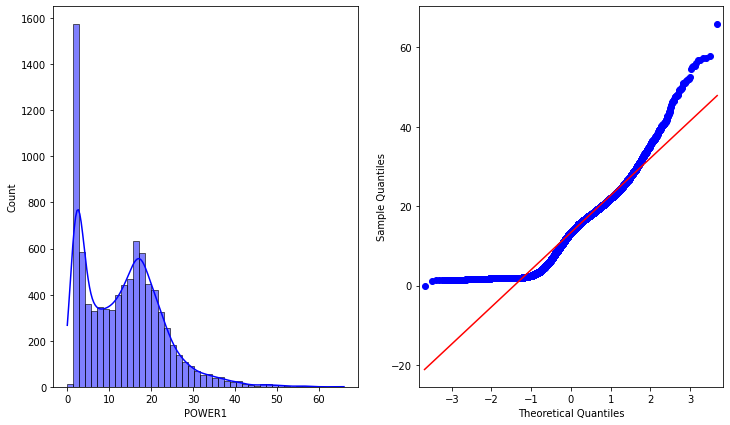
**Implementing The Outlier Detection and Imputation Process**

The outlier identification is an important step since it indicates whether we have a bad data resulted from experiment that may have not run-in proper way. In general, the outlying values should be deleted. In this study the outlier detection process is applied on the heat load since it is the target variable to be predicted.

Identification of outliers requires the normality assumption of our data, where the observations follow an approximately normal distribution. For this reason, the normal probability plot and histogram have been used as methods to test for normality and identifying potential outliers.



**(a) (b)**

**Figure 1. Histogram and normal probability plot used for the normality test: (a) frequency (count) vs heat load, (b) sample quantiles (actual data) vs theoretical quantiles (z-score)**

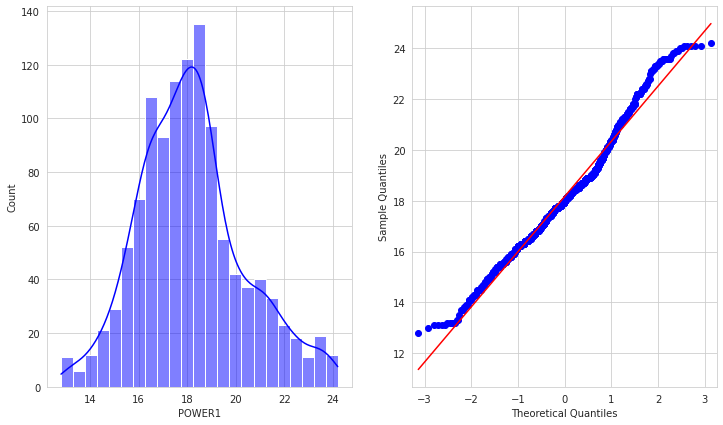
In the sub figure (a) we can see a large deviation from the normal distribution curve. In the normal probability plot (b) the line is largely skewed to the left. As a consequence, we can say that we do not have a normally distributed data.

With the aim of dealing with outliers and reading errors, we propose the interquartile range method (IQR). The aforementioned method measures the spread of our data and represent the range for the middle of the samples.

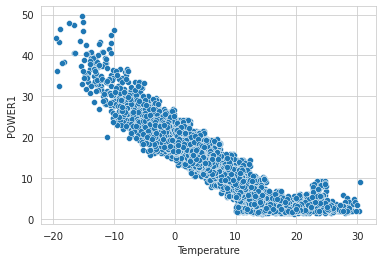
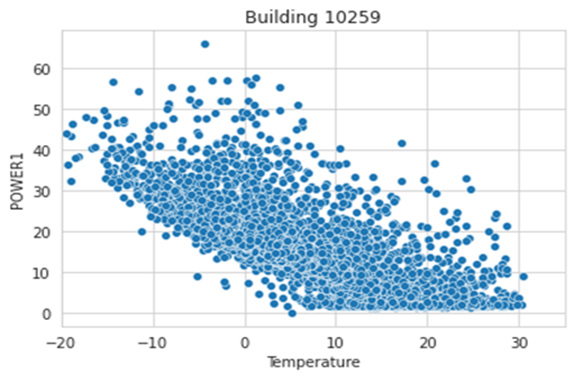
Conversely, we cannot carry out this process directly since our dataset is not randomly distributed due to climatic reasons. As a result, we have performed a preliminary step in which we have divided the data into 21 subsets of temperature. Each temperature subset is marked at an interval of 2.5 °C.

We have taken one of the subsets as an example (figure 2. below) to show how did we apply the outlier detection process using the IQR method. In this context we have selected a range of data lying between 0°C and 2.5°C (representing a subset) in order to make the normality assumption valid.

In order to fill the missing value resulted from the outlier removal, an imputation process have been executed by filling the heat load missing value by the average from the same hour taken from the rest of the data. Knowing that this process has been done for every subset of temperature.



**Figure 2. Histogram (a) and normal probability plot (b) after applying the IQR method to the data subset (between 0°C and 2.5°C)**



**(a) (b)**

**Figure 3. Hourly heat load vs outdoor temperature in building 10259: (a) before data cleaning, (b) after data cleaning**

All observation lying above the third quartile + 1.5\*IQR and lying below the first quartile - 1.5\*IQR is considered as outliers. In our case we have replaced the 1.5 constant with 1 in order to make it less sensitive to the abnormal values. Hence, a total of 4 % of probability of lying in the outlier zone were identified.