**WATER POTABILITY | PRIMARY ANALYSIS**

**|Number of Missing Values**

“pH”: 491

“Sulfate”: 781

“Trihalomethanes”: 162

The dataset has no missing values for the other variables.

**|Initial mean of the variables**

Mean of the variables excluding missing values, without data cleaning.

“pH”: 7,08

“Hardness”: 196,37

“Solids”: 22014,09

“Chloramines”: 7,12

“Sulfate”: 333,78

“Conductivity”: 426,21

“Organic\_carbon”: 14,28

“Trihalomethanes”: 66,40

“Turbidity”: 3,97

**|Number of Outliers**

“pH”: 46

“Hardness”: 83

“Solids”: 47

“Chloramines”: 61

“Sulfate”: 41

“Conductivity”: 11

“Organic\_carbon”: 25

“Trihalomethanes”: 33

“Turbidity”: 19

**|Mean of the variables after data cleaning**

Mean of the variables after data cleaning. The missing values were replaced with the means calculated without the outliers, considering the class.

“pH”: 7,08 (approximately unchanged)

“Hardness”: 196,50 (small increase)

“Solids”: 21624,78 (decreased)

“Chloramines”: 7,12 (approximately unchanged)

“Sulfate”: 333,91 (small increase)

“Conductivity”: 425,50 (small decrease)

“Organic\_carbon”: 14,29 (approximately unchanged)

“Trihalomethanes”: 66,48 (small increase)

“Turbidity”: 3,97 (approximately unchanged)

The means haven’t changed significantly, suggesting that the outliers were not excessively influencing the mean of the variables.