Data Analytics

Energy Consumption in France:

Data Analysis and Predictive Modeling

Adel Chahed

March, 2023

Table des matières

Introduction

Energy consumption is a critical topic in today's society and economy, with implications for climate change, energy transition, and energy sovereignty. Climate change has become an urgent and pressing issue, with the increase in global temperature posing a significant threat to the planet's ecosystem. Therefore, it's important to examine how we consume energy and understand how we can reduce our carbon footprint.

This report focuses on energy consumption in France, which is essential for understanding the country's energy usage but also for gaining insights into how we can predict the consumption in the future. The interest in this topic stems from the recent energy crisis caused by the geopolitical tensions between Russia and Ukraine, which have highlighted the importance of energy security. Given my deep interest in the challenges related to climate change and energy transition, it was clear to me that I should focus my final project on this topic.

To gather data on energy consumption, production, and weather, I used an open data platform called ODRE, which provides access to various energy-related datasets. Additionally, I discovered a dashboard on RTE France’s website that displays real-time energy consumption metrics in France, which I found fascinating. It inspired me to extract the data and perform further analysis. I also extracted a dataset about public holidays in France to see if it’s an important feature for energy consumption.

The goal of this project is to develop a time series model that predicts energy consumption in France based on historical data and identifies the key features that drive energy consumption. The aim of this report is to enhance our understanding of this crucial topic as citizens, and to assist energy management decisions in making more informed choices regarding energy consumption and production.

Plan

1. Data collection: Obtain data from the appropriate sources
2. Data Cleaning: Load CSVs into Python and clean it
3. Create a Database in MySQL
4. Export clean data to MySQL for processing
5. Join datasets into one
6. Data Description Analytics with Python
7. Process data to use appropriate metrics
8. Times Series Model: prediction of the energy consumption
9. Feature Selection and Engineering
10. Prediction of the energy consumption based on appropriate features

Data and data sources

For this project, I am going to use three datasets to analyze the energy consumption in France over time: “Energy Consumption and Production”, “Weather” and “Holidays".

Energy Consumption and Production

This dataset provides real-time data on electricity consumption and production in France, refreshed every hour. The data includes actual consumption, forecasted consumption, production from different energy sources, consumption by pumps in energy transfer stations, physical exchanges at borders, and estimated carbon emissions generated by electricity production in France. The dataset also includes information on the breakdown of production by technology and energy source. The data is provided at a quarter-hourly interval and includes a historical depth of January 2012 to May 2022 (around 400k rows).

The dataset is produced by RTE (French electricity transmission system operator) and is available for free via the éCO2mix tool, which is designed to help consumers better understand and consume electricity.

Column Name	Description
Périmètre	The geographic scope of the data, which covers the entire territory of France
Nature	The type of data, which includes consolidated and final data
Date	The date of the data point
Heure	The hour of the data point
Date et Heure	The date and time of the data point
Consommation (MW)	Actual electricity consumption in megawatts (MW)
Prévision J-1 (MW)	Forecasted electricity consumption for the previous day in MW
Prévision J (MW)	Forecasted electricity consumption for the current day in MW
Fioul (MW)	Production from oil in MW
Charbon (MW)	Production from coal in MW
Gaz (MW)	Production from gas in MW

Nucléaire (MW)	Production from nuclear power plants in MW
Eolien (MW)	Production from wind turbines in MW
Solaire (MW)	Production from solar power in MW
Hydraulique (MW)	Production from hydroelectric power in MW
Pompage (MW)	Consumption by pumps in energy transfer stations in MW
Bioénergies (MW)	Production from bioenergy sources in MW
Ech. physiques (MW)	Physical exchanges with neighboring countries in MW
Taux de CO2 (g/kWh)	Estimated carbon emissions generated by electricity production in France in grams per kilowatt-hour (g/kWh)
Ech. comm. Angleterre (MW)	Commercial exchanges with England in MW
Ech. comm. Espagne (MW)	Commercial exchanges with Spain in MW
Ech. comm. Italie (MW)	Commercial exchanges with Italy in MW
Ech. comm. Suisse (MW)	Commercial exchanges with Switzerland in MW
Ech. comm. Allemagne-Belgique (MW)	Commercial exchanges with Germany and Belgium in MW
Fioul - TAC (MW)	Production from oil in cogeneration units in MW
Fioul - Cogénération (MW)	Production from oil in thermal power plants with cogeneration in MW
Fioul - Autres (MW)	Production from other oil sources in MW
Gaz - TAC (MW)	Production from gas in cogeneration units in MW
Gaz - Cogénération (MW)	Production from gas in thermal power plants with cogeneration in MW
Gaz - CCG (MW)	Production from combined cycle gas turbines in MW
Gaz - Autres (MW)	Production from other gas sources in MW
Hydraulique - Fil de l'eau + éclusée (MW)	Production from hydroelectric power using water flows and lock systems in MW
Hydraulique - Lacs (MW)	Production from hydroelectric power using lakes in MW
Hydraulique - STEP turbinage (MW)	Production from hydroelectric power using pumped storage in MW
Bioénergies - Déchets (MW)	Production from bioenergy sources using waste in MW
Bioénergies - Biomasse (MW)	Production from bioenergy sources using biomass in MW

Bioénergies - Biogaz (MW)	Production from bioenergy sources using biogas in MW

Holidays

This dataset, "Jours fériés en France", provides information on French public holidays as defined by the labor code. The data includes the date of the holiday, the year it occurred, the geographic area it applies to (such as mainland France), and the name of the holiday. The dataset covers holidays for up to 20 years in the past and 5 years in the future from the current year. The data is provided in CSV format and extracted from the Government's open data platform.

