# 3. Outliers detection

Some rules can be applied to filter out some of the CPCs having **unusual behaviors** that could badly impact the performance of the weighting model or the selection of the comparables.

To handle outliers, we developed a dedicated module *outliers_detection.py* in the common code base.

This module allows us to filter outliers based on several methods.

## List of methods

### Mean_Std (Mean and Standard Deviation Method)

This method identifies outliers by calculating the mean and standard deviation of the dataset.

An observation is considered an outlier if it lies beyond a specified number of standard deviations from the mean.

Common thresholds are 2 or 3 standard deviations.

### IQR (Interquartile Range Method)

The IQR method uses the interquartile range, which is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data.

Outliers are typically defined as observations that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. The 1.5 threshold can be adapted.

This method is robust to non-normal distributions.

### Hampel Method

The Hampel method is a robust statistical technique that identifies outliers based on the median and the median absolute deviation (MAD).

An observation is considered an outlier if it deviates from the median by more than a certain threshold, often set at 3 times the MAD.

This method is less sensitive to extreme values compared to the Mean_Std method.

## Flexibility

For each of the described method, we can:

- Adapt the threshold used.
- Detect outliers based on any combination of unit price, volume and sales.
- Decide to exclude only low values, high values or both.

Note that as of now, the outliers detection parameters are **identical for all the product families** of a GBU.

## GBU application

### SpP

For SpP, the outliers are currently based on the IQR method, with the following parameters:

- Threshold of **3**
- Based on **unit price** only
- Excluding only **low values**

Note that this combination only detects outliers for some of the families, especially the ones where the standard deviation is not too high, meaning there are no products with very high prices.

For families with expensive products like Fluids or Tecnoflon FFKM, the lower bound of the IQR can end up being negative, so every CPC with a positive price is kept. If exclusions on low prices need to be done on these families, the outliers parameters should be reviewed accordingly.

For high prices CPC impacting the performance of the models, we decided to exclude them on a product basis as detailed here.

Novecare