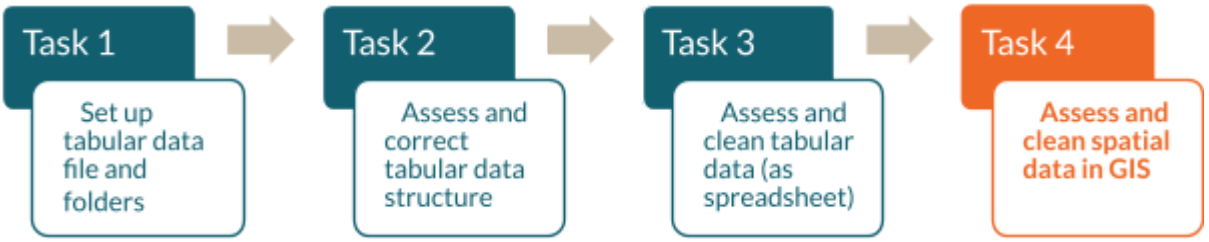
GIS DATA PREPARATION AND MANAGEMENT

Practical Task 4: Using tools in QGIS for data cleaning and management

Task description

In this practical task, you will utilize various tools and techniques within QGIS to clean and manage spatial data effectively. You will address common issues such as misalignment, anomalies, and data inconsistencies.

Tasks include revising metadata records, correcting anomalies based on visual inspection, cleaning up attribute tables, and utilizing processing tools to refine the dataset. Additionally, you will employ rule-based symbology to visually identify and flag errors in the data. Through hands-on exercises, you will reassess the dataset for any lingering issues and finalize it by updating metadata records and exporting the final dataset.



Data rules covered in this task

- **Consistency**: Ensure consistency in data formatting and values across all records to avoid discrepancies and improve data reliability
- **Accuracy**: Validate the accuracy of spatial and attribute data by cross-referencing with reliable sources or using verification tools to minimize errors
- **Completeness**: Verify the completeness of data fields, ensuring all required information is present to support comprehensive analysis and decision-making
- **Relevance**: Assess the relevance of data attributes to the analysis objectives, prioritizing essential information and eliminating unnecessary fields in order to streamline workflows
- **Documentation**: Comprehensively document metadata–including data sources, preprocessing steps, and any corrections made—in order to maintain transparency and facilitate reproducibility in subsequent analyses

GIS Software

This task requires installing the open-source software QGIS on your computer. The instructions provided here are compatible with any version of QGIS, but they are aligned with version 3.34. Minor differences in menu layouts and command names may be encountered, especially in older versions. Feel free to install multiple versions of QGIS for flexibility.

Part 1: Load and visualize data

We have already corrected and finalized the CSV file, and now we will import this cleaned dataset into QGIS to plot the points on the map canvas and conduct a visual review to identify any remaining anomalies.

1.1 Load the CSV and plot points

We will load the CSV file generated in the previous task and plot the points on the map canvas in QGIS. This will allow us to visualize the spatial distribution of the data and proceed with further analysis and corrections. Let's begin by importing the CSV file into QGIS.

- Open QGIS Desktop 3.34
- Go to **Project** > **Open** and search for the QGZ project *GRID3DataMng.qgz* in the downloaded course folder:
  ...\GRID3DataMng\GRID3DataMng.qgz

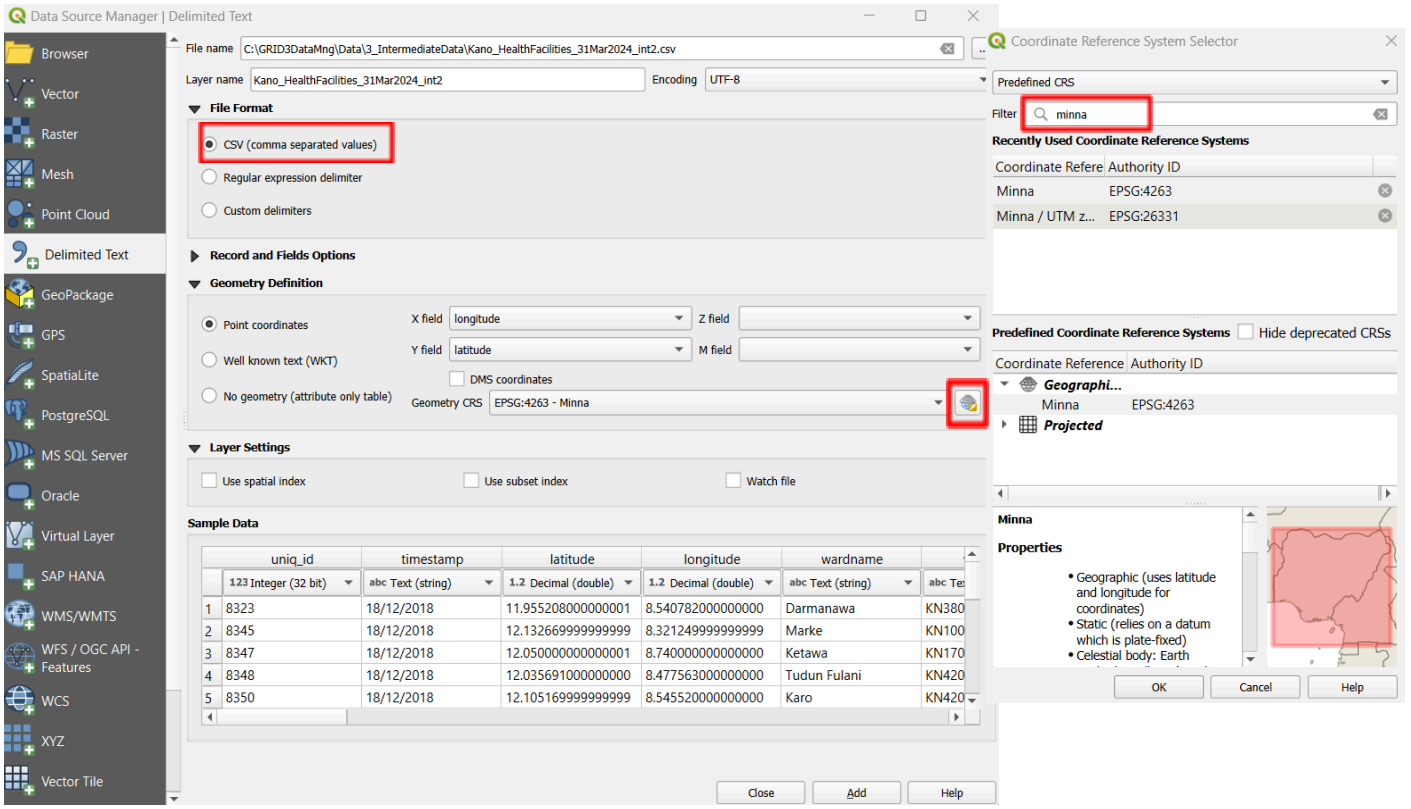
Note: You will find one sample dataset in the QGZ project: a shapefile representing the local government areas (LGAs) within the study area. Additionally, an OpenStreetMap basemap is preloaded, serving as a base layer for visual reference.

- Select **Layer** > **Add Layer** > **Add Delimited Text Layer**
- By clicking the three dots  , you can navigate to your latest CSV file created at the end of task 3 and import it:

...\GRID3DataMng\Data\3_IntermediateData\Kano_HealthFacilities_31Mar2024_int2.csv

Note: The date (31Mar2024) provided in the instructions is simply representative. Feel free to replace it with the date on which you are carrying out the exercise, if desired.

- Choose the CSV file and specify parameters. Make sure the X field is set to longitude and the Y field is set to latitude.
- Select the Nigerian Coordinate Reference System (CRS) code Minna by clicking the button  and searching for it in the filter

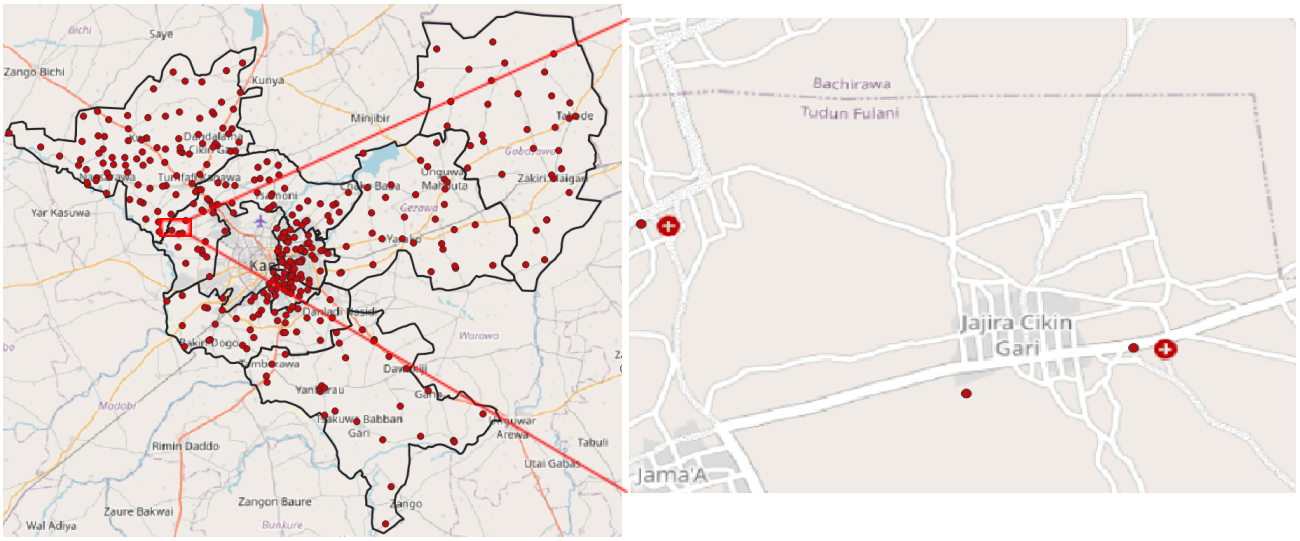


- Click **Add** and then close the box

Note: If you receive a dialog box from QGIS prompting you to select a transformation method, please choose the Nigeria onshore and offshore area of use.

1.2 Ensure correct CRS and initial data inspection

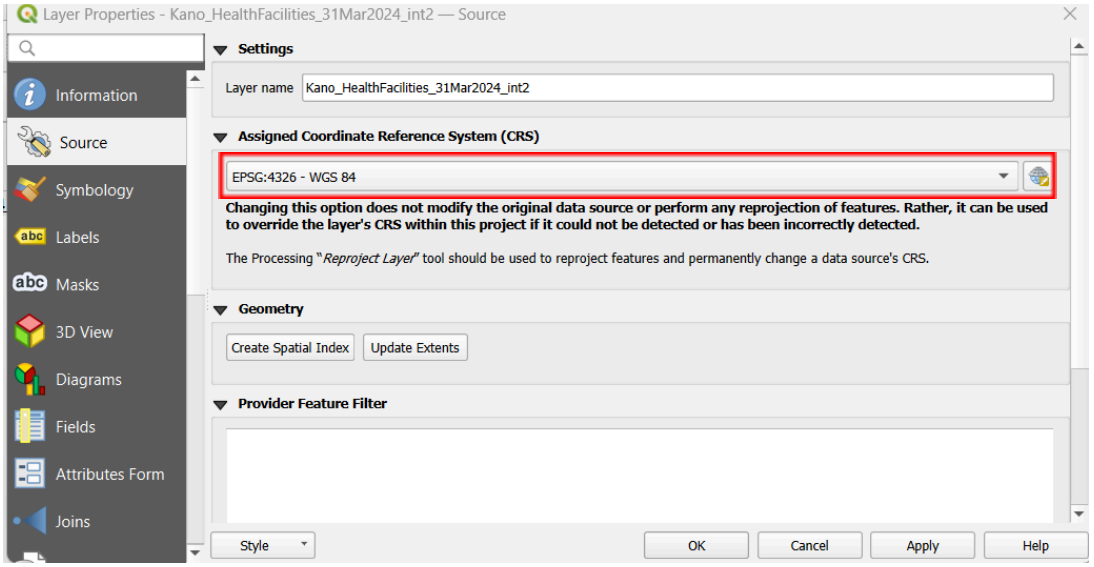
In this step, we will zoom in on specific health facilities to double-check their positions against the OpenStreetMap basemap. For example, let's zoom in on the Ungogo LGA and review the positions of the health facilities.



Let's investigate further. It seems there is a discrepancy between the basemap and our data, which could be due to the CRS we initially chose (Minna). Since these are GPS data, the primary or main CRS used is typically the WGS84 (EPSG:4326).

