Solution Series 5

The topic of this exercise is data cleansing. Thus, the procedures are performed on one example in different steps.

The task is the following: One have to perform an analysis if the average energy prices per day (convention: this means the average energy price per day for 1 hour, energy prices are always given on an hourly basis) are dependent on the weather.

The following original datasets are provided:

RainData.csv	:	rain in liter per square meter per date
TempData.csv	:	average temperature, maximum temperature and minimum temperature in degree Celcius per date
WeatherOtherData.csv	:	average wind speed, average, maximum and minimum humidity, average, maximum and minimum air pressure, length of a day
PowerData_2010.csv	:	Power price per hour e.g. Hour 1 means the price for one hour of energy in the time period 24 – 1 o'clock, Hour 8 means the price for one hour of energy in the time period 7 – 8 o'clock, and so on.

As you are free in choosing a tool the following gives the step by step procedure which errors should be found and how to solve them.

1) Preparing the data

Weather data:

Technically correctness:

i) One has the following values:

AVG_HUMIDITY	MAX_HUMIDITY	MIN_HUMIDITY
81.40	82,00	81,00 %

Therefore, first, correct "," to "." and deleting the "%" in MIN_HUMIDITY

ii) The second technical correction is in air pressure

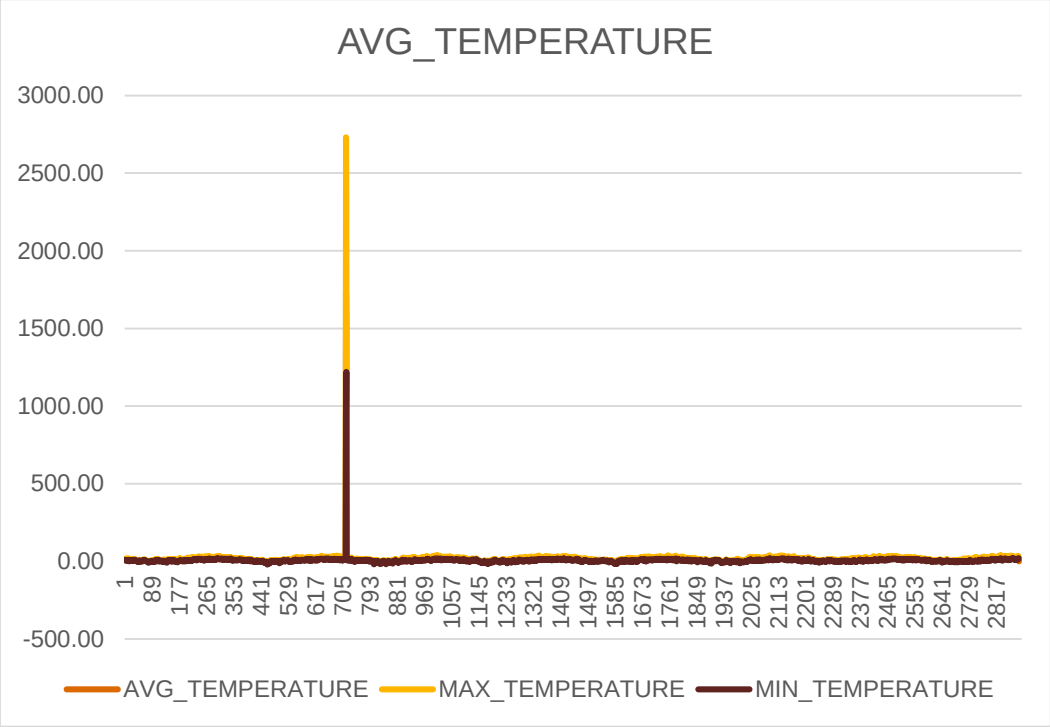
AVG_AIR_PRESSURE	MAX_AIR_PRESSURE	MIN_AIR_PRESSURE
1021.64	1021,80	1021,40 hPa

Correct "," to "." and delete "%" in MIN_AIR_PRESSURE

Consistent data:

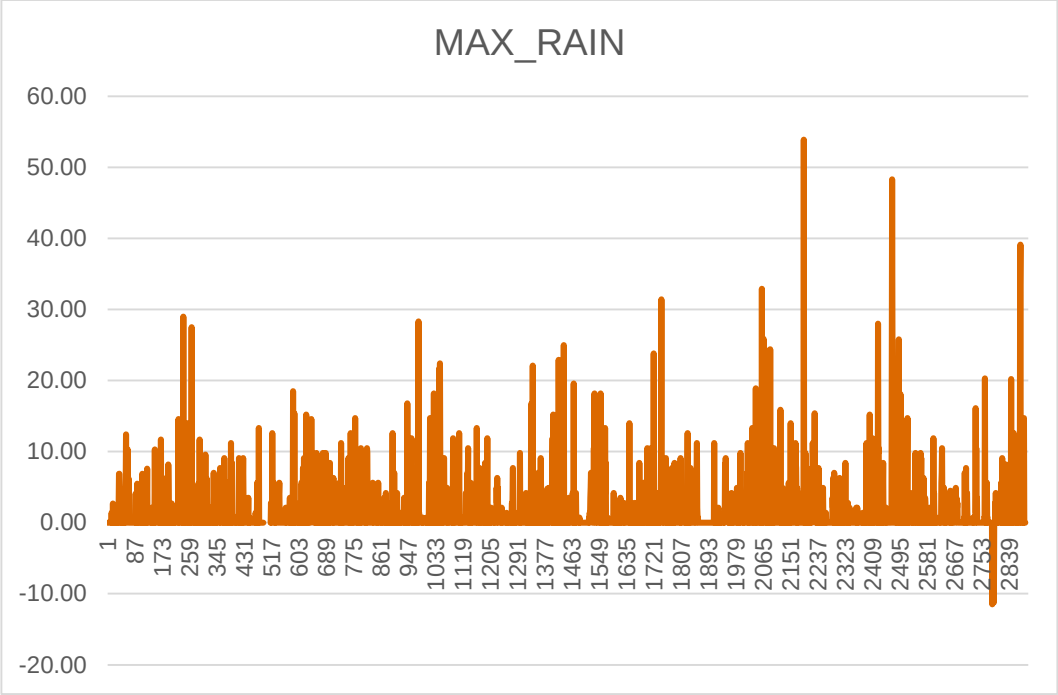
Temperature data:

i) On the dates 29.08.2008, 30.08.2008 and 31.08.2008 the average temperature is missing. Replace them e.g. with the mean of MAX_TEMPERATURE and MIN_TEMPERATURE.
ii) On the dates 18.09.2009, 19.09.2009 and 20.09.2009 there the data are too high by a factor 100, i.e. the "." was shifted by two digits.
iii) On the dates 19.03.2010 - 27.03.2010, there are NA in the average temperature. Replace them e.g. . with the mean of MAX_TEMPERATURE and MIN_TEMPERATURE.



Rain data:

i) On the dates 07.02.2009 - 28.02.2009 there are no entries in MAX_RAIN. As there nearly one month with no rain data. If this would be used as "0" in the climate chart then chart would be wrong. Thus, omit February 2009 to generate the climate chart.
ii) On the dates 20.05.2015 - 25.05.2015 there are negative entries for the rain amount. One possibility is to replace these values by an estimate based on average rain in this month.



Pressure data:

i) On the dates 02.08.2015, 03.08.2015 and 04.08.2015 the data in AVG_AIR_PRESSURE and MIN_AIR_PRESSURE are not plausible and for MAX_AIR_PRESSURE on the 03.08.2015. They can just be omitted.

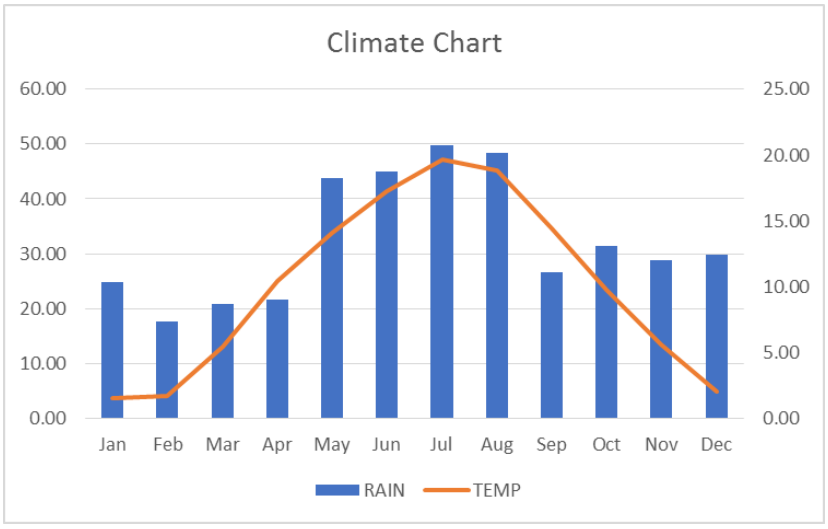
Finally, add new data fields “year” and “month” as they are needed for doing the plots.

Power data:

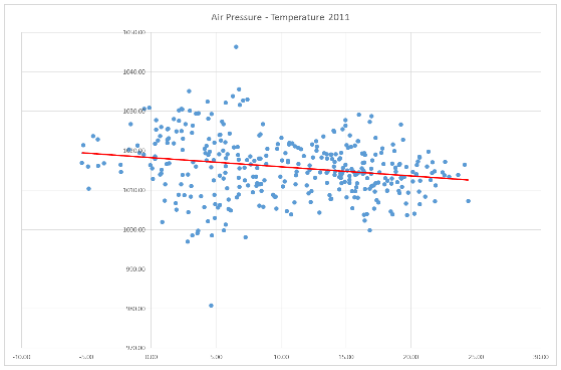
The hourly price data has to be aggregated to a daily base price on an hourly basis which means just to take the average over all the hours and include a new column in the data.

