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.

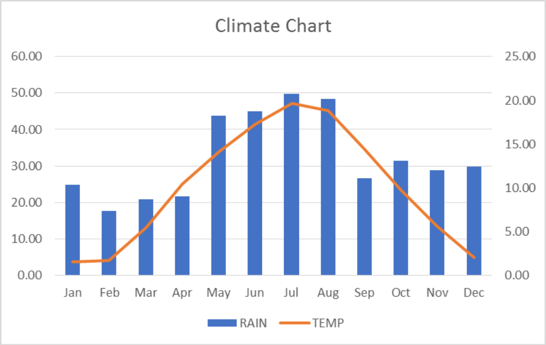
You are free to choose the tool (R, Python, Matlab, Julia, Excel, SPSS, SAS and so on) that you want to use to solve this exercise. In the exercise lecture an short introduction to Python, pandas (Python Data Analysis Library) and Anaconda (package and environment manager for Python) are given.

1) Aggregate all weather data sets to one database.

2) Review and clean up weather data.
   - i) Technically correct data like "," and ".", different notation for the same variable, e.g. for air pressure (hint: there is a second variable which has to be cleansed up)
   - ii) Consistent data: clean up the data and describe what type of error it is.
     Hint:
     - 3 types of errors to correct in the temperature data
     - 2 types of errors to correct in the rain data
     - 1 type of error to correct in the air pressure data
   - iii) Add new data fields "year" and "month".

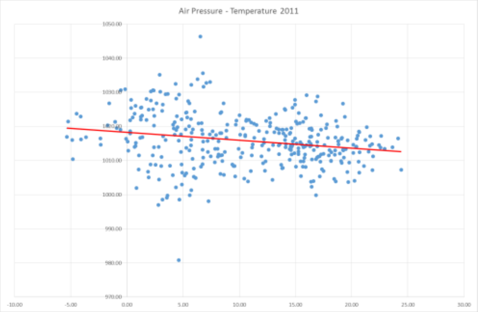
3) Create a climate chart:

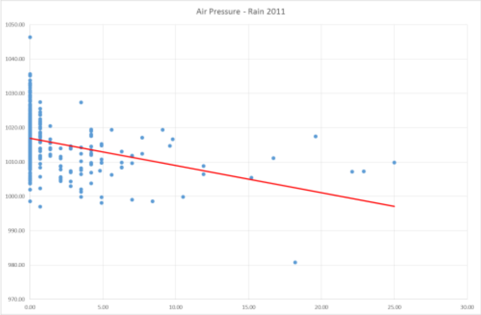
Hint:



4) Is there a correlation of air pressure and temperature or air pressure and rain in the year 2011?

Hint:





What are other methods to test if there are any relationship between these variables?

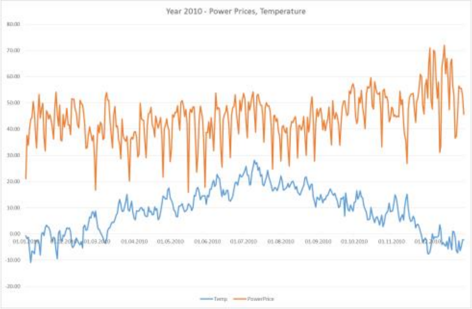
5) If this would be a large dataset, which data would you delete before processing further on?

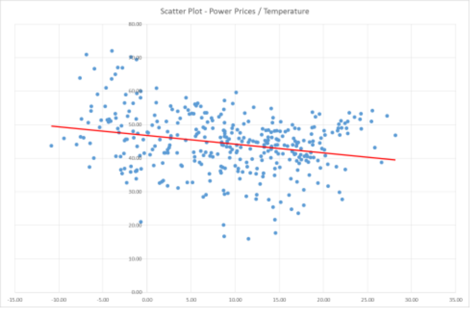
6) Review and clean up the Power Data
   i) Technical correct data: ---
   ii) Consistent data: ---
   iii) Add a new field: average energy price per day (hint: see convention above, and it has to be arbitrage-free)

7) Aggregate the weather data and the energy price data

8) Can you detect a dependency of temperature with power prices? What would you expect and why? (hint: only 2010 data are available for the power data)

Hint:


Year 2010 - Power Prices, Temperature
Temp
PowerPrice


Scatter Plot - Power Prices / Temperature