PROJECT R : Times Series Analysis

TIME SERIES ANALYSIS IN R ( PROJECT)

In mathematics, a time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.

In the broadest definition , a time series is any data set where the values are measured at different points in time. Many time series are uniformly spaced at a specific frequency, for example, hourly weather measurements, daily counts of web site visits, or monthly sales totals. Time series can also be irregularly spaced and sporadic, for example, timestamped data in a computer system’s event log or history of 91 emergency calls.

Working with a time series of energy data we’ll see how techniques such as time-based indexing, resampling, and rolling windows can help us explore variations in electricity demand and renewable energy supply over time.

Aspects of the data set:

• The data set: Open Power Systems Data

• Time series data structures

• Time-based indexing

• Visualizing time series data

• Seasonality

• Frequencies

• Trends

Electricity production and consumption are reported as daily totals in gigawathh-hours. The columns of the data file are

\* date -> the date (yyyy-mm-dd format)

\*consumption -> electricity consumption in GWh

\*Wind w->wind power production in GWh

\* Solar -> Solar power production in GWh

\*Wind+Solar -> Sum of wind and solar power production in GWh

also in this project we will explore how electricity consumption and production in Germany have varied over time.

Questions:

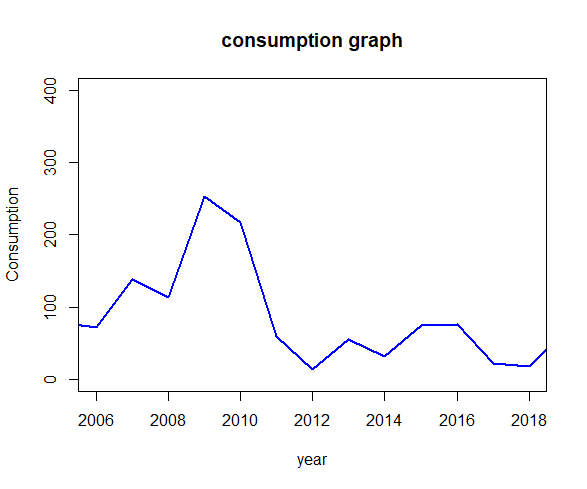
\* when is electricity consumption typically highest and lowest?

\* how do wind and solar power production vary with seasons of the year?

\* what are the long-term trends in electricity consumption, solar power and wind power?

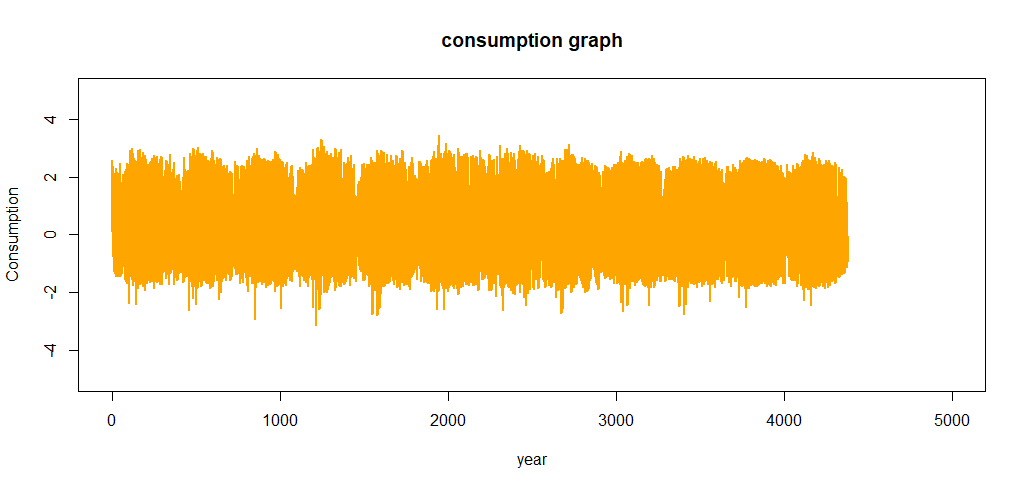
\* how do wind and solar power production compare with electricity consumption and how has this ratio changed over time ?

After wrangling / cleaning of data/pre-processing of data and creating a data frame, we can visualize the data.

question 1 : when is electricity consumption typically highest and lowest ?

