**Methodology**

As part of this paper’s methodological analysis of the provided data, the core procedure has consisted of a linear bivariate regression, as well as multiple regression analysis. One of the key aspects of our analysis consisted in creating the necessary granularity in the data to obtain relevant and actionable insights regarding energy and water prices across different cities in the country. Thus, not only have we made a division on the location, but also on the type of

As a first step, we have cleaned our data set from any incomplete entry, as an initial look at the set showed that water and energy bill price data (*totener*) was lacking in a good number of entries. This reduced our total sample from *n = 7789* to *n = 7093.*

Secondly, the elimination of outliers before the data processing was minimal, as the original data skewness was expected to have a larger right-tail, due to the observation of natural numbers. Not only that, but the team expected, prior to the analysis and data partition, to see a lot of variance and skewness. Thus:

* The team considers that our data is sufficiently large to ignore the existence of outliers as means to acquire a more accurate and representative price trend of each city – meaning, we expect some cities to have a more acute increase in bills price than others. We believe this difference may be obfuscated by eliminating many outliers.
* In the current real estate market, there is not much regulation for the top-end, but this subsection will still be smaller than the bottom one, meaning the skewness will be large.
* We expected the number of people per house to increase in smaller properties, with more modest energy and water consumption price. Removing outliers may potentially obfuscate these differences.

The goal of data partitioning, which made up for the next step in our process, was to smoothen out the variance of most, if not all, properties listed. In order to increase such granularity, we included the type of contract that ties residents to the property – whether they *rent* or *own* the property.

This leaves us with eight subsets of data: four of which are *rentals* in *The Hague, Amsterdam, Utrecht* and *Rotterdam;* and another four who are *owner-occupied* properties, in the same cities.

Not only is our intention to use descriptive statistics to gauge an overall trend, but also to create a linear model in which we are able to predict the expenditure on bills based off these following factors: *City, Type of Contract, Number of people in the household, House surface.*