Column Name	Description
date	Date of the holiday
annee	Year of the holiday
zone	Geographic area the holiday applies to (e.g., "Métropole" for mainland France)
nom_jour_ferie	Name of the holiday

Weather

The weather dataset includes daily historical data on maximum power (in GW) required to cover the peaks of French gross consumption. It also includes values of reference and average temperatures. The dataset is provided by RTE and Météo-France, updated monthly, and covers the entire territory of France. The temporal granularity is daily, and the depth of history ranges from 2012 to M-1 (previous month). The dataset is licensed under the Open License v2.0 (Etalab) and the data source is available on the ODRE platform. The dataset is related to electricity energy and the networks of distribution and transport.

Column Name	Description
Date	Date of the day for which the data is recorded
Pic journalier consommation (MW)	Maximum power (in MW) required to cover the peaks of French gross consumption on that particular day
Température moyenne (°C)	Average temperature (in °C) on that particular day
Température référence (°C)	Reference temperature (in °C) on that particular day

Data collection

The above data was collected by exporting the CSVs respectively from each platform. Then the data was parsed properly as shown in the next section, using Python.

Data cleaning and Exploratory data analysis

Fixing the Datetime Type

In the three datasets, I had to fix some data types. The most important feature to fix was the datetime and date columns.

```
# Combine the "Date" and "Heure" columns into a new "Datetime" column
df['Datetime'] = pd.to_datetime(df['Date'] + ' ' + df['Heure'])
```

Creating New Features

For all the dataset I have created more features about time to be able to match the rows in SQL. We have a datetime column in the Energy Consumption and Production dataset and each row of the table corresponds to an hour of a day. But in the other datasets, each row corresponds to a day, so we have a lot more rows in the first dataset.

```
# Creating "Year", "Month", "Day", "Weekday" "Hour" columns from "Datetime" column :
df['Year'] = df['Datetime'].dt.year
df['Month'] = df['Datetime'].dt.month
df['Day'] = df['Datetime'].dt.day
df['Weekday'] = df['Datetime'].dt.weekday
df['Hour'] = df['Datetime'].dt.hour
```

Dropping & Creating Columns

I dropped many low variance columns, with unique values. For example, columns regarding the geographical area (which is only France) or the kind of data (consolidated or final data). I also dropped very specific columns useless for our analysis in order to simplify our datasets. This includes columns regarding very specific kind of technologies of different energy sources and the name of each public holiday. Moreover, I have created a calculated field “Temperature Deviation (°C)” in the Weather dataset which is the difference between the average temperature and the reference temperature.

Dropping & Creating Rows

In our main dataset, the temporal granularity is quarter-hourly to half-hourly. But the quarter-hourly rows were empty so I dropped them. So, the temporal granularity became half-hourly. In the other datasets, I had to drop rows to match the time range of our main dataset. In the Holidays dataset there was only public holidays, so I have created a flag feature with 1 value for holidays. So, I had to create new rows of “normal” days (that are not holidays) with the flag feature equal to 0 between the holidays inside the time range.

```
# Defining the time range
start_date = pd.to_datetime('2012-01-01')
end_date = pd.to_datetime('2022-05-31')

# Creating a DataFrame with all dates in the time range
all_dates = pd.DataFrame(pd.date_range(start=start_date, end=end_date), columns=['date'])

# Merging the original DataFrame and the new DataFrame with all dates
df = pd.merge(all_dates, df, on='date', how='left')

# Replacing missing values in the 'Flag Holiday' column with 0
df['Flag Holiday'].fillna(0, inplace=True)

# Sorting the DataFrame by date
df.sort_values('date', inplace=True)

# Reseting the index of the sorted DataFrame
df.reset_index(drop=True, inplace=True)
```

NaN Handling

In the main dataset, I got around 700 rows with missing values out of 180,000 records. I replaced the null values with the median values of the columns.

Outliers

I tried to analyze the dataset in order to detect outliers, but I wasn't really convinced by the idea of removing them and probably corrupting the data. The outliers seem to occur exclusively in winter, probably because of cold waves. I preferred to keep these data in order to study the evolution of the number of days with extreme temperature deviation and to see if climate change can play a role in the evolution of energy consumption.

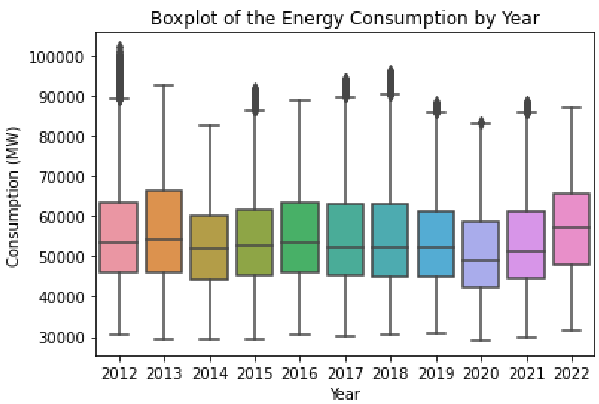
SQL connection

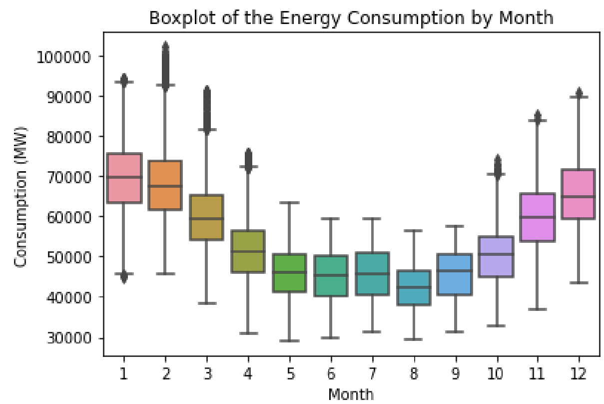
After the cleaning of the data, I was able to go ahead and insert the tables into MySQL. The reasons for that choice are explained in the next section. I connect the Python file with the MySQL database that I already established using PyMySQL, and SQLalchemy. Thus, I end up with 3 tables in MySQL, one for each of ‘holidays’, ‘energy_consumption_in_france’, ‘weather’. Once the data has been added and cleaned successfully, the subsequent chapter will involve analysis using SQL. However, at present, we will shift our attention back to EDA with Python.