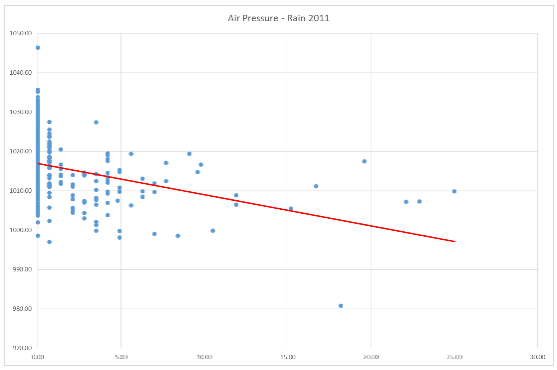
2) Plots and correlations

Climate diagram:



Correlation of air pressure and temperature or air pressure and rain in the year 2011:



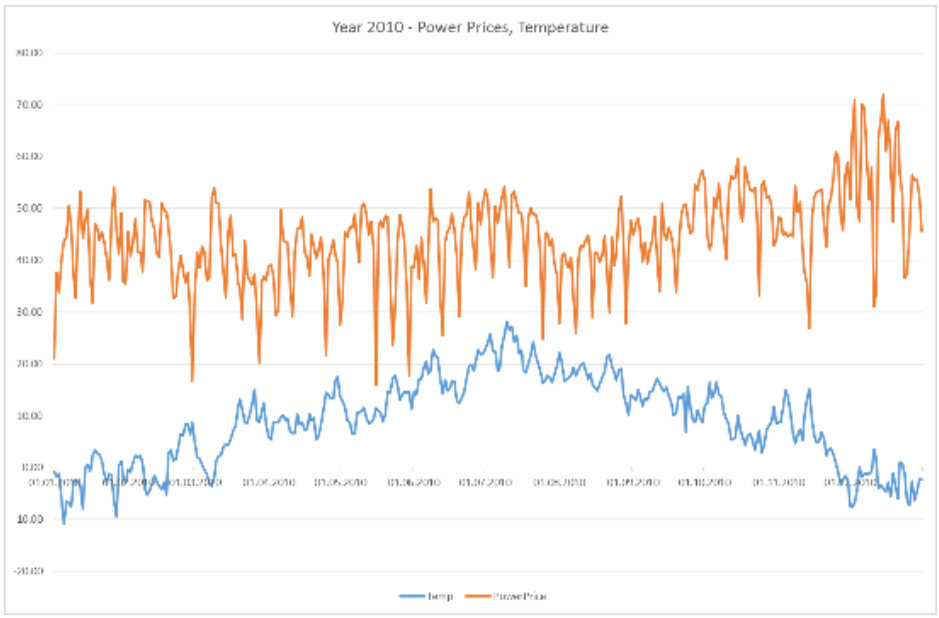


Correlation Air Pressure / Temperature: -0.21

Correlation Air Pressure / Rain: -0.35

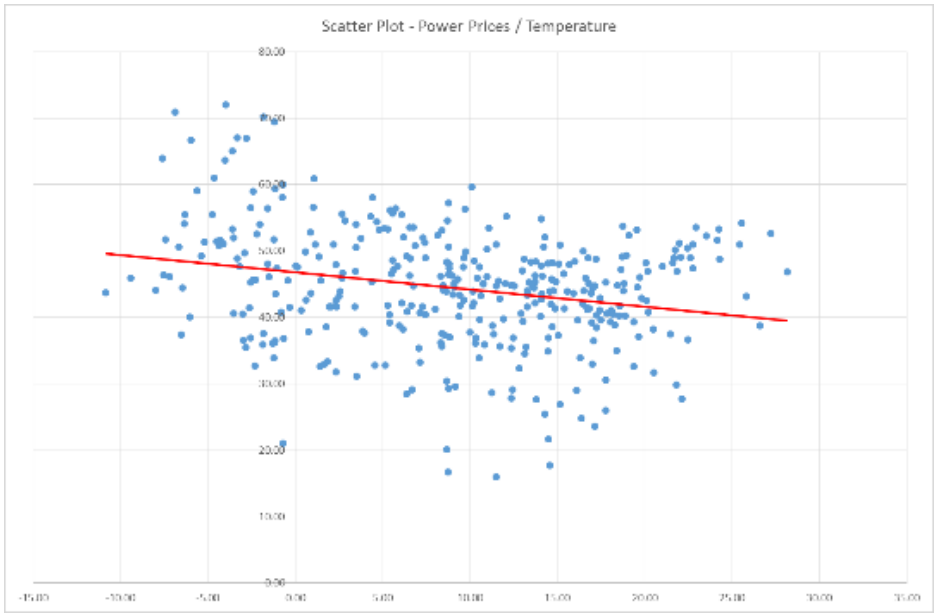
3) Dependencies between temperature and power prices:
Our expectation would be that during periods with low temperature more energy is used for heating and thus, the energy price is rising. I.e. there should be a negative correlation.

Plots (as only 2010 data are available for the power prices, only the year 2010 is plotted):



In the time series plot indeed one can see in the right hand side of the plot an increase in the power prices when the temperature are decreasing. Nevertheless, one have expected a stronger dependencies than one can see in the plot.

Further, the peaks down are the weekly seasonality of the week-ends.



Correlation Power Prices / Temperature: -0.25