Next, we will modify the CRS to WGS84 (EPSG:4326) and observe if it resolves the shift issue.

- Right-click on the layer *Kano_HealthFacilities_31Mar2024_int2* in the Layers panel
- Select **Properties** > **Source**
- Adjust the **Assigned Coordinate Reference System (CRS)** to **WGS84**



It appears that modifying the CRS to WGS84 (EPSG:4326) has resolved the issue. As you can see in the screenshot below, the health facilities now align perfectly with the basemap, confirming that the correct CRS is WGS84.

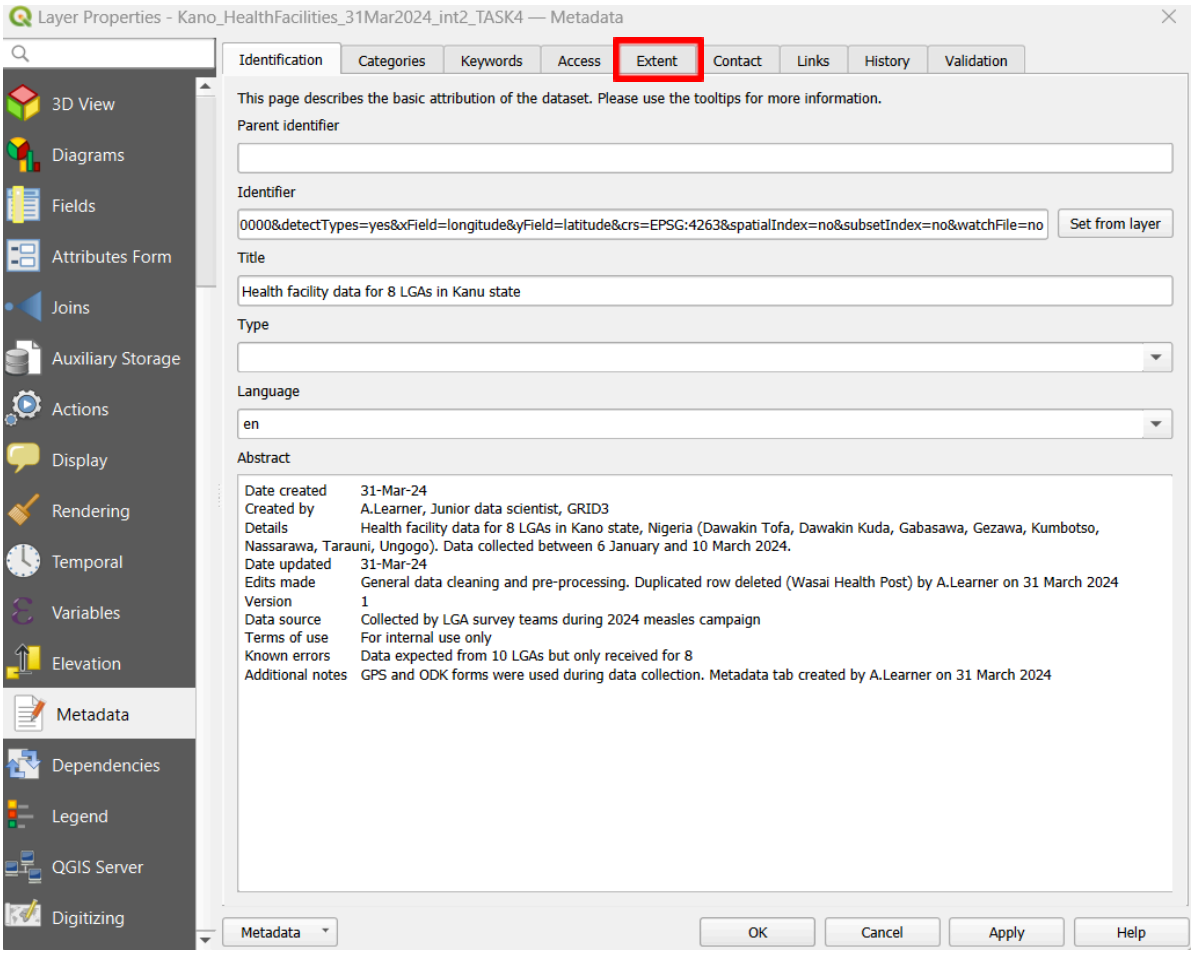
Part 2: Metadata and detailed data inspection

We will begin by updating the QGIS metadata record. It is crucial to maintain accurate metadata throughout our workflow to ensure transparency and reproducibility. You will notice that we already have a metadata record from the previous task, which we will use as a foundation for updating the QGIS metadata. Let's proceed with updating the metadata to reflect the latest changes and additions to our dataset.

2.1 Start QGIS metadata record

In QGIS, metadata can be managed directly within the software by accessing the layer properties.

- Right-click on the layer in the Layers panel
- Select **Properties** > **Metadata**.
- Here you can input key information about the dataset, including data source, acquisition date, preprocessing details, and any other relevant metadata. You can start by simply copying and pasting what you initially wrote on the Metadata spreadsheet in earlier tasks.

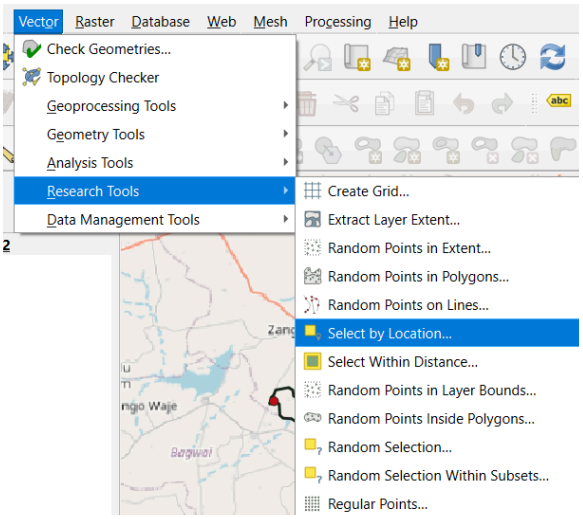


- Make sure to add the CRS from the Extent tab. You can have it automatically added by clicking **Set CRS from Layer**.

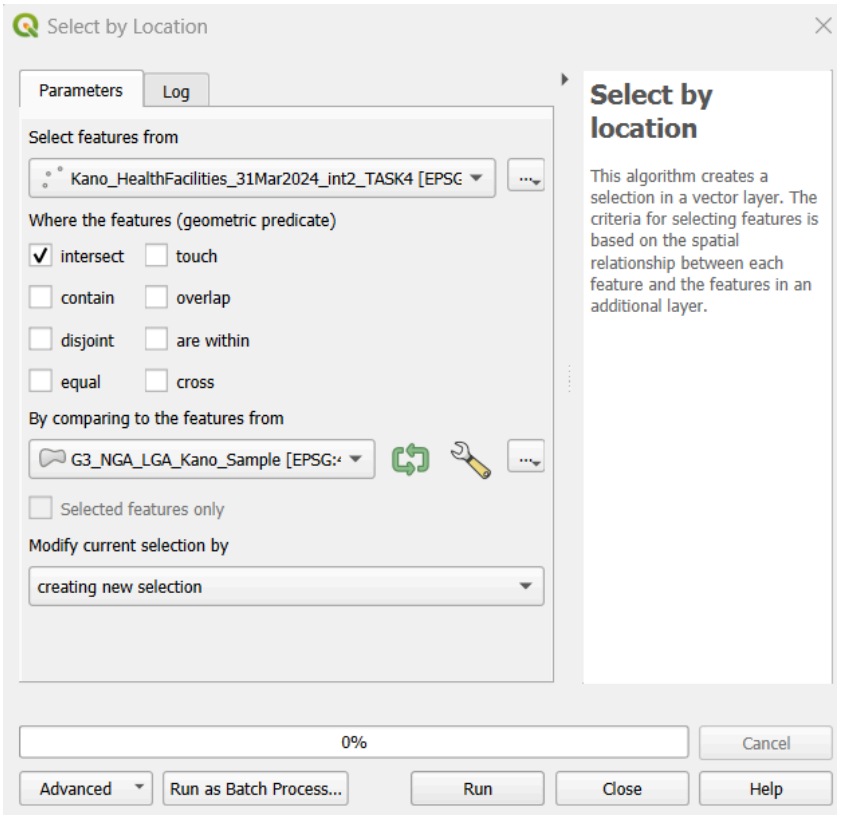
2.2 Identifying health facilities outside LGAs via spatial analysis

In this step, we will use a **Select by Location** operation to compare health facilities against LGA boundaries, identifying those facilities that are outside their designated LGAs. This visual assessment helps pinpoint discrepancies that need further investigation. Always critically assess the data, bearing in mind that LGA boundaries are operational and may not perfectly align with official boundaries.

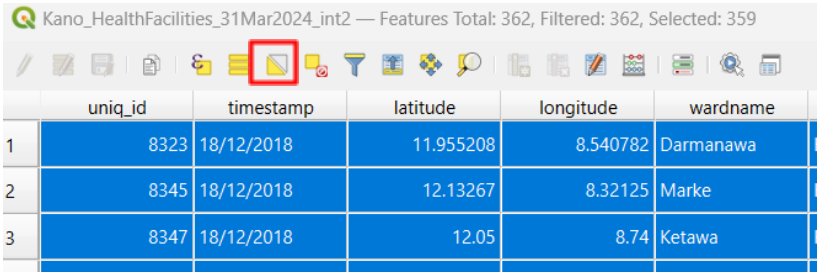
- Go to the **Vector** menu > **Research Tools** > **Select by Location**



- We will intersect our recently imported point layers with the polygon boundaries of the LGAs. Select the *Kano_HealthFacilities_31Mar2024_int2* layer in the **Select features from** option. Ensure the **Intersect** option is selected. For **By comparing to the features from**, choose the *G3_NGA_LGA_Kano_Sample* layer. Click **Run**.

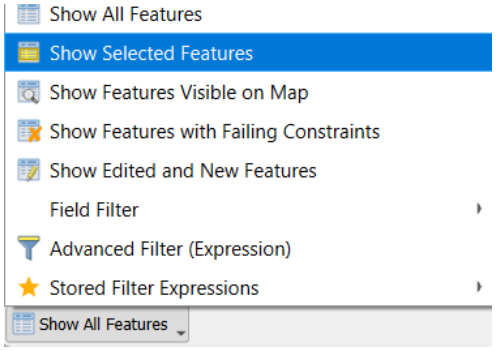


- You will now find in the attribute table the exact number of the selected points–and therefore the exact number of the points that fall inside the LGA boundaries. If there are any that fall outside the boundaries, they will be left unselected. Right-click on our layer and open the attribute table.

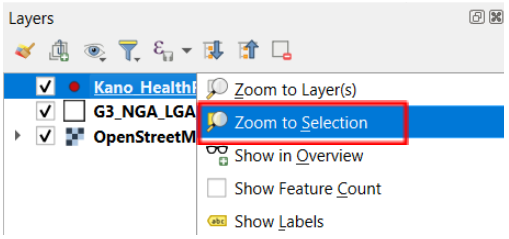


	uniq_id	timestamp	latitude	longitude	wardname
1	8323	18/12/2018	11.955208	8.540782	Darmanawa
2	8345	18/12/2018	12.13267	8.32125	Marke
3	8347	18/12/2018	12.05	8.74	Ketawa

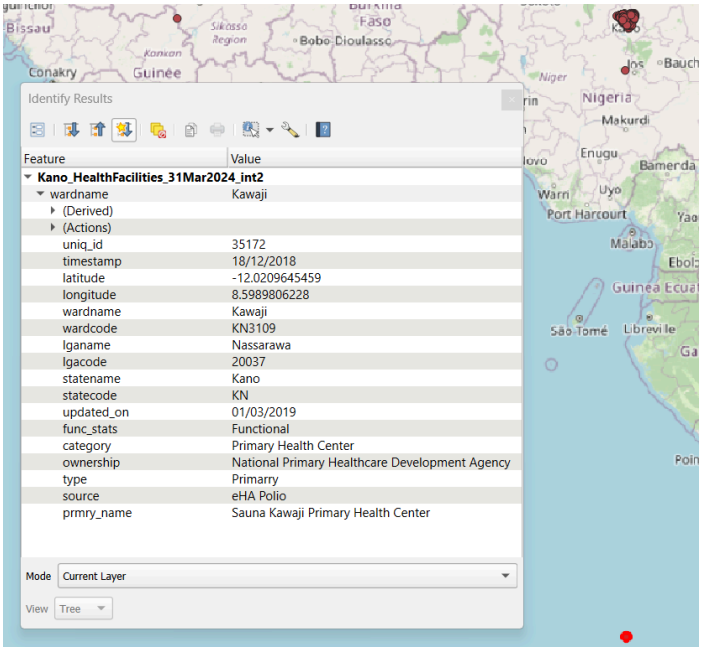
- To easily spot the points that fall outside the LGA boundaries, **Invert the selection** and show only the selected features. Click on the filter located in the lower left corner of the table labeled **Show All Features**, then choose the option **Show Selected Features**.