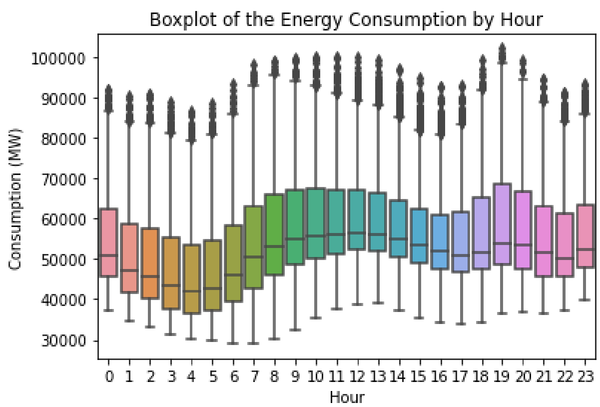
```
# Defining a function that reads the datasets into a pandas dataframe and writes them to a MySQL database:

def load_dataset_to_mysql(dataset_path, table_name):
    # Read dataset into a Pandas DataFrame
    df = pd.read_csv(dataset_path)

    # Define a database connection string
    connection_string = 'mysql+pymysql://root[redacted]@localhost/final_project_energy'

    # Create an SQLAlchemy engine to connect to MySQL
    engine = create_engine(connection_string, echo=True)

    # Write the DataFrame to MySQL
    df.to_sql(name=table_name, con=engine, if_exists='replace', index=False)
```

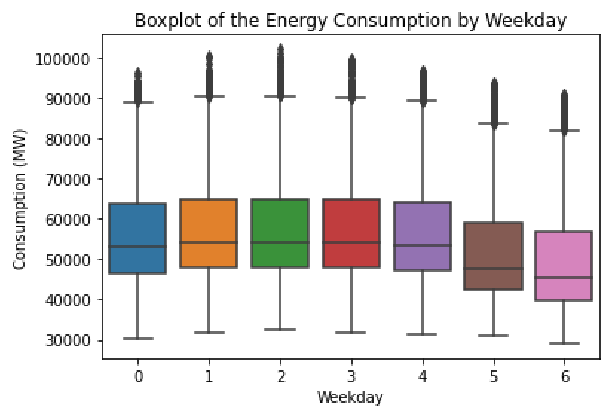
```
# Calling the function for each dataset you want to load into MySQL:
# Load the electricity consumption dataset
load_dataset_to_mysql('/Users/adel/Desktop/Final_project_Flow/GITHUB/energy_consumption_france.csv', 'electricity_consumption')

# Load the holidays dataset
load_dataset_to_mysql('/Users/adel/Desktop/Final_project_Flow/GITHUB/all_days.csv', 'holidays')

# Load the weather dataset
load_dataset_to_mysql('/Users/adel/Desktop/Final_project_Flow/GITHUB/weather.csv', 'weather')
```

EDA with Python









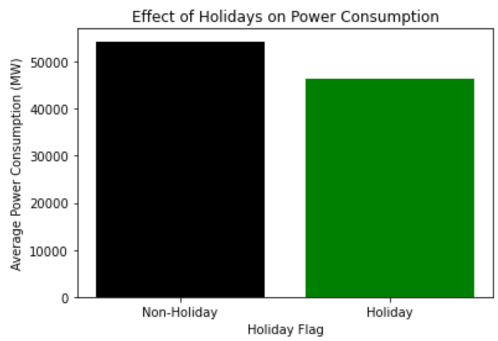
These boxplots allow us to visualize the evolution of the average energy consumption over time: per year, per month, per hour of the day and per day of the week (Monday is 0).

We observe that the energy consumption remains relatively stable over the years, with slight variations, especially in 2020 with the lockdown.

