This project is the last part of one big SQL project. I started with part 1 where I got a feel of the data. Thus knowing each table and the column within it. In the second part of the project, I concentrated my efforts on finding patterns in the data set. Also I looked out for inconsistencies and did a bit of data cleaning. The third part of the data had a little data cleaning as well but the main agenda was to audit our data to make sure it was valid and that we can therefore proceed to use it for decision making. This part of the project, which is the final part, will also have a bit of data analysis and I will concentrate on finding ways to track our work. In terms of providing solutions to the water crisis in Maji dogo.

Let's summarize the data we need, and where to find it:

- All of the information about the location of a water source is in the location table, specifically the town and province of that water source.
- water_source has the type of source and the number of people served by each source.
- visits has queue information, and connects source_id to location_id. There were multiple visits to sites, so we need to be careful to
include duplicate data (visit_count > 1).
- well_pollution has information about the quality of water from only wells, so we need to keep that in mind when we join this table.

The first thing I want to find out is whether there are any specific provinces, or towns where some sources are more abundant?

To answer the question, we will need province_name and town_name from the location table. We also need to know the type_of_water_source and number_of_people_served from the water_source table.

The problem is that the location table uses location_id while water_source only has source_id. So we won't be able to join these tables directly. But the visits table maps location_id and source_id. So if we use visits as the table we query from, we can join location where the location_id matches, and water_source where the source_id matches.

Before we can analyze, we need to assemble data into a table first. It is quite complex, but once we're done, the analysis is much simpler!

Start by joining locations to visit and then add water source table. add the visits.visit_count = 1 as a filter, to select rows where visits.visit_count = 1.

Last one! Now we need to grab the results from the well_pollution table.
This one is a bit trickier. The well_pollution table contained only data for well. If we just use JOIN, we will do an inner join, so that only records
that are in well_pollution AND visits will be joined. We have to use a LEFT JOIN to join theresults from the well_pollution table for well
sources, and will be NULL for all of the rest.

So this table contains the data we need for this analysis. Now we want to analyze the data in the results set. We can either create a CTE, and then query it, or in my case, I'll make it a VIEW. I'll call it the combined_analysis_table.

This view creates a "table" that pulls all of the important information from different tables into one.

The last analysis
We're building another pivot table! This time, we want to break down our data into provinces or towns and source types. If we understand where the problems are, and what we need to improve at those locations, we can make an informed decision on where to send our repair teams.

province_totals is a CTE that calculates the sum of all the people surveyed grouped by province. If you replace the query above with this one:

```
SELECT
*
FROM
Province_totals;
```

You should get a table of province names and summed up populations for each province.

The main query selects the province names, and then like we did last time, we create a bunch of columns for each type of water source with CASE statements, sum each of them together, and calculate percentages.

We join the province_totals table to our combined_analysis_table so that the correct value for each province's pt.total_ppl_serv value is used.

Finally we group by province_name to get the provincial percentages.

Patterns that i see:

• Looking at the river column, Sokoto has the largest population of people drinking river water. We should send our drilling equipment to Sokoto first, so people can drink safe filtered water from a well.

• The majority of water from Amanzi comes from taps, but half of these home taps don't work because the infrastructure is broken. We need to send out engineering teams to look at the infrastructure in Amanzi first. Fixing a large pump, treatment plant or reservoir means that thousands of people will have running water. This means they will also not have to queue for water, so we improve two things at once.

Let's aggregate the data per town now. You might think this is simple, but one little town makes this hard. Recall that there are two towns in Maji Ndogo called Harare. One is in Akatsi, and one is in Kilimani. Amina is another example. So when we just aggregate by town, SQL doesn't distinguish between the different Harare's, so it combines their results.

To get around that, we have to group by province first, then by town, so that the duplicate towns are distinct because they are in different towns.

Here the temporary table calculates town_totals which returns three columns:

province_name,
town_name,
total_ppl_serv.

In the main query we select the province_name and the town_name and then calculate the percentage of people using each source type, using the CASE statements.
Then we join town_totals to combined_analysis_table, but this time the town_names are not unique, so we have to join town_totals, but we check that both the province_name and town_name matches the values in combined_analysis_table.

Then we group it by province_name, then town_name.

There are still many gems hidden in this table. For example, which town has the highest ratio of people who have taps, but have no running water?

We can see that Amina has infrastructure installed, but almost none of it is working, and only the capital city, Dahabu's water infrastructure works.

Insights from the first project till now:
Ok, so let's sum up the data we have.
A couple of weeks ago we found some interesting insights:
1. Most water sources are rural in Maji Ndogo.

2. 43% of our people are using shared taps. 2000 people often share one tap.