- Right-click on our layer and select **Zoom to Selection** to visualize the extent of the points selected in the layer

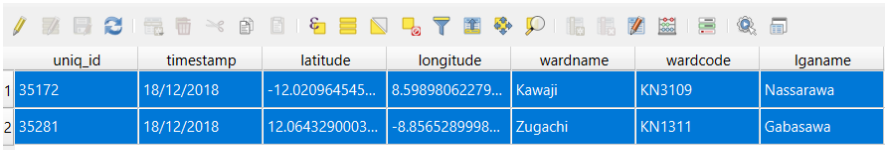


- You can use the **Identify feature** tool to investigate and understand what might have gone wrong with them



Upon initial inspection, two points appear to be located outside the country boundaries. On closer review, we observe that the points have negative latitude or longitude values, which could be causing the error. Let's correct these values and observe the outcome.

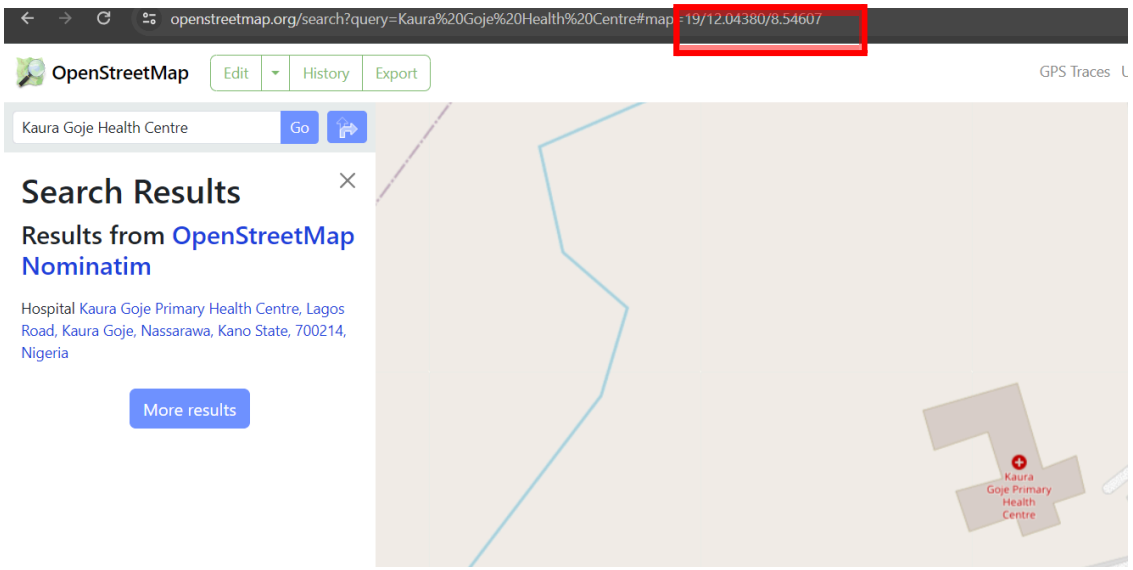
- Return to the attribute table and locate the points with these two errors by using their unique IDs. This allows you to easily find them in your CSV file. Remove the "-" sign from the respective latitude and longitude columns, then save the changes. These modifications will also be applied to the imported file.

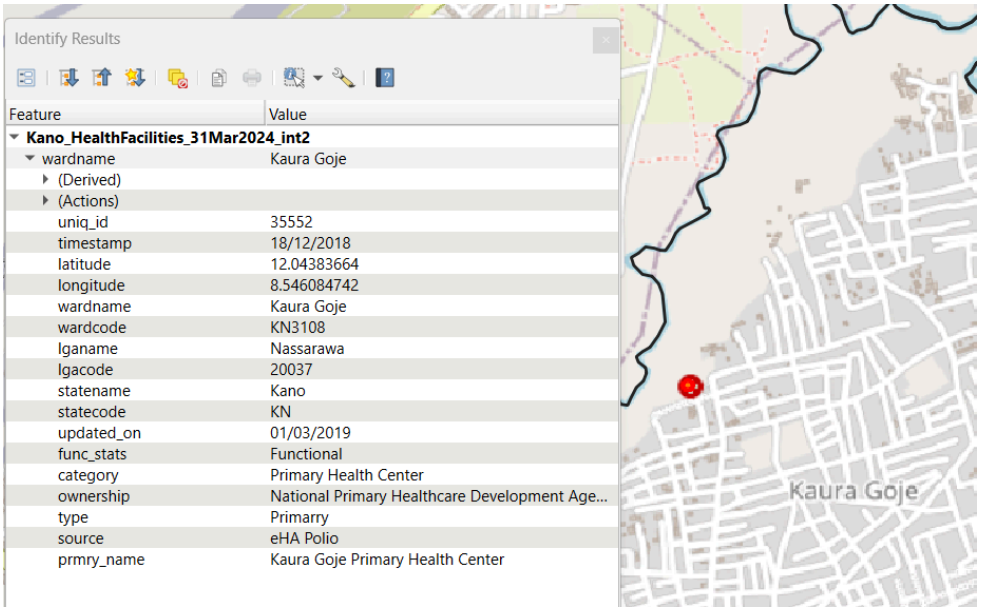


	uniq_id	timestamp	latitude	longitude	wardname	wardcode	lganame
1	35172	18/12/2018	-12.020964545...	8.59898062279...	Kawaji	KN3109	Nassarawa
2	35281	18/12/2018	12.0643290003...	-8.8565289998...	Zugachi	KN1311	Gabasawa

- Zoom in to inspect the third health facility located outside our LGAs, then use the Identify tool to examine it. Instead of speculating on the error's source, let's verify its location by searching for the facility on OpenStreetMap (OSM).
- Upon searching for Kaura Goje Health Centre, OSM provides its precise coordinates (12.04380/8.54607; see next screenshot on where to easily read the coordinates). Based on this, we can confidently adjust the latitude from 10 to 12 in our dataset.

Note: While OSM provides precise coordinates for the Kaura Goje Health Centre, learners may also consider double-checking the location on other map providers (such as Google Maps or Bing Maps) for additional confirmation.





It's a match! It fits in the same LGA and even exactly on top of the health facility mapped by OSM.

Note: This analysis may reveal health facilities that are significantly distant from their designated LGAs as specified in the database, as well as facilities that appear just outside LGA boundaries. It is important to keep in mind that LGA boundaries may have inherent inaccuracies. Therefore, facilities located slightly outside these boundaries do not necessarily indicate incorrect location data. These minor deviations might be due to anomalies within the LGA boundaries themselves, rather than errors in the facility locations.

2.3 Creation of an anomaly tracking sheet

In data management processes, it's crucial to maintain thorough documentation of any issues encountered, including errors, discrepancies, and anomalies. The creation of a bug sheet documentation provides a structured approach to recording and tracking such issues, facilitating effective communication and resolution within the data management team.

Begin by setting up a dedicated document, such as a spreadsheet, specifically for recording data-related issues. This document will serve as the bug sheet for documenting and tracking identified problems.

To initiate the creation of the Anomaly Tracking Sheet, set up the following columns in your preferred spreadsheet software and fill them out accordingly:

- **Unique ID:** Record the unique identifier of the point or feature where the issue was identified. Use the Identify tool in QGIS or any other method to find the unique identifier of the affected feature.
- **Date of Identification:** Enter the date when the issue was first identified
- **Issue Found:** Describe the anomaly or discrepancy that was detected
- **Resolved?:** Indicate whether the issue has been resolved (Yes/No)
- **Comments:** Provide any additional comments or notes related to the identified issue, including details about resolution steps or further investigations needed

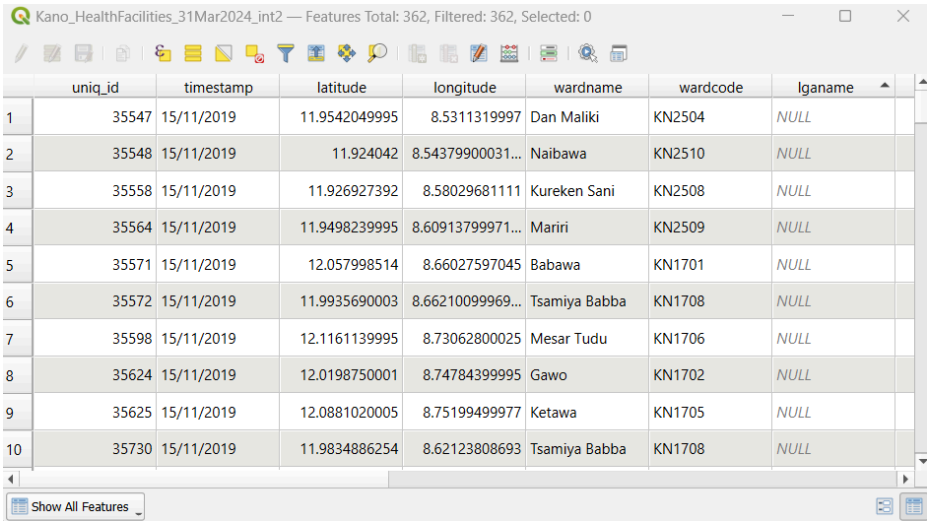
Here's an example of an anomaly tracking sheet documenting the issues we've identified and resolved so far; you can name it *bugsheet.xlsx* and save it in the *5_Outputs* folder. It's crucial to keep updating this sheet as we progress through the exercise to ensure comprehensive documentation.

Unique ID	Date of identification	Issue found	Resolved?	Comments
35172	31/03/2024	Wrong Latitude record	Yes	Found a "-" sign before the latitude that shouldn't be there, deleted it and it fixed the issue
35281	31/03/2024	Wrong Longitude record	Yes	Found a "-" sign before the longitude that shouldn't be there, deleted it and it fixed the issue
35552	31/03/2024	Wrong Latitude record	Yes	Through a research in OSM, the latitude was corrected from 10 to 12

2.4 Close inspection of the attribute table

Next, we will thoroughly review the attribute table in QGIS, which contains information on each feature in our dataset. While some of this information is crucial–such as facility type and location–other data may be unnecessary. You can explore, filter, and sort these fields to pinpoint and rectify inconsistencies or errors. This meticulous inspection is key to enhancing the accuracy and completeness of our data prior to further analysis.

- Right-click again on our layer and click **Open Attribute Table**
- Quickly inspect the table by clicking on the title of each column. You can use each column's values to sort the table from low to high by clicking once, and in reverse order by clicking again. This helps to identify irregularities or anomalies.



	uniq_id	timestamp	latitude	longitude	wardname	wardcode	lganame
1	35547	15/11/2019	11.9542049995	8.5311319997	Dan Maliki	KN2504	NULL
2	35548	15/11/2019	11.924042	8.54379900031...	Naibawa	KN2510	NULL
3	35558	15/11/2019	11.926927392	8.58029681111	Kureken Sani	KN2508	NULL
4	35564	15/11/2019	11.9498239995	8.60913799971...	Mariri	KN2509	NULL
5	35571	15/11/2019	12.057998514	8.66027597045	Babawa	KN1701	NULL
6	35572	15/11/2019	11.9935690003	8.66210099969...	Tsamiya Babba	KN1708	NULL
7	35598	15/11/2019	12.1161139995	8.73062800025	Mesar Tudu	KN1706	NULL
8	35624	15/11/2019	12.0198750001	8.74784399995	Gawo	KN1702	NULL
9	35625	15/11/2019	12.0881020005	8.75199499977	Ketawa	KN1705	NULL
10	35730	15/11/2019	11.9834886254	8.62123808693	Tsamiya Babba	KN1708	NULL

- For example, you may notice some missing LGA names, indicated by "*NULL*" entries. In the next section, we will explore tools that can help us to efficiently rectify this issue.

Part 3: Utilizing GIS processing

Let's dive into using GIS processing tools to enhance our dataset.