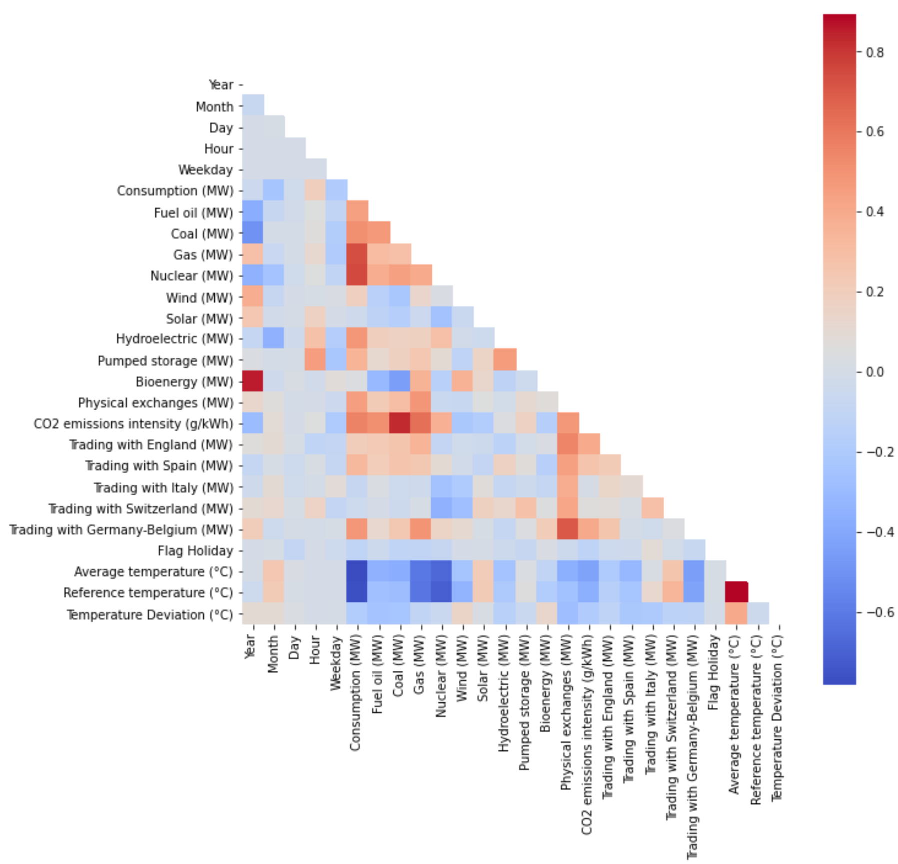
Some years seem to have a large number of outliers, this is mainly due to cold waves in winter (2012 is the most affected year).

We notice a seasonal pattern for the months, this was to be expected. In France, we observe that the consumption peaks are in winter (some hot countries may have the opposite pattern because of air conditioning). The boxplot of the months confirms that it is indeed in winter that the vast majority of outliers are found.

The boxplots on the hour and weekdays illustrate that we consume more energy on days and hours when we are working. Indeed, we observe a decrease in consumption on weekends and nights.

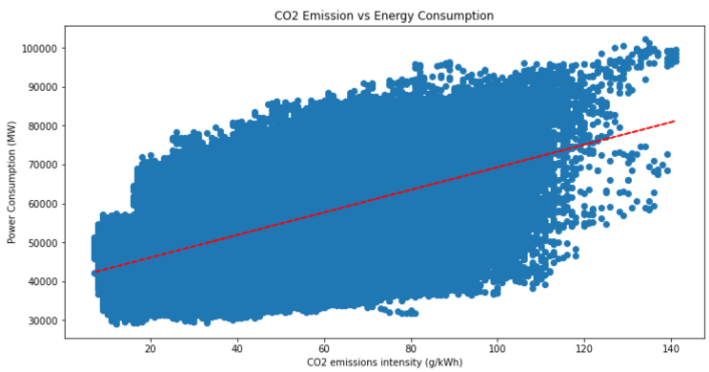


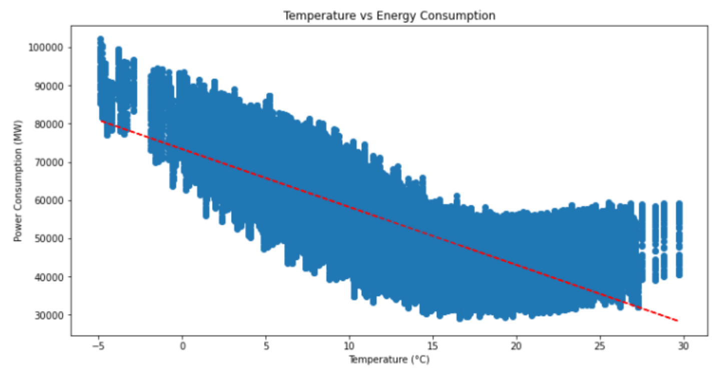
We consume less energy on public holidays.



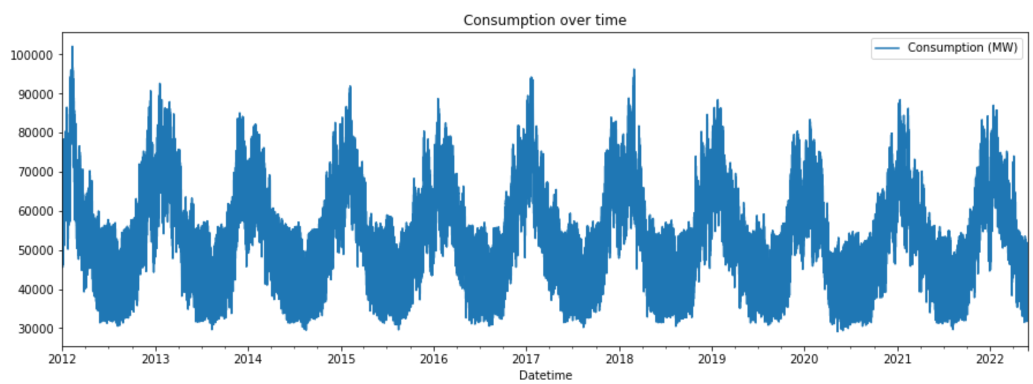
In this heat map the column "Consumption (MW)" is correlated with energy sources, which is logical, but mostly with gas and nuclear. This may indicate that these are probably important sources in the energy mix. The consumption is also correlated with the CO2 emission intensity. We also observe that there is a strong negative correlation between average and reference temperature and consumption. This means that we consume more energy when it is cold, which is in agreement with the previous charts.

Apart from that, we can also observe that there is a strong correlation between CO2 emissions and fuel oil, coal and gas.