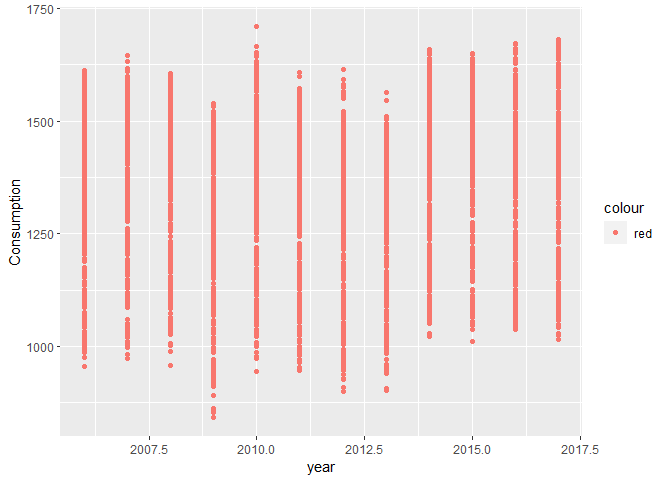
Here we can deduce obviously that consumption of electricity is reduced comparatively in 1 december 2010 where it attended its peak with 1709.568 GWH. In previous years, the consumption still not stable but it doesn’t increase that much.

Now if we want to look at the data and understand it better we can take a log value:

both of this graphs were built using plot library which has chosen pretty good locations (every two years) and labels (the years) for the x axis , which is helpful and made us insight patterns.

However with so many data points, the line plot is crowded and hard to read.

Instead we can use the ggplot library that make plots more readable and discoverable.

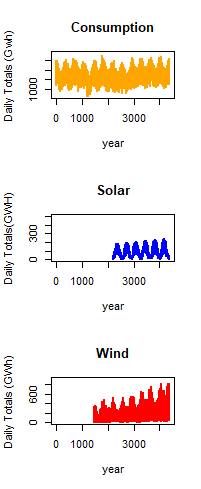


here we can see that in 2010, the consumption was too high however it was in its lowest in 12 april 2009 with 842.395 GWH.

question 2 : how do wind and solar power production vary with seasons of the year ?

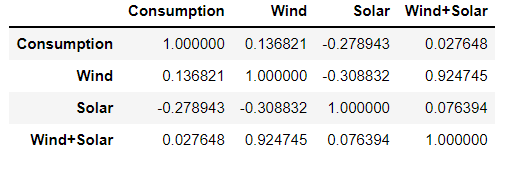
Multiple plots in the same plot helps distinguish the main patterns and differences based on the same horizontal axis which is the year and the different vertical axis which are consumption, solar and wind.

This will lead us to understand the correlation and the variability of consumption with climatic conditions over year.

The following plot is 3 plots in only one plot to helps us recognize relationships and patterns between the data and explore these patterns and focus on specific areas where there is a robust and dominant dependency between variables.

We can already see some interesting patterns emerge:

* Electricity consumption is highest in winter, presumably due to electic heating and increased lighting usage, and lowest in summer.
* Electricity consumption appears to split into two clusters – one with oscillations centered roughly around 1150 GWh, and another with fewer and more scattered data points, centered roughly around 1150 GWh.We might guess that these clusters correspond with weekdays and weekends.
* Solar power production is highest in summer where sunlight is most abundant, and lowest in winter.
* Wind power production is highest in winter, presumably due to stronger winds and more frequent storms and lowest in summer.
* There appears to be a strong increasing trend in wind power production over the years.
* All three time series clearly exhibit periodicity –often referred to as seasonality in time series analysis –in which a pattern repeats again and again at regular time intervals. The consumption, solar and wind time series oscillate between high and low values on a yearly time scale, corresponding with the seasonal changes in weather over the year. However, seasonality in general does not have to correspond with the meteorological seasons. For example, retail sales data often exhibits yearly seasonality with increased sales in November and December, leading up to the holidays.
* Seasonality can also occur on other time scales. The plot above suggests there may be some weekly seasonality in Germany’s electricity consumption, corresponding with weekdays and weekends.

 We tried to visualize correlation between the 4 variables so we used plt in python and the result returned :