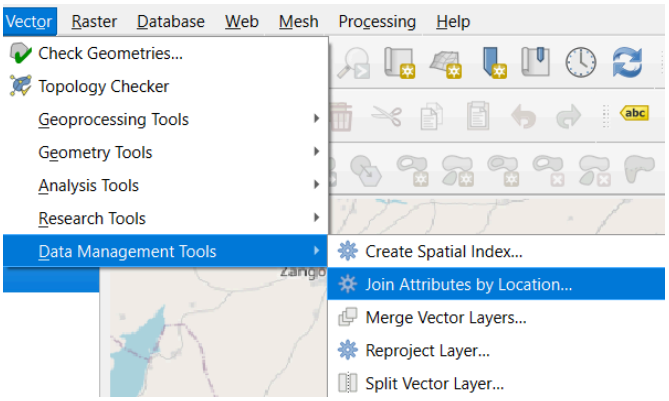
3.1 Attribute table updates and data corrections

Let's now address some remaining data inconsistencies and errors, beginning with missing LGA names.

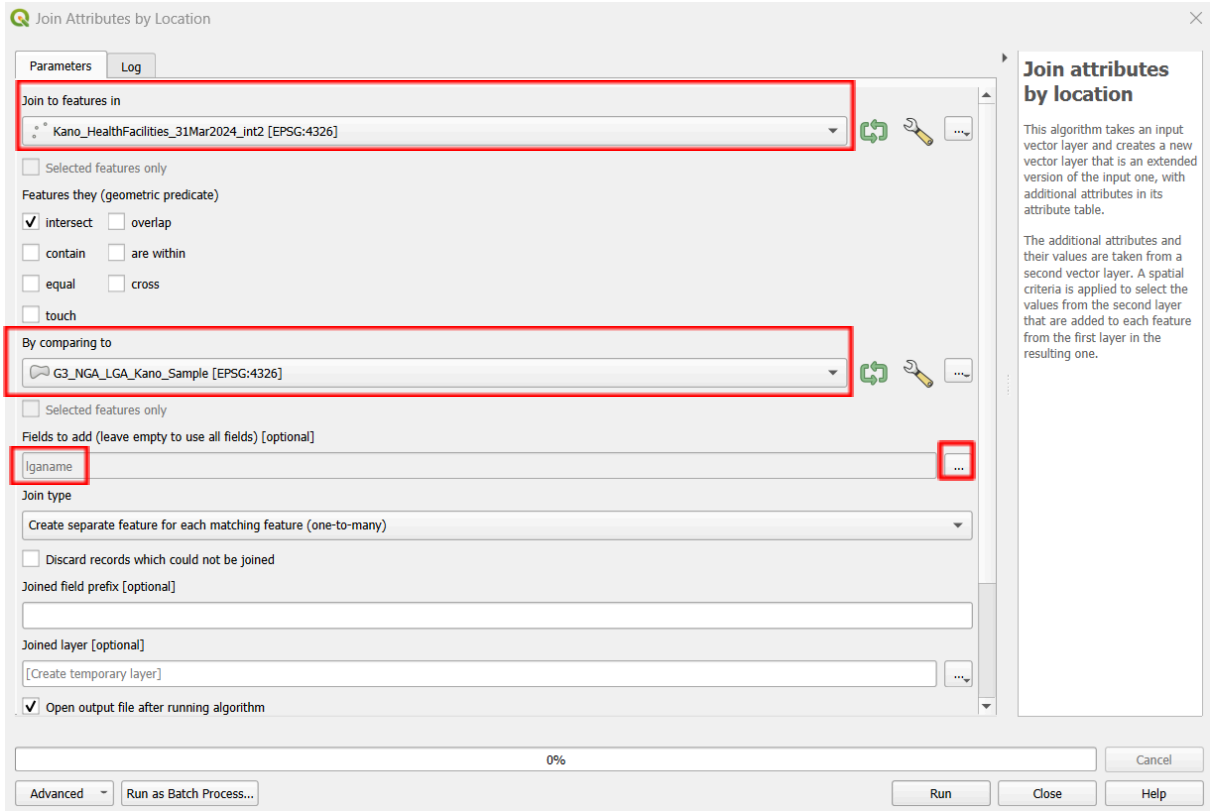
One option to handle this is through the **Join Attributes by Location** tool in QGIS. This tool compares features, such as those in the Kano Health Facilities layer, with those in the LGA layer (*G3_NGA_LGA_Kano_Sample*), assigning attributes from the latter to the former based on spatial proximity.

Note: The tools described in Section 3 and Section 4 may serve similar functions to those commonly used in spreadsheet software like Excel. However, these tools are applied within a GIS context, offering an alternative method to accomplishing similar tasks using spatial data.

- In the vector menu, inside the **Data Management Tools**, click on **Join Attributes by Location**

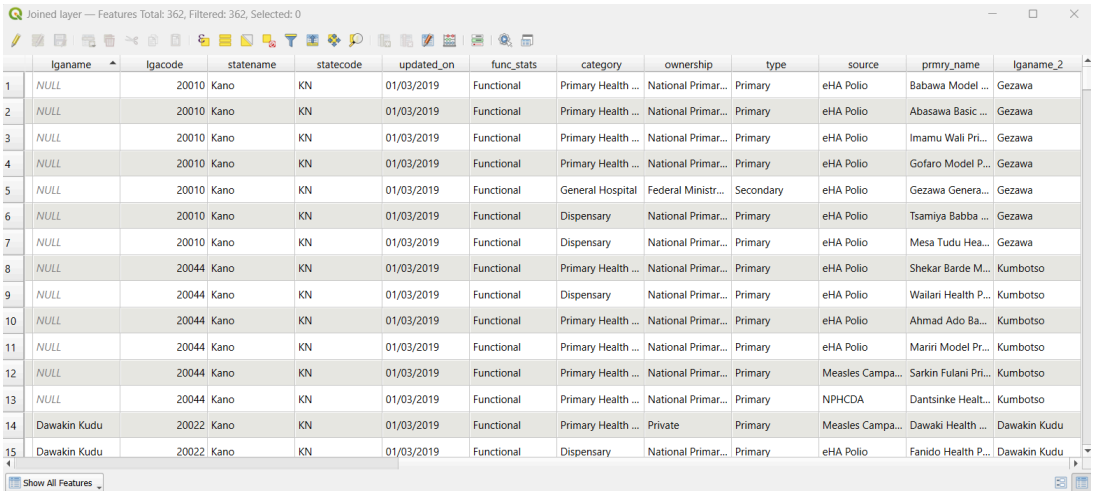


- Select the Kano Health Facilities layer as the Join layer and ensure that the option **Intersect** is selected. Then, in the **By comparing to** field, choose the *G3_NGA_LGA_Kano_Sample* layer.
- In the **Fields to add** section, select only the lganame column, as it appears to be the only missing field in our table. And finally, click **Run**.



- We have created a temporary layer that we can review by inspecting the attribute table once again. You should expect to find a new field containing the LGA names of all health facilities as a result of this technique.

Note: Temporary layers can be lost if your laptop freezes or crashes. Regularly save the QGIS QGZ project and your files.



	lganame	lgacode	statename	statecode	updated_on	func_stats	category	ownership	type	source	prmry_name	lganame_2
1	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Babawa Model ...	Gezawa
2	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Abasawa Basic ...	Gezawa
3	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Imamu Wali Pri...	Gezawa
4	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Gofaro Model P...	Gezawa
5	NULL	20010	Kano	KN	01/03/2019	Functional	General Hospital	Federal Ministr...	Secondary	eHA Polio	Gezawa Genera...	Gezawa
6	NULL	20010	Kano	KN	01/03/2019	Functional	Dispensary	National Primar...	Primary	eHA Polio	Tsamiya Babba ...	Gezawa
7	NULL	20010	Kano	KN	01/03/2019	Functional	Dispensary	National Primar...	Primary	eHA Polio	Mesa Tudu Hea...	Gezawa
8	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Shekar Barde M...	Kumbotso
9	NULL	20044	Kano	KN	01/03/2019	Functional	Dispensary	National Primar...	Primary	eHA Polio	Wailari Health P...	Kumbotso
10	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Ahmad Ado Ba...	Kumbotso
11	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	eHA Polio	Mariri Model Pr...	Kumbotso
12	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	Measles Campa...	Sarkin Fulani Pri...	Kumbotso
13	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health ...	National Primar...	Primary	NPHCDA	Dantsinke Healt...	Kumbotso
14	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Primary Health ...	Private	Primary	Measles Campa...	Dawaki Health ...	Dawakin Kudu
15	Dawakin Kudu	20022	Kano	KN	01/03/2019	Functional	Dispensary	National Primar...	Primary	eHA Polio	Fanido Health P...	Dawakin Kudu

3.2 Comparing multiple LGA name fields

After running processing tools to update the attribute table, you may find yourself with two LGA name fields. To ensure data consistency, you can compare these fields to identify any discrepancies.

- **Open the table:** Right-click on the layer in the Layers panel and select **Open Attribute Table**

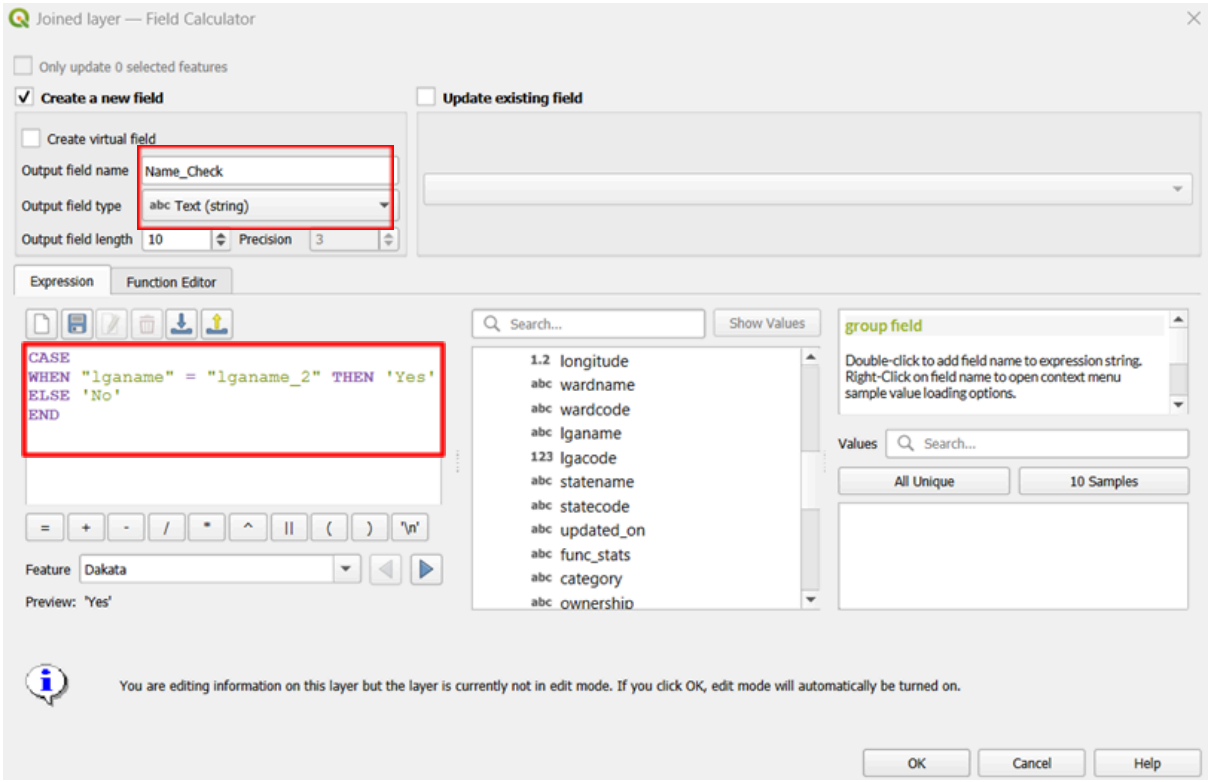
Note: Make sure all attributes are deselected by clicking the icon "Deselect features from all layers"

- **Create a new field for comparison:** Go to the Field Calculator by clicking on the icon  or selecting **Field Calculator** from the menu. Create a new field called Name_Check or a similar name and as a type choose **Text (string)**.
- **Comparison expression:** In the Field Calculator expression builder, enter the following SQL expression:

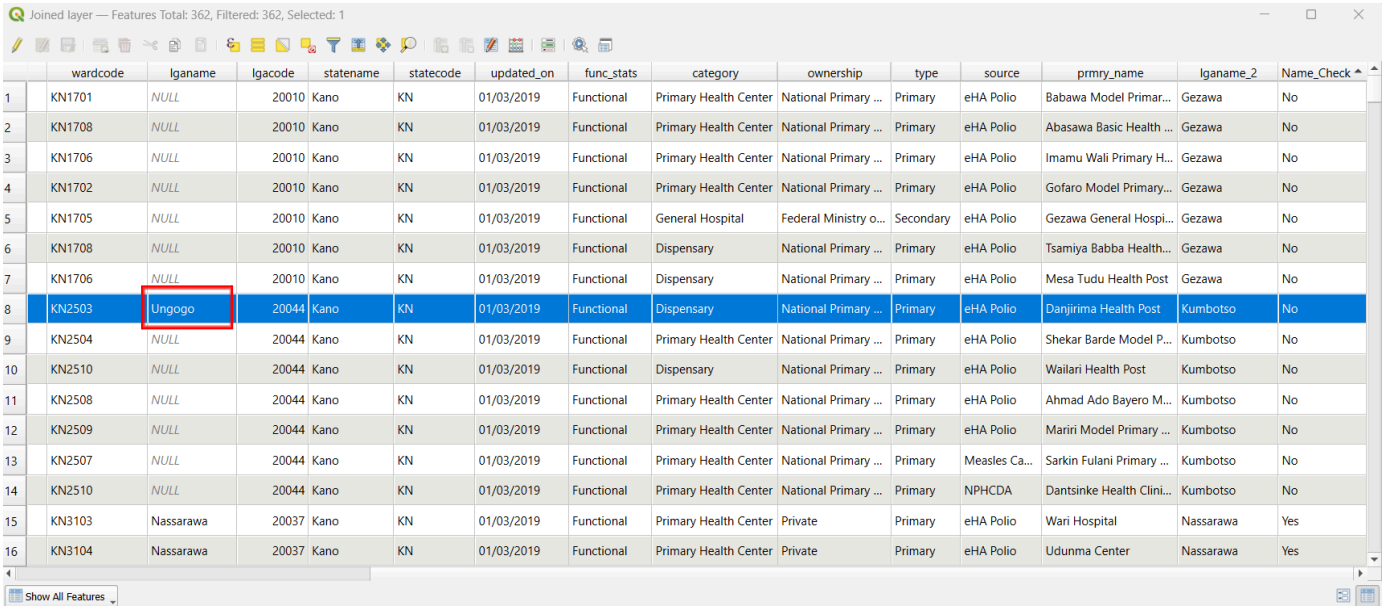
```
CASE
WHEN "lganame" = "lganame_2" THEN 'Yes'
ELSE 'No'
END
```

Note: This expression compares the values in the two LGA name fields. If they match, it returns "Yes"; otherwise, it returns "No."

- **Apply the expression:** Click **OK** to apply the expression. A new field will be added to the attribute table indicating whether the LGA name fields match each feature.



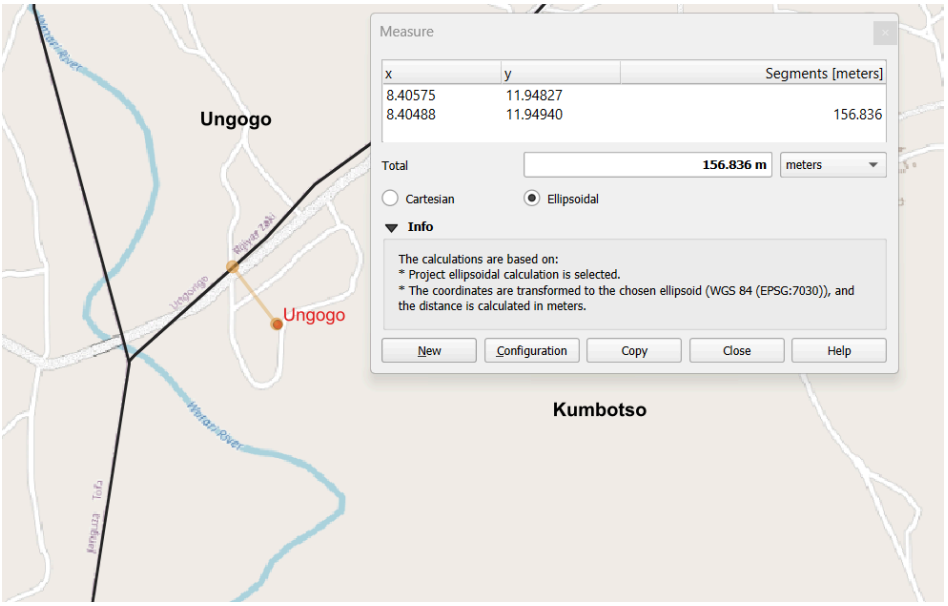
- **Review the results:** Inspect the Name_Check field to see the comparison results. Features with "Yes" indicate matching LGA name values, while "No" indicates discrepancies. Sort the table on this field to easily identify and address any inconsistencies.



	wardcode	lganame	lgacode	statename	statecode	updated_on	func_stats	category	ownership	type	source	prmry_name	lganame_2	Name_Check
1	KN1701	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Babawa Model Primar...	Gezawa	No
2	KN1708	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Abasawa Basic Health ...	Gezawa	No
3	KN1706	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Imamu Wali Primary H...	Gezawa	No
4	KN1702	NULL	20010	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Gofaro Model Primary...	Gezawa	No
5	KN1705	NULL	20010	Kano	KN	01/03/2019	Functional	General Hospital	Federal Ministry o...	Secondary	eHA Polio	Gezawa General Hospi...	Gezawa	No
6	KN1708	NULL	20010	Kano	KN	01/03/2019	Functional	Dispensary	National Primary ...	Primary	eHA Polio	Tsamiya Babba Health...	Gezawa	No
7	KN1706	NULL	20010	Kano	KN	01/03/2019	Functional	Dispensary	National Primary ...	Primary	eHA Polio	Mesa Tudu Health Post	Gezawa	No
8	KN2503	Ungogo	20044	Kano	KN	01/03/2019	Functional	Dispensary	National Primary ...	Primary	eHA Polio	Danjirima Health Post	Kumbotso	No
9	KN2504	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Shekar Barde Model P...	Kumbotso	No
10	KN2510	NULL	20044	Kano	KN	01/03/2019	Functional	Dispensary	National Primary ...	Primary	eHA Polio	Wailari Health Post	Kumbotso	No
11	KN2508	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Ahmad Ado Bayero M...	Kumbotso	No
12	KN2509	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	eHA Polio	Mariri Model Primary ...	Kumbotso	No
13	KN2507	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	Measles Ca...	Sarkin Fulani Primary ...	Kumbotso	No
14	KN2510	NULL	20044	Kano	KN	01/03/2019	Functional	Primary Health Center	National Primary ...	Primary	NPHCDA	Dantsinke Health Clini...	Kumbotso	No
15	KN3103	Nassarawa	20037	Kano	KN	01/03/2019	Functional	Primary Health Center	Private	Primary	eHA Polio	Wari Hospital	Nassarawa	Yes
16	KN3104	Nassarawa	20037	Kano	KN	01/03/2019	Functional	Primary Health Center	Private	Primary	eHA Polio	Udunma Center	Nassarawa	Yes

After running the comparison expression, we can identify a point with an LGA name of Ungogo in our facility data that is actually located in Kumbotso LGA in our boundary data. To locate this point, click on the far left of the row in the attribute table and select **Zoom to Selection** to highlight it on the map.

Use the measure tool  to see how near the point is to the closest boundary.



We notice that the point is located close to the boundary between Ungogo and Kumbotso, around 156 meters. We have to make a judgment about how to resolve the anomaly. Characterizing the type of mismatch can help.

- If the point in the facility data is far from its designated LGA according to the boundary data, we have very high certainty that there is an error in the facility data. Either the location information

is wrong or the LGA attribute is wrong.
    - If we have independent confirmation that the facility is in the correct place, we correct the LGA value based on the boundary data
    - If such confirmation is not possible, we flag the point as requiring review and describe the issue in the bugsheet
- If the point in the facility data is close to its designated LGA according to the boundary data, we do not know if there is an error or not. The boundary file may be inaccurate or overly generalized.
    - We do not make any changes in either the facility data or the boundary data
    - We describe the discrepancy in the bugsheet

In this case, the anomaly is of the second type, so we make no corrections and describe the issue in the bugsheet.

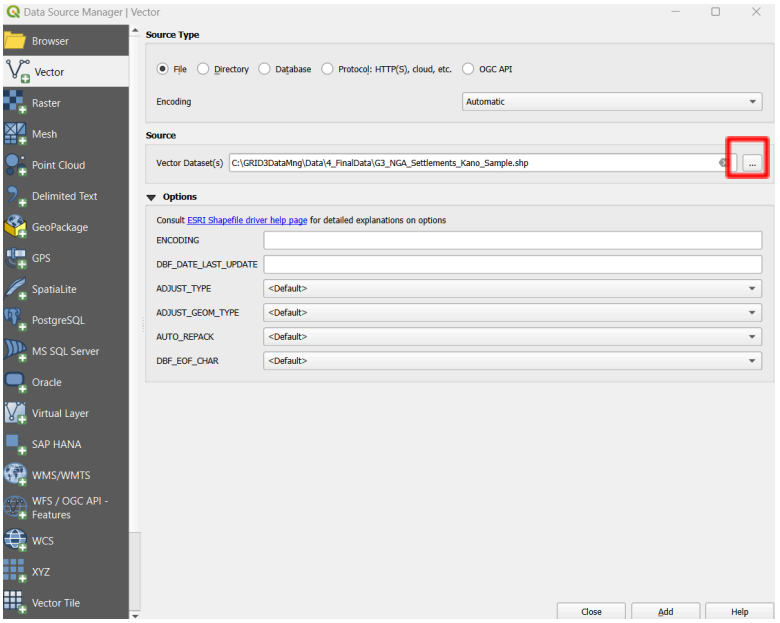
Disclaimer: *The boundary data utilized in this exercise, including LGAs, are operational and may be outdated. It is important to acknowledge that boundary data are subject to change due to various factors, and the boundaries presented here are for operational purposes only. Learners should exercise caution and verify the accuracy of boundary data as needed for their specific use cases.*

3.3 Analyzing health facility points against settlement extents

In this step, we'll assess how health facility points align with settlement extents, which delineate the built-up areas around human settlements sourced from GRID3 data. Settlement extents categorize regions into Built-up Areas (BUA extents), Small Settlement Areas (SSA extents), and Hamlets (hamlet extents), based on building density. Using QGIS's **Select Within Distance** tool, we'll analyze the spatial relationship between health facilities and settlement extents by specifying a distance threshold to identify points within the proximity of the settlement extents.

First, let's import the settlement extents layer:

- Go to **Layer** > **Add Layer** > **Add Vector Layer**. Browse and select the file located at *...\GRID3DataMng\Data\4_FinalData\G3_NGA_Settlements_Kano_Sample.shp*
- Click **Add** to import the layer and close this window

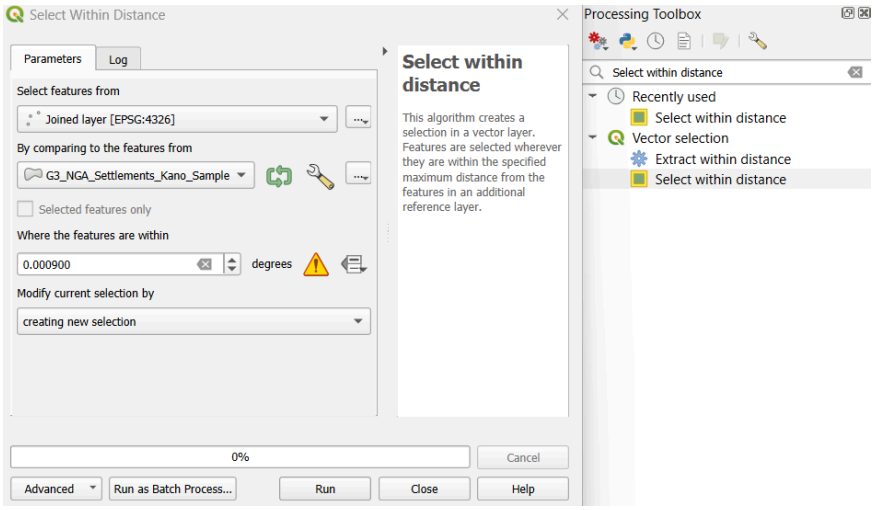


Next, run the **Select Within Distance** tool:

- Go to the **Processing Toolbox**. Search for the **Select Within Distance** tool. Double-click on the tool to open its parameters.
- Select the health facility points layer (Joined Layer, or however you have named it) as the input layer
- Choose the settlement extents layer as the **Select features from** layer
- Set the distance to 0.0009 degrees (equivalent to approximately 100 meters at this location)

Note: If you want to specify distance directly in meters instead of degrees, you need to convert the data into a projected CRS. Read more about this in course A005: Understanding Coordinate Reference Systems (CRS) using QGIS. *In the task outlined here, it is not necessary to specify meters instead of degrees.*

- Click **Run** to execute the tool



After identifying health facility points that are within settlement extents, we will further refine our analysis by examining points that fall outside settlement extents. This allows us to pinpoint potential inaccuracies that may require further investigation.

