potential fluctuates within different months of the year in Spain (Figure 9).

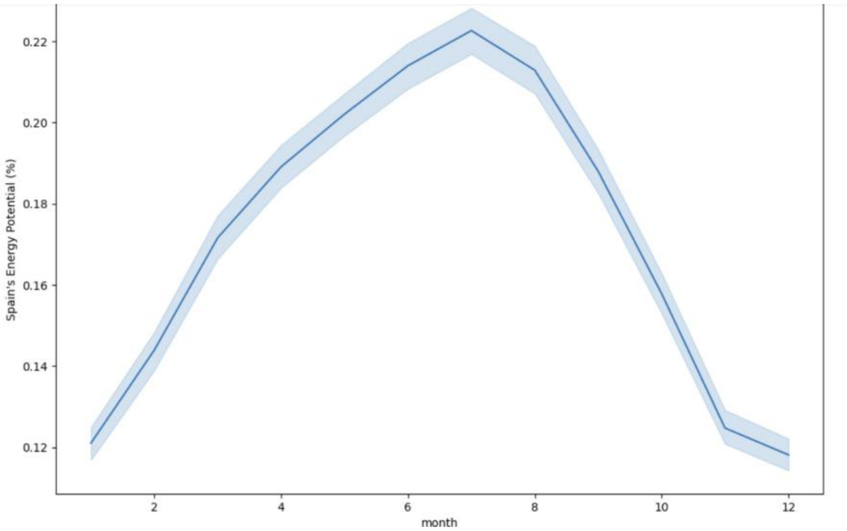


Figure 9 Spain's seasonality of energy potential by months

Key takeaways:

- The hottest months (April to August) coincide with spring and summer, with the highest energy potential.
- The coldest months (December to March) coincide with winter, showing the lowest energy potential.

4.6. Regional Trends Using NUTS2 Dataset

To further validate these trends, I analyze regional seasonality using the NUTS2 dataset (Figure 10).

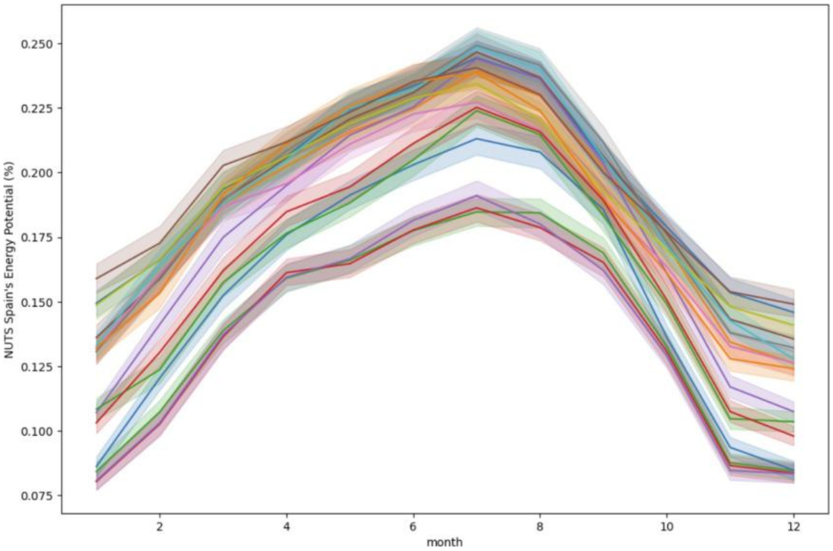


Figure 10 Spain's seasonality of energy potential by months (NUTS2 system)

The NUTS2 dataset confirms that different regions of Spain follow a similar seasonal trend in solar energy potential.

5. Data Dictionary

Below is the description of the dataset, sourced from Kaggle. The data was made available by the European Commission's STETIS Program.

Axis	Type	Description
Columns	str	European Country Codes
Rows	float	Hourly estimates of an area's energy potential for 1986-2015 as a percentage of a power plant's maximum output

Table 1 Data Dictionary

6. References

The dataset used in this analysis was obtained from Kaggle:

Sohier, Trevor. *30 Years of European Solar Generation*. 2018. Avail- able at: https://www.kaggle.com/datasets/sohier/30-years-of-european-solar-generationhttps://www.kaggle.com/datasets/sohier/30-years-of-european-solar-generation.