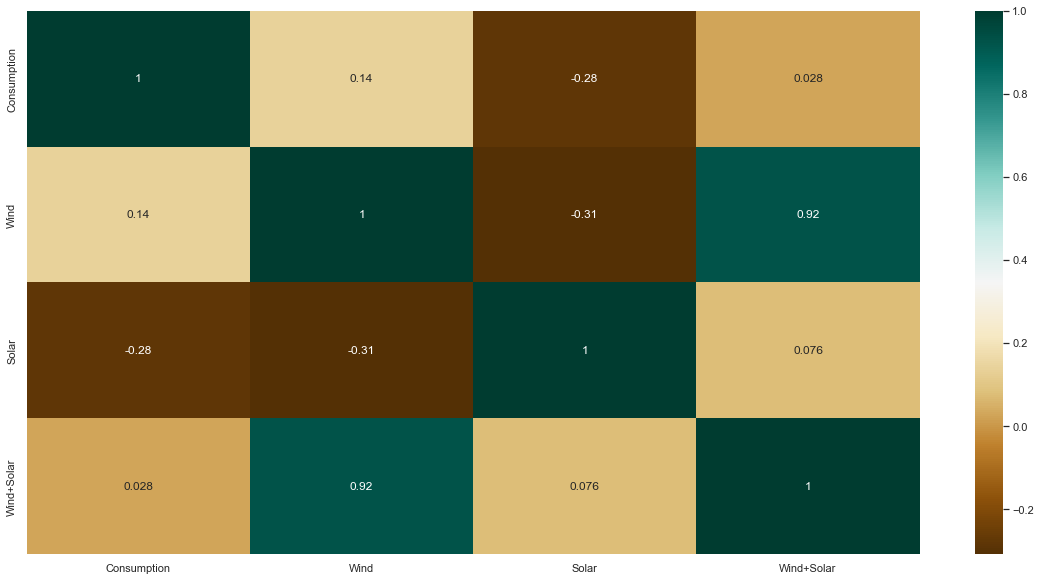
This table visualizes the dependency between variables, the more we get nearer to 1 the more is the dependency between variables .

It refers to a positive correlation.

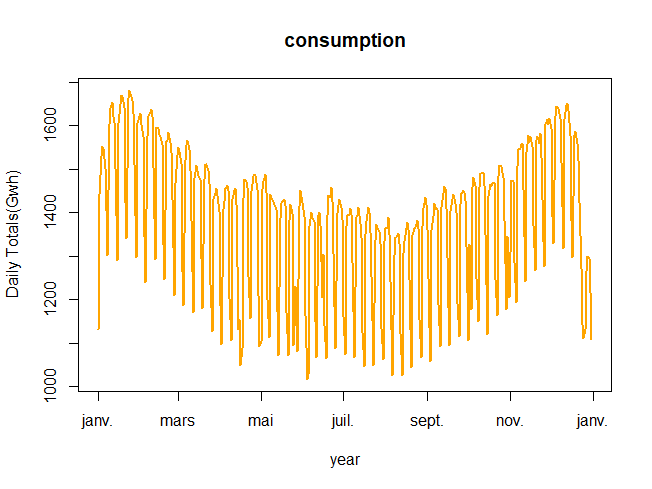
The less we get nearer to 1 the less is the dependency between variables.

The more we get to -1 the more the dependency exist but the correlation is negative. We can take the example of solar and consumption which have a negative correlation . This means when the solar’ s intensity increase the consumption of electricity decrease.

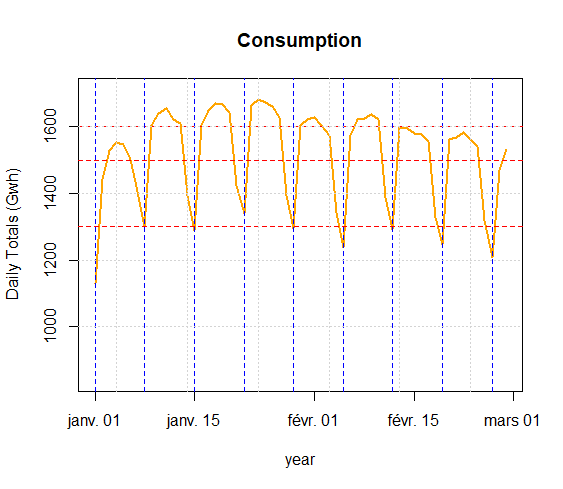
The following heat map shows the relative intensity of values captured by analyzing the correlation between variable and assign each value a color presentation.

Consumption over year is more affected by the wind than solar by a probability 0.14 compare to -0.28.

* **Now we will plot the time series in a single year to investigate further**

**Here we will work on year 2017:**

Now we can clearly see the weekly oscillations. Another interesting feature that becomes apparent at this level of granularity is the drastic decrease in electricity consumption in early January and late December, during the holidays.

* **Let’s zoom in further and look at just January and February.**

Here we can see that consumption reduces by the end of the week and increases in the middle of the week to attend its peak. Here in our graph we added solid horizontal and vertical lines to dissect the data which means to cut up the data methodically for the purpose to study its internal parts.