- After the initial selection is made, navigate to the selection menu in the attribute table; as you did earlier, choose the **Invert Selection** option to select those outside settlement extents
- Upon completion, you should observe a total of 19 points resulting from the selection process. We will review them in the next step based on the satellite imagery.

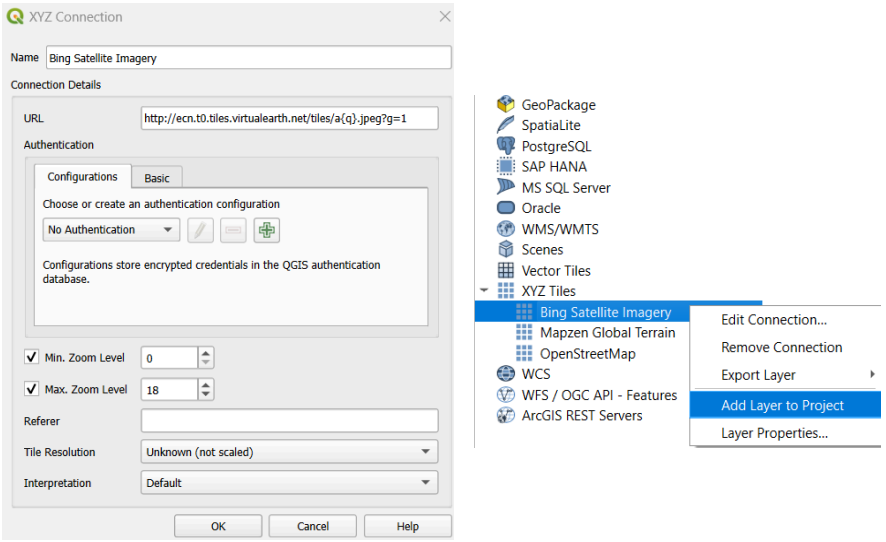
3.4 Analyzing against satellite imagery

Satellite imagery helps us make judgments about facilities located outside the settlement extent polygons.

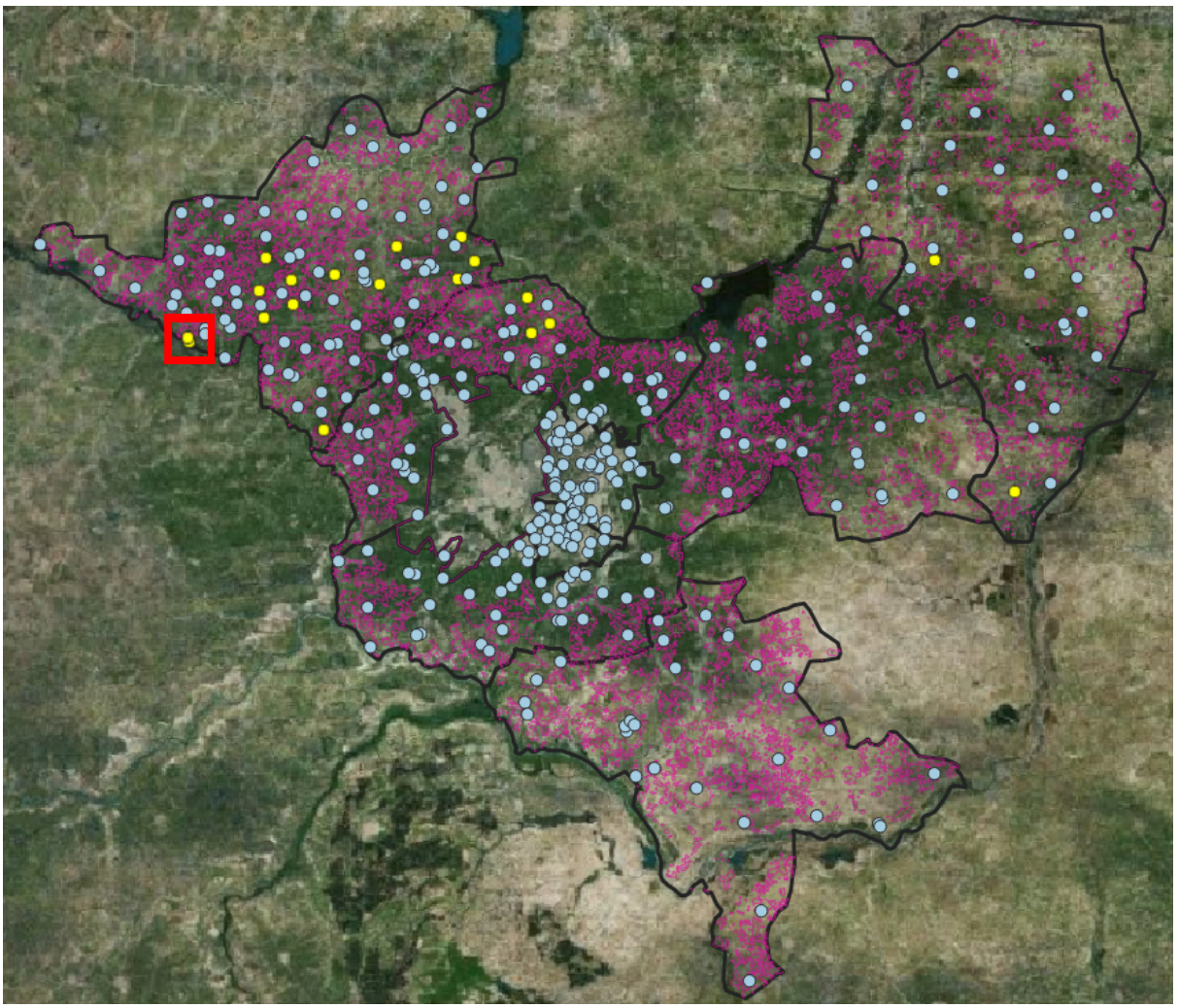
Let's load Bing Satellite Imagery:

- Go to XYZ Tiles in the browser panel
- Right-click and choose **New Connection**
- Name it "Bing Satellite Imagery"

- Enter the following URL: http://ecn.t0.tiles.virtualearth.net/tiles/a{q}.jpeg?g=1
- Click **OK** to add the layer
- Right-click the newly added layer and click **Add Layer to Project**



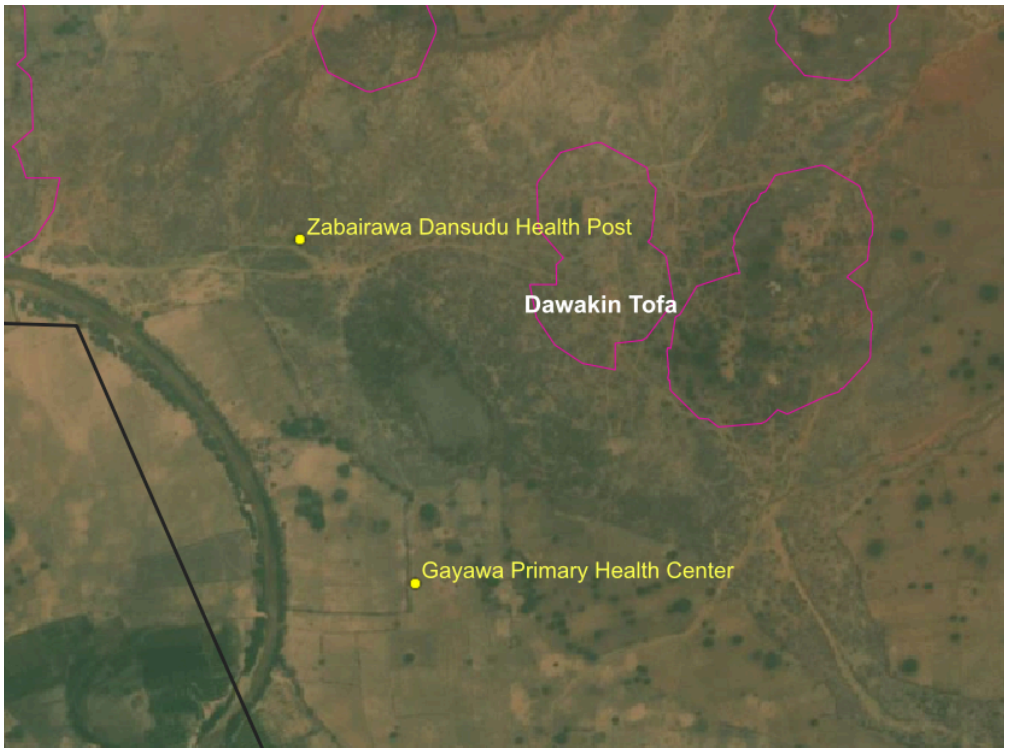
Zoom in to one of the selected health facilities. Remember, they will be highlighted in yellow.



There are several reasons why a facility point might be located outside of a settlement polygon:

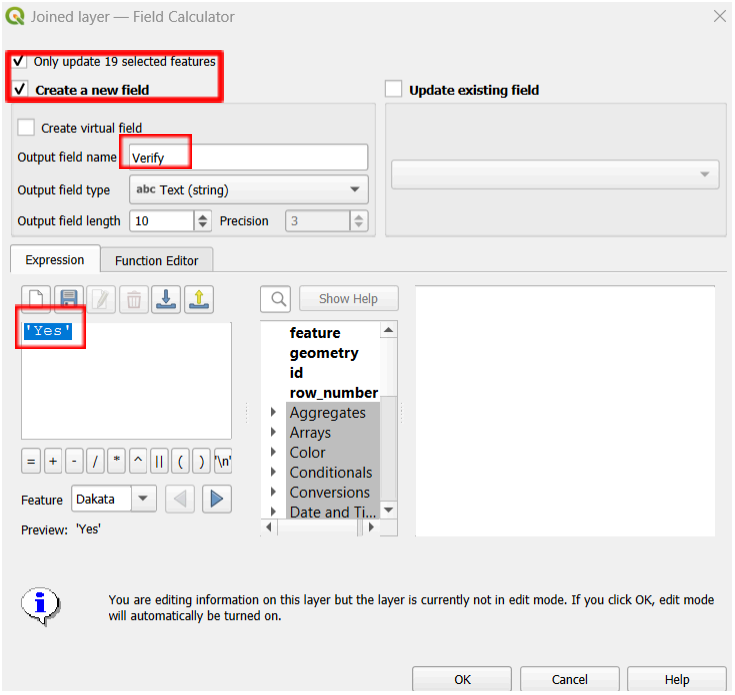
1. Data inaccuracies: This includes data entry errors (or geocoding mistakes) that lead to incorrect facility locations
2. Dynamic geographical changes: Settlement extents may not reflect recent developments due to factors like population movement or changes in settlement patterns, as they were derived from imagery captured at a specific point in time
3. Operational changes: Facilities might have relocated, leading to discrepancies in recorded data without corresponding updates in the dataset

Look for visual indicators–like nearby infrastructure, terrain characteristics, or land use patterns–to corroborate the placement of health facilities within the context of their surroundings.



Upon closer investigation, it appears that these points are dubious; the imagery fails to corroborate their validity. There are two plausible scenarios: either the point data are erroneous, or the imagery is outdated, warranting cross-verification with alternative sources or local knowledge. Regardless, these points require tagging for additional scrutiny.

To accomplish this, reopen the attribute table and access the **Field Calculator**. Create a new field (you can name it "Verify") and only update the selected features that merit further investigation with a designation of "Yes".