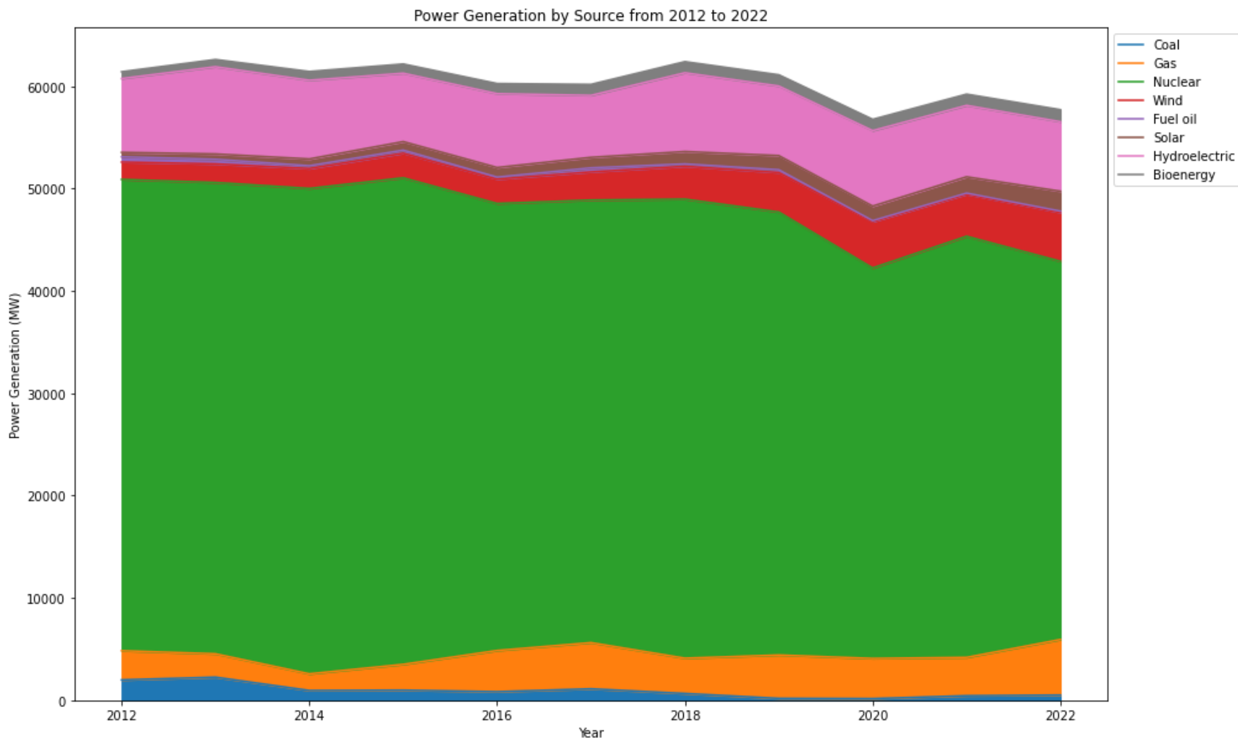




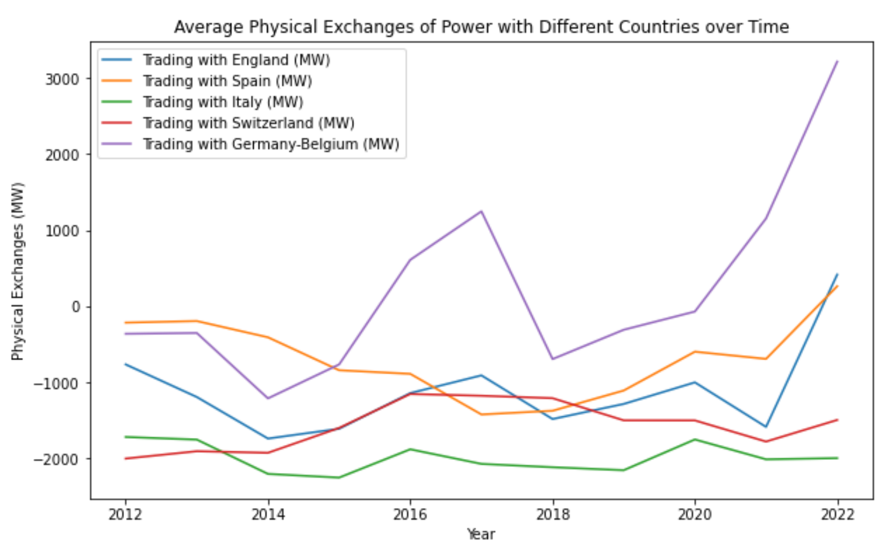
Indeed, energy consumption is correlated with temperature (it increases with the cold in France) and CO2 emissions.

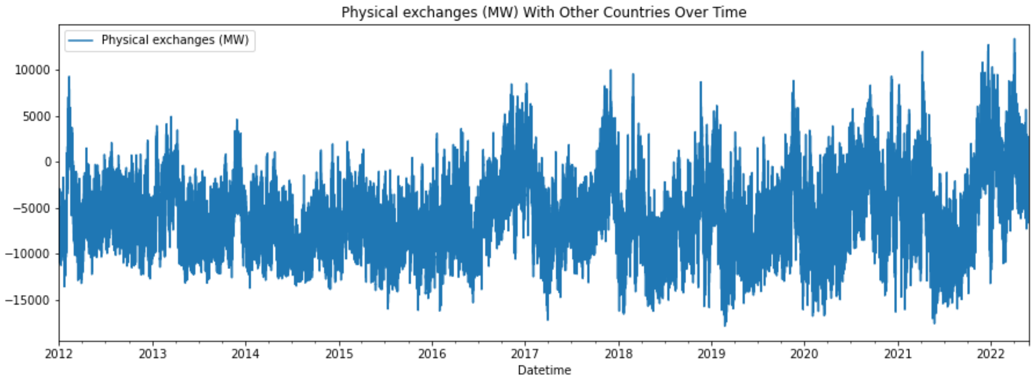


The chart representing energy consumption over time reaffirms our previous observation from the boxplots and provides a clearer illustration of the seasonal pattern.

