The purpose of this project is to dive deep into a survey that was conducted about "Maji Dogo". A community that is faced with a lot of water problems. My aim is to understand the data and remove errors from it as much as possible, thereby making it more reliable for data driven decision making about how to solve the water crisis in the community.

The first thing I did is to familiarize myself with the data by looking up the various tables and their columns.

I then check the various water sources that are involved in our data set

An important note on the home taps: About 6-10 million people have running water installed in their homes in Maji Ndogo, including
broken taps. If we were to document this, we would have a row of data for each home, so that one record is one tap. That means our
database would contain about 1 million rows of data, which may slow our systems down. For now, the surveyors combined the data of
many households together into a single record.

For example, the first record, AkHa00000224 is for a tap_in_home that serves 956 people. What this means is that the records of about 160 homes nearby were combined into one record, with an average of 6 people living in each house 160 x 6 ≈ 956. So 1 tap_in_home or tap_in_home_broken record actually refers to multiple households, with the sum of the people living in these homes equal to number_of_people_served.

I then check water sources that have long queue times which I capped at atleast 8 hours or 500 min.

The quality of our water sources is the whole point of this survey.

Scores have been assigned to each source from 1, being terrible, to 10 for a good, clean water source in a home. Shared taps are not rated as high, and the score also depends on how long the queue times are. All this info is stored in the water quality table.

One obvious water source that was polluted was wells. But that part of the data has issues with some of the wells being declared as being clean while they were not.

Thus either they had chemicals or biological pollutants in them. Therefore I wrote a query to correct the situation. Based on the results, each well was classified asClean,Contaminated: Biological or Contaminated: Chemical. It is important to know this because wells that are polluted with bio- or
Other contaminants are not safe to drink.

To make sure the query I wrote was correct, I first had to make a copy of the well pollution table and run it in it first.