3. 31% of our population has water infrastructure in their homes, but within that group,

4. 45% face non-functional systems due to issues with pipes, pumps, and reservoirs. Towns like Amina, the rural parts of Amanzi, and a couple of towns across Akatsi and Hawassa have broken infrastructure.

5. 18% of our people are using wells of which, but within that, only 28% are clean. These are mostly in Hawassa, Kilimani and Akatsi.

6. Our citizens often face long wait times for water, averaging more than 120 minutes:
- Queues are very long on Saturdays.
- Queues are longer in the mornings and evenings.
- Wednesdays and Sundays have the shortest queues

Plan of action

1. We want to focus our efforts on improving the water sources that affect the most people.
 Most people will benefit if we improve the shared taps first.

2. Wells are a good source of water, but many are contaminated. Fixing this will benefit a lot of people.

3. Fixing existing infrastructure will help many people. If they have running water again, they won't have to queue, thereby shorting queue times for others. So we can solve two problems at once.

4. Installing taps in homes will stretch our resources too thin, so for now if the queue times are low, we won't improve that source.

5. Most water sources are in rural areas. We need to ensure our teams know this as this means they will have to make these repairs/upgrades in rural areas where road conditions, supplies, and labour are harder challenges to overcome.

Practical solutions:

1. If communities are using rivers, we will dispatch trucks to those regions to provide water temporarily in the short term, while we send out crews to drill for wells, providing a more permanent solution. Sokoto is the first province we will target.

2. If communities are using wells, we will install filters to purify the water. For chemically polluted wells, we can install reverse osmosis (RO) filters, and for wells with biological contamination, we can install UV filters that kill microorganisms , but we should install RO filters too. In
the long term, we must figure out why these sources are polluted.

3. For shared taps, in the short term, we can send additional water tankers to the busiest taps, on the busiest days. We can use the queue time pivot table we made to send tankers at the busiest times. Meanwhile, we can start the work on installing extra taps where they are needed.
According to UN standards, the maximum acceptable wait time for water is 30 minutes. With this in mind, our aim is to install taps to get queue times below 30 min. Towns like Bello, Abidjan and Zuri have a lot of people using shared taps, so we will send out teams to those
towns first.
4. Shared taps with short queue times (< 30 min) represent a logistical challenge to further reduce waiting times. The most effective solution, installing taps in homes, is resource-intensive and better suited as a long-term goal.

5. Addressing broken infrastructure offers a significant impact even with just a single intervention. It is expensive to fix, but so many people can benefit from repairing one facility. For example, fixing a reservoir or pipe that multiple taps are connected to. We identified towns like Amina, Lusaka, Zuri, Djenne and rural parts of Amanzi seem to be good places to start.

A practical plan

Our final goal is to implement our plan in the database.
We have a plan to improve the water access in Maji Ndogo, so we need to think it through, and as our final task, create a table where our teams
have the information they need to fix, upgrade and repair water sources. They will need the addresses of the places they should visit (street

address, town, province), the type of water source they should improve, and what should be done to improve it.

We should also make space for them in the database to update us on their progress. We need to know if the repair is complete, and the date it was completed, and give them space to upgrade the sources. Let's call this table Project_progress.

At a high level, the Improvements are as follows:

1. Rivers → Drill wells
2. wells: if the well is contaminated with chemicals → Install RO filter
3. wells: if the well is contaminated with biological contaminants → Install UV and RO filter
4. shared_taps: if the queue is longer than 30 min (30 min and above) → Install X taps nearby where X number of taps is calculated using X = FLOOR(time_in_queue / 30).
5. tap_in_home_broken → Diagnose local infrastructure

To make this simpler, we can start with this query:

```
–– Project_progress_query
SELECT
location.address,
location.town_name,
location.province_name,
water_source.source_id,
water_source.type_of_water_source,
well_pollution.results
FROM
water_source
LEFT JOIN
well_pollution ON water_source.source_id = well_pollution.source_id
INNER JOIN
visits ON water_source.source_id = visits.source_id
INNER JOIN
```

```
location ON location.location_id = visits.location_id
```

It joins the location, visits, and well_pollution tables to the water_source table. Since well_pollution only has data for wells, we have to join those records to the water_source table with a LEFT JOIN and we used visits to link the various id's together.

First things first, let's filter the data to only contain sources we want to improve by thinking through the logic first.
1. Only records with visit_count = 1 are allowed.
2. Any of the following rows can be included:
a. Where shared taps have queue times over 30 min.
b. Only wells that are contaminated are allowed -- So we exclude wells that are Clean
c. Include any river and tap_in_home_broken sources.

With this table now, we can start solving our water problems.