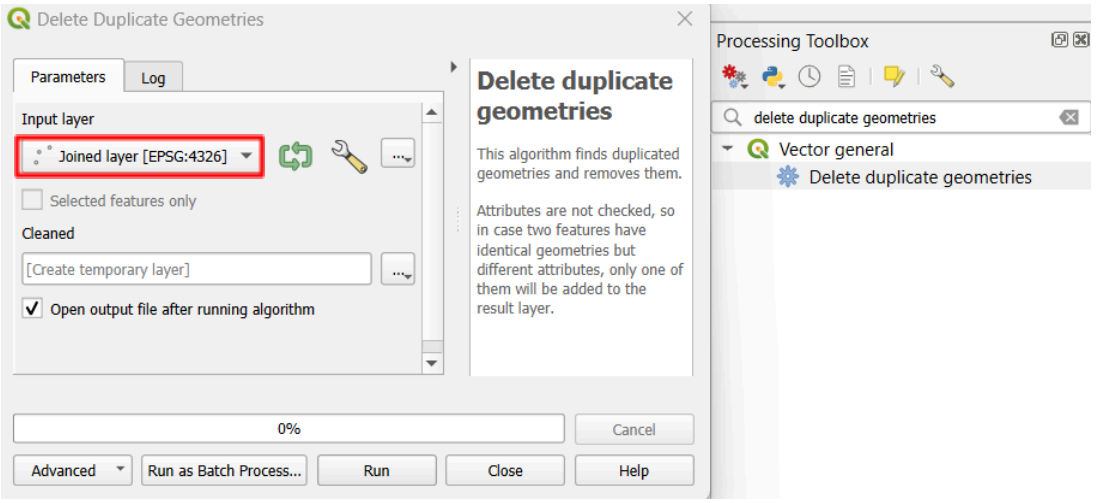


- Save the edits and close the edition mode by deselecting the **Toggle Edition** 
- Remember to note the identified features in the bug sheet as "not resolved, requiring further review", and also note the exact satellite image it was reviewed on.

3.5 Applying geoprocessing techniques

There are many geoprocessing techniques available to detect anomalies in our dataset. As an example, we will demonstrate the use of the **Delete Duplicate Geometries** tool.

- In the Processing Toolbox, search for Delete Duplicate Geometries. Select the joined layer created earlier as the Input layer and click **Run**.



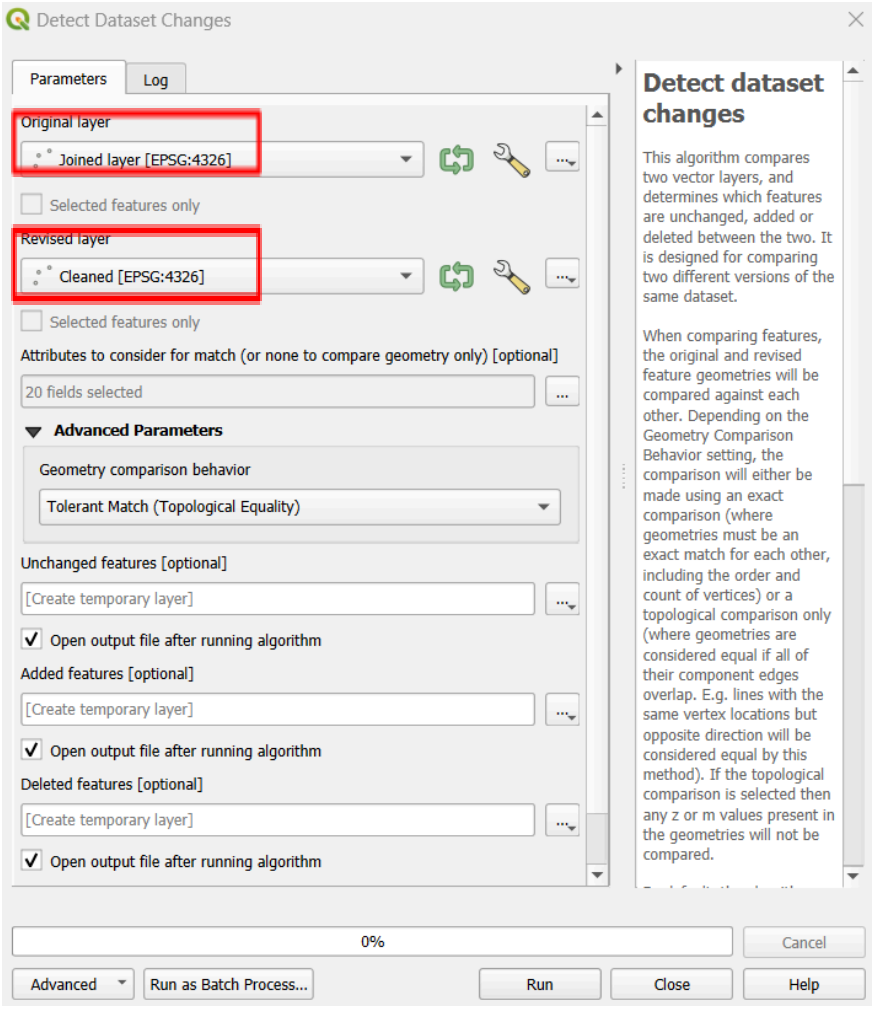
- You will receive the following message:

  {'DUPLICATE_COUNT': 2, 'OUTPUT': 'Cleaned_29279c4d_3efb_4655_ba15_029d2beb4790', 'RETAINED_COUNT': 360}

  This confirms that the tool identified and deleted 2 duplicate entries while retaining 360 health facilities.

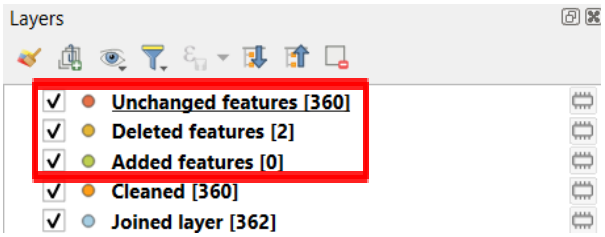
In order to identify features that were deleted during the previous operation, we'll utilize the **Detect Dataset Changes** tool. This tool helps us detect any modifications made to our dataset, including deletions, additions, or modifications to existing features. We'll use it to pinpoint any features that were removed, allowing us to track changes effectively.

- In the Processing Toolbox, search for **Detect dataset changes**. Select the Joined layer created earlier as the **Original layer**, the Cleaned layer as the **Revised layer** and click **Run**.



After running the **Detect dataset changes** tool, you'll get three layers: *Unchanged features*, *Deleted features*, and *Added features*. Check the *Deleted features* layer to spot the removed features. Verify these changes with the original attribute table to confirm whether they were indeed duplicates. Record any discrepancies in the bugsheet.

*Tip: You can right-click on any layer in QGIS and select **Show Feature Count** for a quick view of the number of features in each layer. This helps assess the extent of changes made.*



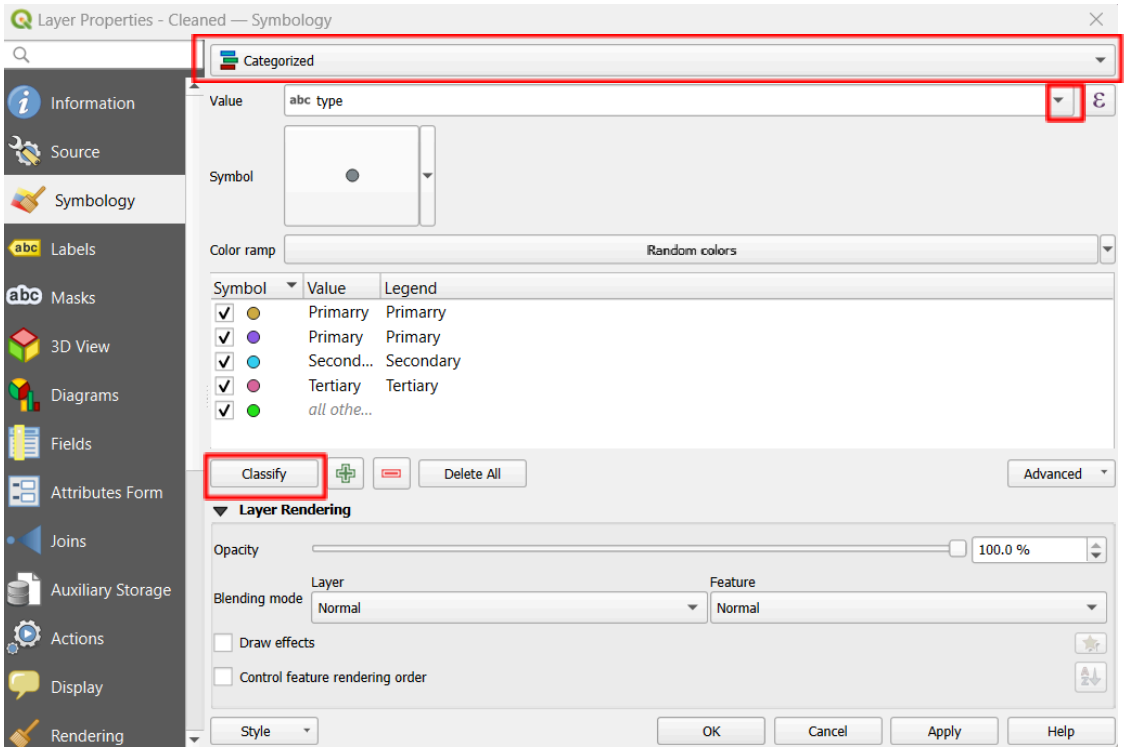
Part 4: Reassessment and finalization

Now we will review the dataset for remaining anomalies or errors. We will inspect the map canvas and attribute table to see if there are further options for improving data integrity and completeness. After a final assessment of the corrections, we will update the metadata record and export the final dataset for analysis.

4.1 Reassess map canvas and attribute table

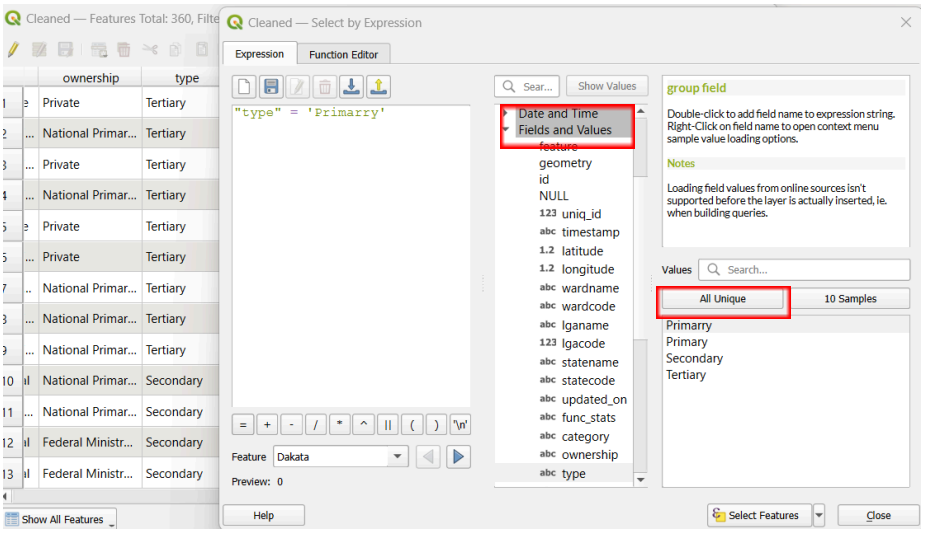
We will check the map canvas and attribute table for remaining errors. We will then symbolize the data based on health facility types in order to visually identify remaining issues. Finally, we will make adjustments to improve data integrity and completeness before finalizing the dataset.

- Right-click on the latest layer created (automatically named “Cleaned”, or whatever you have named it), click on Properties and go to Symbology
- In the type of symbology, choose **Categorized** > and as a **Value** choose **type** as the category > click on **Classify**. This will randomly assign colors to the health facilities based on their type.

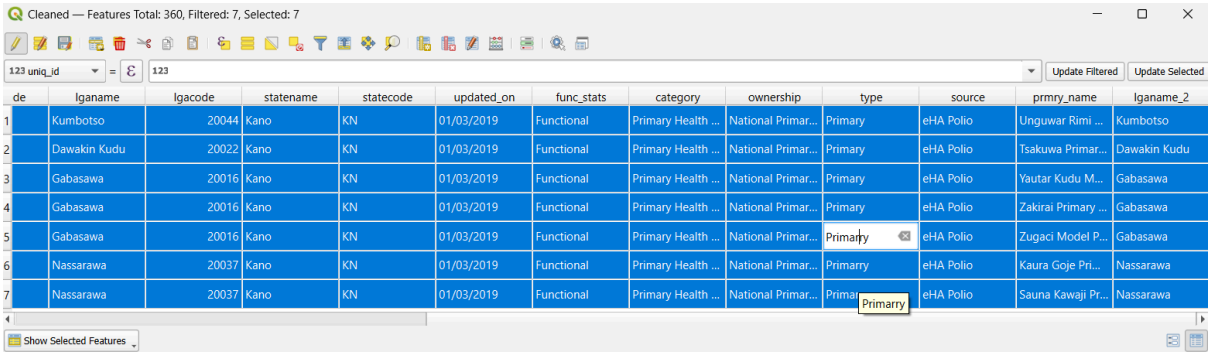


- We notice that there are four (4) facility types instead of the expected three (3), likely due to a typing error ("Primarry" instead of "Primary"). We will correct this directly in the attribute table, since we are working on a temporary file.