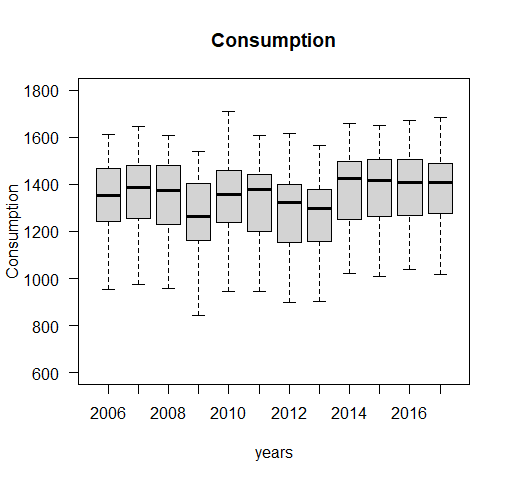
Seasonality:

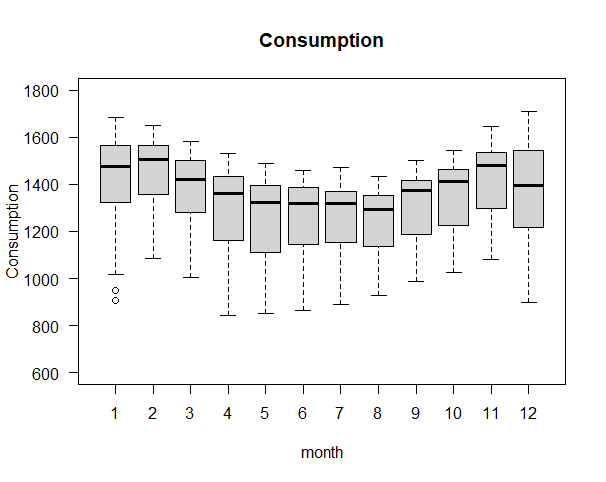
Next, let’s further explore the seasonality of our data with box plots to group the data by different time periods and display the distribution for each group.

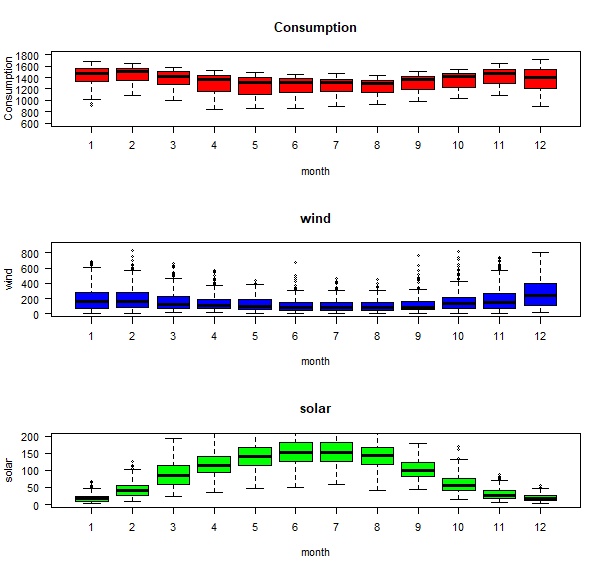
Utility of using box plot is that they provide a quick visual summary of the variability of values in a dataset. They show the median, upper and lower quartiles, minimum and maximum values, and any outliers in the dataset.

Outliers can reveal mistakes or unusual occurrences in data.

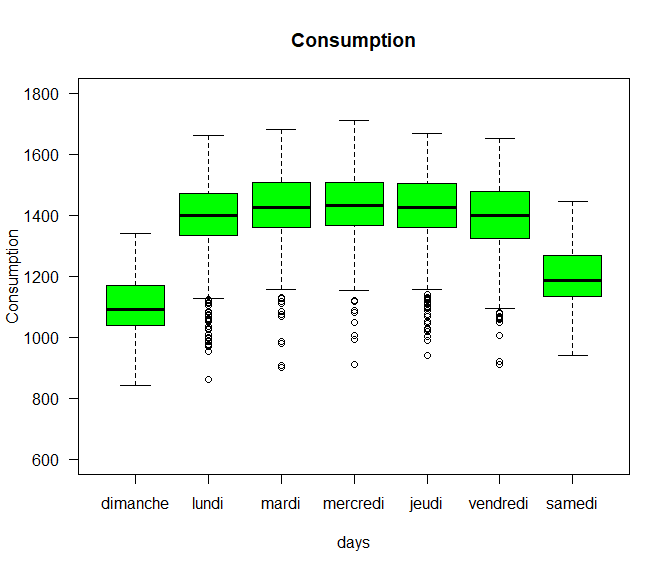
They enable the researchers to quickly identify mean values, the dispersion of the data set and signs of skewness.







