**Data Filtering Based on Total Dissolved Solids (TDS):** One of the key steps in our data cleaning process is filtering out samples with a TDS (Total Dissolved Solids) of less than 35,000 ppm (sea water). TDS is a measure of the combined content of all organic and inorganic substances contained in a liquid; in this case, the produced water from various sources like coalbed methane, shale gas, and conventional hydrocarbon wells. I filtered out lower TDS samples because they typically require less complex treatment processes and pose a lower environmental risk. Our focus should be on high TDS samples, which are more challenging in terms of treatment and management. By doing this, I narrowed down the dataset from 114,944 rows to those that are more relevant for studying high salinity produced water.

**Molarity Calculations:** I converted the concentrations of elements (like Na, Ca, Cl, etc.) from ppm (parts per million) to Molarity. Molar concentration is defined as the number of moles of a substance per liter of solution. A mole is a standard unit in chemistry that is used to measure the amount of a substance. One mole of any substance contains the same number of entities (like molecules or atoms) as there are in 12 grams of carbon-12. This number is approximately 6.022×1023, known as Avogadro's number. The formula for calculating molar concentration is:

Molarity (M)=

For example, if you dissolve 1 mole of salt (NaCl) in 1 liter of water, the molarity of that salt solution is 1 M. The formula I used to convert ppm or mg/L to molarity is:

Molarity (M) =

This conversion is essential for accurately comparing the concentrations of different elements in the water.

**Applying Conditions for Filtering:** I implemented specific criteria to refine the dataset further. These criteria focused on the molar concentration relationships between various elements, such as Sodium (Na), Calcium (Ca), Chlorine (Cl), etc. The rationale behind this step was to identify and exclude samples with chemical profiles that did not meet these conditions. The exclusion of such samples is critical because their chemical characteristics suggest that they might be unnatural or have problematic analyses. Essentially, these samples are considered less reliable or valid for the purposes of this study.

**Charge Balance Calculations:** In simple terms, charge balance is making sure that the total positive charges in water (from cations) are equal to the total negative charges (from anions). This balance is crucial because if it's off, it might mean there are mistakes in how we measured the ions in the water. I calculated this balance for each water sample by adding up all the cations and anions and then seeing how close they are to being equal. This was done as a percentage difference, which I called the “Calculated Charge Balance”. I also checked this against the ‘chargebalance’ data from the USGS. Since they matched up well, I used the USGS's data to further clean our dataset. Generally, in water studies, a charge balance between -10 and +10 is seen as good enough. It means the positive and negative charges are pretty balanced. So, I only kept data where the charge balance was within this -10 to +10 range.