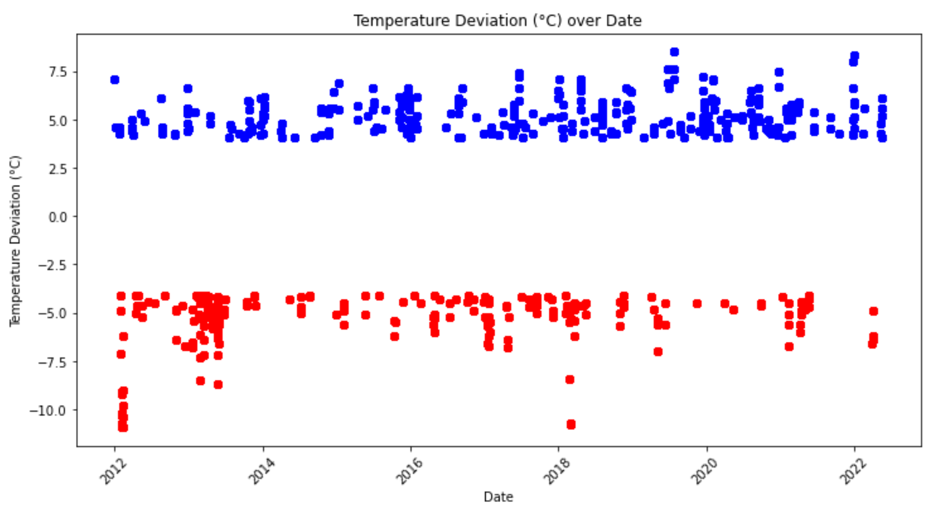


This area chart of the share of the types of energy sources in France shows us that nuclear is the main source of energy, and its share in the energy mix seems to be decreasing in recent years. Then comes hydroelectricity whose share seems to be stable in the mix. Wind power is our third source of energy and we observe a growth in its share in the mix during the last decade, this is also the case for solar energy. The share of gas seems to be increasing since 2014.





The two charts above illustrate the evolution of energy exchanges with neighboring countries. When the values are negative, France is exporting energy and when they are positive, it is importing electricity. We observe that France was an exporter on average during this decade but we can observe a trend towards imports. The production capacity seems to be increasingly overwhelmed in the winters. Imports mainly from Germany and Belgium, this may leave us that the French regions bordering these countries do not have close or efficient production capacity.



This scatter plot illustrates the number of days when the Temperature Deviation was significant (+4°C or -4°C). We observe that indeed 2012 was a year with strong waves of exceptional cold. But there is a slight tendency towards more extreme values in heat and less in cold.

SQL or NoSQL

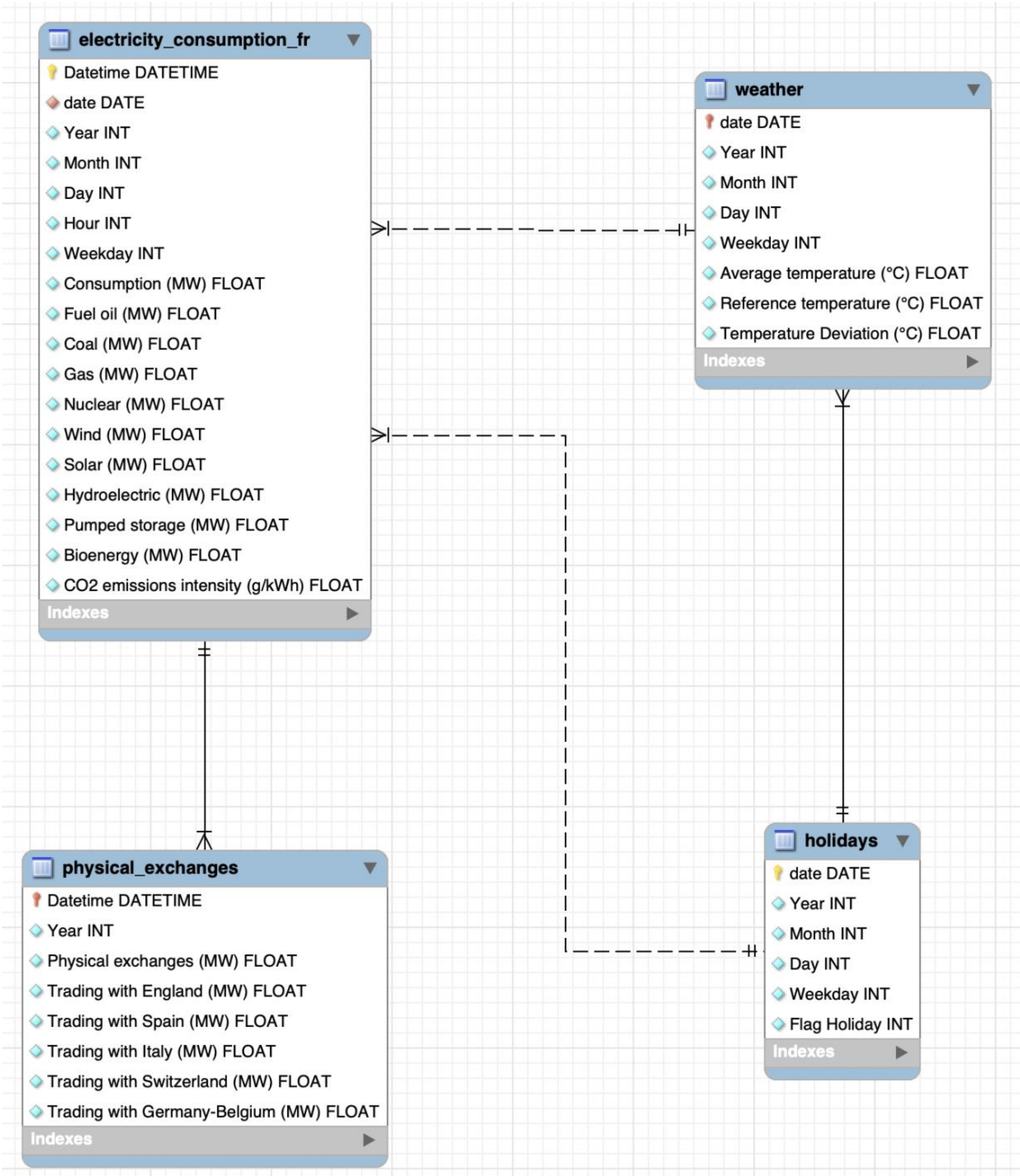
After conducting a thorough examination of the differences, benefits, and drawbacks of each database type, the primary distinctions can be summarized as follows:

- SQL databases are relational and rely on structured query language and predefined schemas. In contrast, NoSQL databases are non-relational and have dynamic schemas designed for unstructured data.
- SQL databases are vertically scalable, while NoSQL databases are horizontally scalable.
- SQL databases use a table-based model, whereas NoSQL databases are available in various forms, including document, key-value, graph, or wide-column stores.
- SQL databases are ideal for multi-row transactions, while NoSQL databases are better suited for handling unstructured data such as documents or JSON.

Given that structured tables with predefined schemas are used in our case, it seems appropriate to use SQL. Additionally, if we anticipate adding more data or sources in the future, SQL would be preferable due to its vertical scalability and ability to perform multi-row transactions.

ERM

The diagram for our entity relation model can be seen as follows:



The database consists of four tables: "electricity_consumption_fr", "physical_exchanges", "weather" and "holidays."

The "electricity_consumption_fr" table includes fields for date and time, year, month, day, hour, weekday, and various energy sources' consumption in France, including fuel oil, coal, gas, nuclear, wind, solar, hydroelectric, pumped storage, bioenergy, and CO2 emissions intensity.

The "physical_exchanges" table includes fields for date and time, year, and physical exchanges of electricity between France and other countries, including England, Spain, Italy, Switzerland, and Germany-Belgium.

The "weather" table contains information on the weather conditions in France, including the average temperature, reference temperature, and temperature deviation.

The "holidays" table includes fields for date, year, month, day, weekday, and a flag indicating whether the day is a holiday or not.

MySQL Queries

The MySQL code used to create the database prior to pushing in the tables:

```
CREATE DATABASE IF NOT EXISTS `final_project_energy` ;
USE `final_project_energy` ;
```

After I imported the datasets with PyMySQL, I had to clean the tables: Fixing the data types of the columns, adding primary and foreign keys, and creating a new column ‘date’ for the energy consumption table. I divided the main table into two separate tables, namely electricity_consumption_fr and physical_exchanges.

Examples of data cleaning query:

```
ALTER TABLE weather
MODIFY date date NOT NULL,
MODIFY Year int NOT NULL,
MODIFY Month int NOT NULL,
MODIFY Day int NOT NULL,
MODIFY Weekday int NOT NULL,
MODIFY `Average temperature (°C)` float NOT NULL,
MODIFY `Reference temperature (°C)` float NOT NULL,
MODIFY `Temperature Deviation (°C)` float NOT NULL,
ADD PRIMARY KEY (date),
ADD CONSTRAINT `fk_weather_holidays` FOREIGN KEY (date) REFERENCES holidays(date);
```

```
-- Create the physical_exchanges table
CREATE TABLE physical_exchanges (
  Datetime datetime NOT NULL,
  date date NOT NULL,
  Year int NOT NULL,
  `Physical exchanges (MW)` float NOT NULL,
  `Trading with England (MW)` float NOT NULL,
  `Trading with Spain (MW)` float NOT NULL,
  `Trading with Italy (MW)`float NOT NULL,
  `Trading with Switzerland (MW)` float NOT NULL,
  `Trading with Germany-Belgium (MW)` float NOT NULL,
  PRIMARY KEY (Datetime)
);
```

The average electricity consumption by sources of energy:

```
-- The average electricity consumption by sources of energy
SELECT 'Fuel oil' AS source, ROUND(AVG(`Fuel oil (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Coal' AS source, ROUND(AVG(`Coal (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Gas' AS source, ROUND(AVG(`Gas (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Nuclear' AS source, ROUND(AVG(`Nuclear (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Wind' AS source, ROUND(AVG(`Wind (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Solar' AS source, ROUND(AVG(`Solar (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Hydroelectric' AS source, ROUND(AVG(`Hydroelectric (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Pumped storage' AS source, ROUND(AVG(`Pumped storage (MW)`)) AS average_consumption
FROM electricity_consumption_fr
UNION ALL
SELECT 'Bioenergy' AS source, ROUND(AVG(`Bioenergy (MW)`)) AS average_consumption
FROM electricity_consumption_fr
GROUP BY source
ORDER BY average_consumption DESC;
```

source	average_consumption
Nuclear	43862
Hydroelectric	7204
Gas	3402
Wind	2945
Solar	1039
Bioenergy	968
Coal	936
Fuel oil	281
Pumped storage	-787

The average production of pumped storage is negative probably because more energy is used to pump water uphill than is generated by releasing it downhill.