These box plots confirm the yearly seasonality that we saw in earlier plots and provide some additional insights:

* Although electricity consumption is generally higher in winter and lower in summer, the median and lower two quartiles are lower in December and January compared to November and February , likely due to businesses being closed over the holidays. We saw in the time series for the year 2017, and the box plot confirm that is consistent pattern throughout the years.
* While solar and wind power production both exhibit a yearly seasonality, the wind power distributions have many outliers, reflecting the effects of occasional extreme wind speeds associated with storms and other transient weather conditions.
* **Let’s group the electricity consumption time series by day of the week, to explore weekly seasonality.** 

As expected, electricity consumption is significantly higher on weekdays than on weekends.

The low outliers on weekdays are presumably during holidays.

Time series with strong seasonality can often be well represented with models that decompose the signal into seasonality and a long-term trend, and these models can be used to forecast future values of the time series.

Frequencies:

To understand frequency of data we should take care of missing values using Fill column so that the missing can be forward filled, meaning that the last value repeats through the missing rows until the next non-missing value occurs. Filling missing values can be done in different directions.

In this case we’re doing a time series analysis which requires uniformly spaced data without any missing, so we first convert our time series to the specified frequency and fill the missing with an appropriate method.

In our situation the specified frequency wasn’t afforded from the beginning that’s why we did calculate it using frequency monthly, weekly, and yearly.

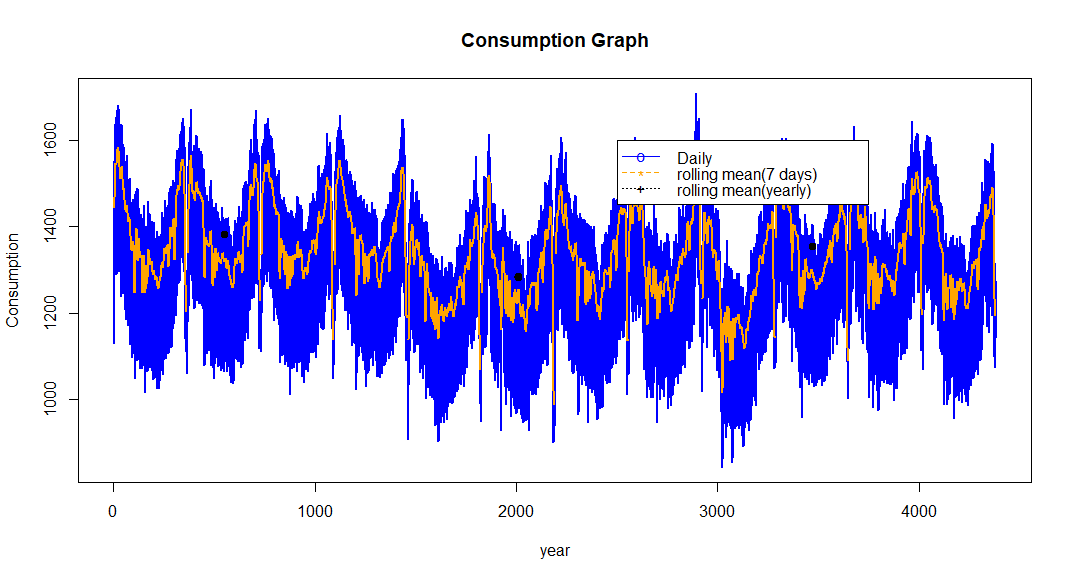
Here we’ve chosen day wise frequency seeing that we’ve noticed that our missing is linked to specific dates and not years or months.

Une image contenant texte

Description générée automatiquementBelow we can see the result of our null values cleaning and transforming:

Trends:

Time series data often exhibit some slow, gradual variability in addition to higher frequency variability such as seasonality and noise. An easy way to visualize theses trends is with rolling or moving means at different time scales.

A rolling mean tends to smooth a time series by averaging out variations at frequencies much higher than the window size and averaging out any seasonality on a time scale equal to the window size. This allows lower-frequency variations in the data to be explored. Since our electricity consumption time series has weekly and yearly seasonality, let’s look at rolling means on those two-time scales.

We can see that the 7-day rolling mean has smoothed out all the weekly seasonality while preserving the yearly seasonality.

The 7-day rolling mean reveals that while electricity consumption is typically higher in winter and lower in summer, there is a dramatic decrease for a few weeks every winter at the end of December and beginning of January, during the holidays.

Looking at the 365\_day rolling mean time series, we can see that the long-term trend in electricity consumption is pretty flat, with a couple of periods of anomalously low consumption around 2009 and 2012-2013.