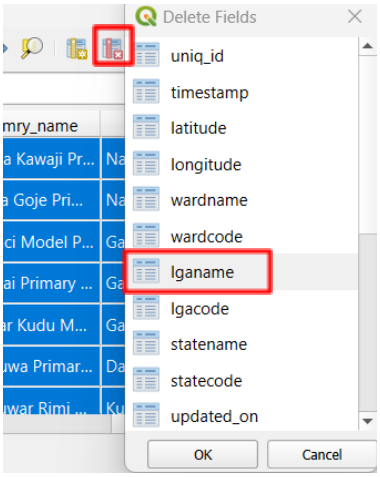
- Right-click on our last layer > **Open attribute table** > Select the **Select feature using expressions** tool; let's select only the health facilities that have the value "Primarry". (You can select the fields directly from the Fields and Values menu, and then click "All unique" to get the unique values of the types).



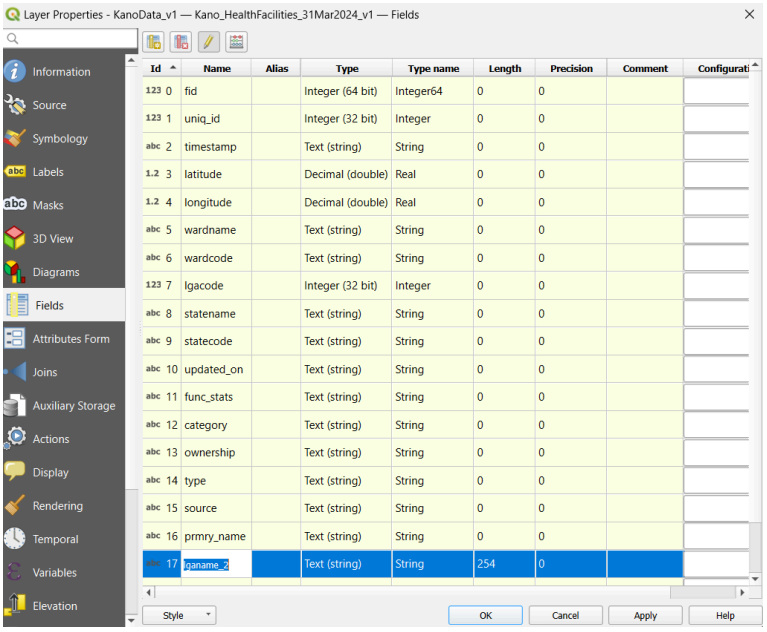
- Use the same technique in Part 2.2 to only see the selected features. Activate the edition mode by clicking on the icon to **Toggle editing** and directly modify the word "Primarry" back to "Primary" on the type field. Note the features changed in the bugsheet (while you're making the changes) as well.



- While we're here, we can also delete the old lganame field that had that error and leave the last one (i.e. the one created automatically). Click on the **Delete fields** tool , select lganame, and click **OK**.



- Let's also clean the lganame_2 we created earlier and rename it back to lganame. Right click on the layer and go to **properties**. On the **Fields** tab, find the lganame_2 column and rename it to lganame. Add this information to the bugsheet as well, where the LGA names were missing or wrong.



- **Save** the edits by clicking the save button  and click again on the toggle editing to close the edition mode.

4.2 Export final data

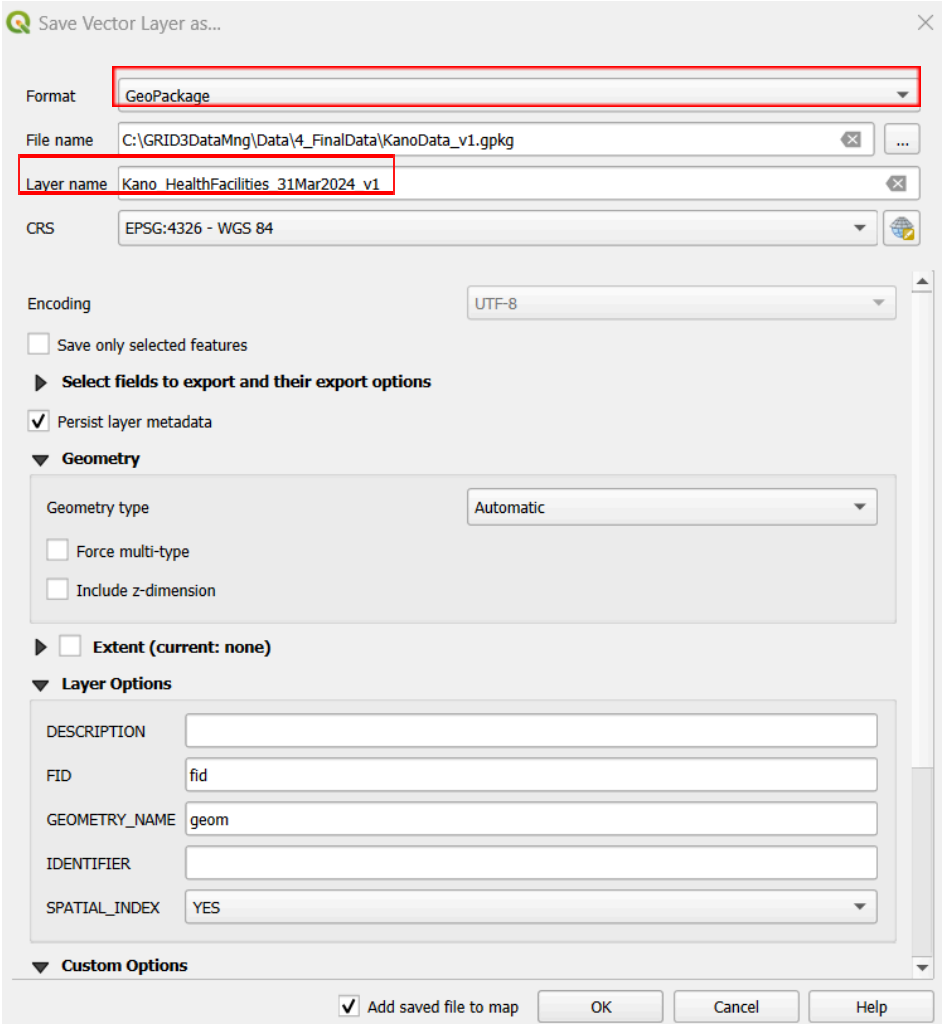
To preserve our work and make it accessible, we will save the final output as a **GeoPackage**. GeoPackages serve as comprehensive databases that can be easily shared among teams and partners.

- Right-click on the last created layer (Cleaned) > **Export** > **Save feature as**
- In the file format, choose **GeoPackage**
- In the output folder and name, save it in the 4_FinalData folder as: *...\GRID3DataMng\Data\4_FinalData\KanoData_v1.gpkg*

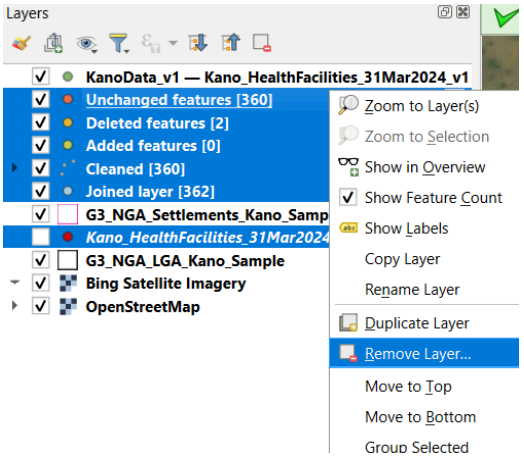
- Remember the naming conventions and name the layer

  Kano_HealthFacilities_31Mar2024_v1

- You may leave other options as standard

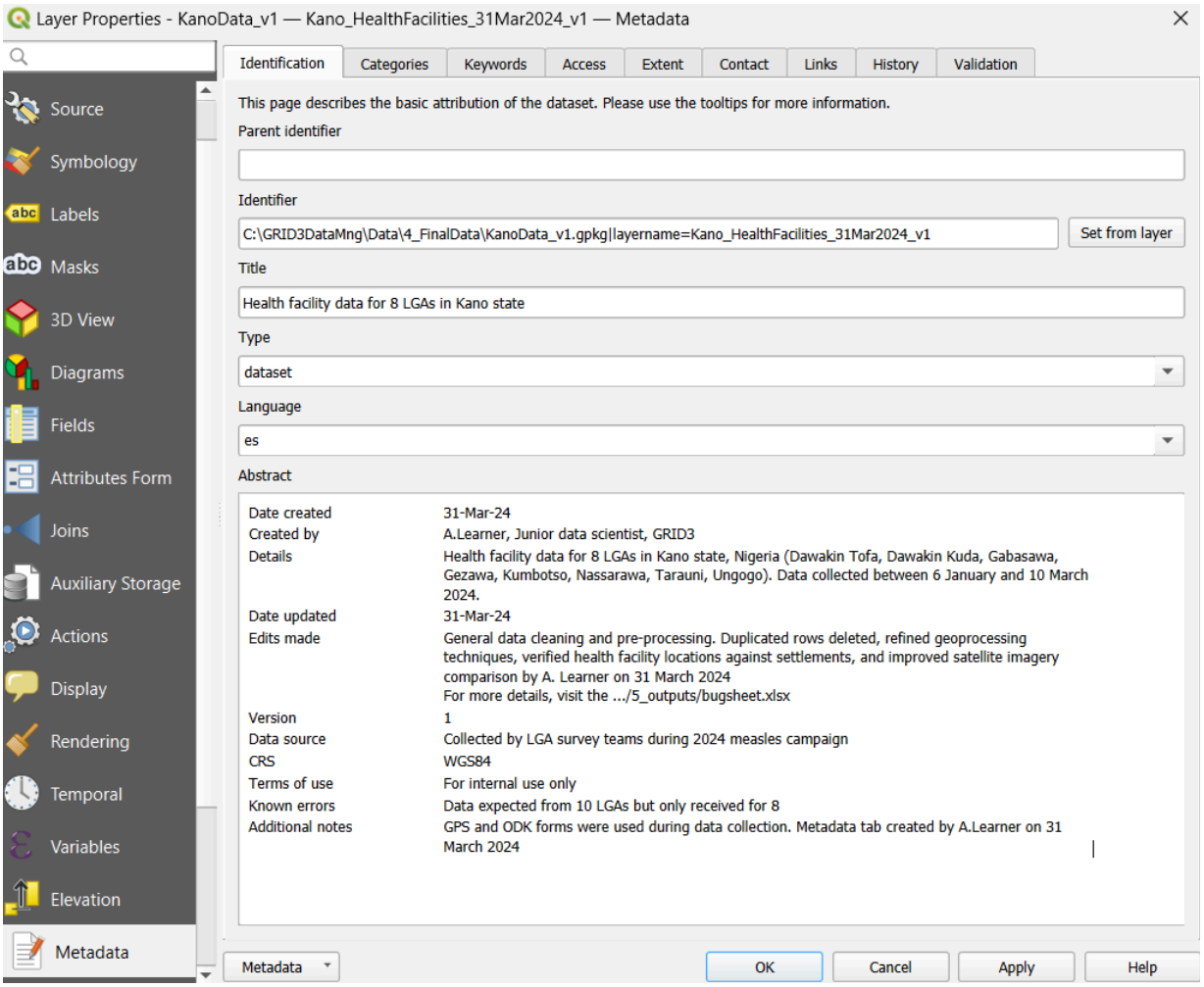


- Clean up your layers panel by removing any unnecessary files



4.3 Update metadata record and Save project

You are now required to finalize the metadata record for the dataset with the applied updates and modifications. Refer back to Part 2.1 for instructions on accessing the metadata tab and updating the metadata record; this step is crucial to ensure that all changes made during processing are accurately documented. Please ensure to finalize the metadata record before continuing and indicate where you have noted all of your edits.



- Save the project once again, and you're done!

Summary

Learning checklist

You have now completed the following steps, demonstrating good practice in data preparation and management for GIS:

- Addressed common issues such as misalignment, anomalies, and data inconsistencies
- Corrected anomalies based on visual inspection
- Employed rule-based symbology to identify and flag errors in the data
- Enhanced identification of erroneous health facility entries via updated satellite imagery and settlement extents
- Revised metadata records

You have now completed all four tasks in the GRID3 GIS data preparation and management course.