The average trading (MW) exchanges with other countries:

```sql
-- The average trading (MW) exchanges with other countries
SELECT 'Trading with England (MW)' AS country, ROUND(AVG(`Trading with England (MW)`)) AS average_trading
FROM physical_exchanges
UNION ALL
SELECT 'Trading with Spain (MW)' AS country, ROUND(AVG(`Trading with Spain (MW)`)) AS average_trading
FROM physical_exchanges
UNION ALL
SELECT 'Trading with Italy (MW)' AS country, ROUND(AVG(`Trading with Italy (MW)`)) AS average_trading
FROM physical_exchanges
UNION ALL
SELECT 'Trading with Switzerland (MW)' AS country, ROUND(AVG(`Trading with Switzerland (MW)`)) AS average_trading
FROM physical_exchanges
UNION ALL
SELECT 'Trading with Germany-Belgium (MW)' AS country, ROUND(AVG(`Trading with Germany-Belgium (MW)`)) AS average_trading
FROM physical_exchanges
GROUP BY country
ORDER BY average_trading DESC;
```

country	average_trading
Trading with Germany-Belgium (MW)	55
Trading with Spain (MW)	-733
Trading with England (MW)	-1204
Trading with Switzerland (MW)	-1572
Trading with Italy (MW)	-1992

The average physical exchanges by Year:

```sql
-- The average physical exchanges by Year
SELECT Year, ROUND(AVG(`Physical exchanges (MW)`)) AS average_physical_exchanges
FROM physical_exchanges
GROUP BY Year
ORDER BY average_physical_exchanges DESC;
```

Year	average_physical_exchanges
2022	88
2017	-4678
2016	-4811
2021	-5207
2020	-5210
2012	-5254
2013	-5607
2019	-6676
2018	-7275
2015	-7373
2014	-7748

The Average temperature (°C) and the date of the top 10 days with the higher average Consumption (MW):

```
-- The Average temperature (°C) and the date of the top 10 days with the higher average Consumption (MW)
SELECT weather.`date`, ROUND(AVG(weather.`Average temperature (°C)`), 1) AS `Average temperature (°C)`,
ROUND(AVG(electricity_consumption_fr.`Consumption (MW)`), 1) AS `Average Consumption (MW)`
FROM electricity_consumption_fr
INNER JOIN weather ON electricity_consumption_fr.`date` = weather.`date`
GROUP BY weather.`date`
ORDER BY `Average Consumption (MW)` DESC
LIMIT 10;
```

date	Average temperature (°C)	Average Consumption (MW)
08/02/2012	-4.9	94096.7
09/02/2012	-3.8	92868.1
07/02/2012	-4.8	91176.3
10/02/2012	-2.9	90262.6
28/02/2018	-3.5	89600.1
03/02/2012	-4.6	89247.9
06/02/2012	-3.3	88151.9
11/02/2012	-4.8	87951.2
02/02/2012	-3.5	87524
27/02/2018	-3.7	87012.2

The average energy consumption (MW) and average CO2 emission (g/kWh) by flag holiday:

```sql
-- The average energy consumption (MW) and average CO2 emission (g/kWh) by flag holiday
SELECT holidays.`Flag Holiday`,
    ROUND(AVG(electricity_consumption_fr.`Consumption (MW)`) ,0) AS avg_consumption,
    ROUND(AVG(electricity_consumption_fr.`CO2 emissions intensity (g/kWh)`) ,0) AS avg_co2
FROM holidays
INNER JOIN electricity_consumption_fr ON holidays.`date` = electricity_consumption_fr.`date`
GROUP BY holidays.`Flag Holiday`;
```

Flag Holiday	avg_consumption	avg_co2
1	46372	32
0	54350	48

Creating a temporary table that merges all the tables in order extract it for the next steps of the project:

```sql
CREATE TEMPORARY TABLE temp_electricity_consumption AS
SELECT ecf.*,
       h.`Flag Holiday`,
       w.`Average temperature (°C)`,
       w.`Reference temperature (°C)`,
       w.`Temperature Deviation (°C)`
FROM electricity_consumption_fr AS ecf
JOIN holidays AS h ON ecf.date = h.date
JOIN weather AS w ON ecf.date = w.date;
```

Conclusion

This report has provided a comprehensive analysis of energy consumption in France, including factors such as holidays and weather, using various data sources and analytical methods. Through exploratory data analysis with Python, we have been able to identify patterns in energy consumption and production over time, including seasonal variations and the impact of public holidays and the temperature. Additionally, we have discovered that despite the decreasing share of nuclear energy in the energy mix in recent years, it remains the dominant source of energy in France, while wind and solar energy are increasing in their share.

The findings from this report can provide useful insights for us as consumers and citizens. By using the historical data and identifying the key features that drive energy consumption, it is possible to be more informed and aware about energy consumption and production in the future. Furthermore, the report highlights the importance of energy security, sovereignty and the need to transition to more sustainable and renewable energy sources to reduce carbon emissions and mitigate the effects of climate change.

In conclusion, this report has contributed to the body of knowledge on energy consumption in France and provided valuable insights for people and energy management professionals. With the use of open data platforms and analytical tools, it is possible to gain a deeper understanding of energy consumption and production patterns, which can help in making informed decisions about energy management and production.

The next steps in this project are to develop a time series model that predicts energy consumption in France based on historical data.

Annexe

Project plan

The project plan was created in Jira and can be accessed here.

Github

The Github repository of the project can be accessed here.