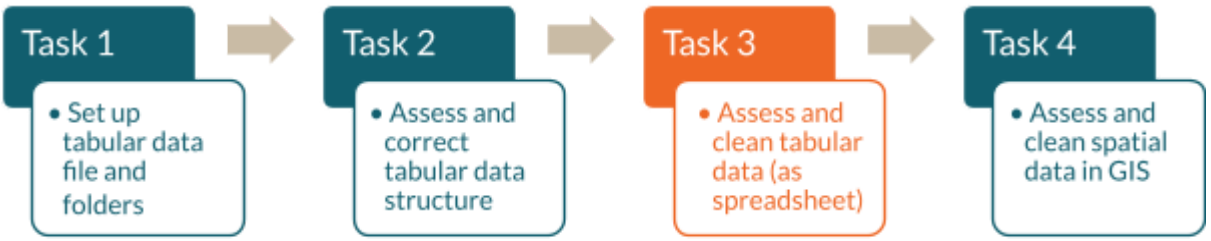
GIS DATA PREPARATION AND MANAGEMENT

Practical Task 3: Cleaning tabular data in preparation for use in GIS

Task description

This task will walk you through some common data errors and inconsistencies that are likely to appear in raw data files. You will look at how to identify these types of errors and think about which factors should be considered when making amendments.



It is important to note that the steps covered in this task are not a set process that will work for all data files. The skills learned will help you address common errors found in tabular data, ahead of its use for GIS analysis, but every dataset will require different levels of pre-processing and cleaning.

Data rules covered in this task

- Not all data will have the same errors! A good understanding of how the data were collected will help you to identify potential problems.
- Do not delete any data unless the process has been approved as part of your organization's data management protocol.
- If you make any changes to the data, create new intermediate data files.

Part 1: Categorical data standardization

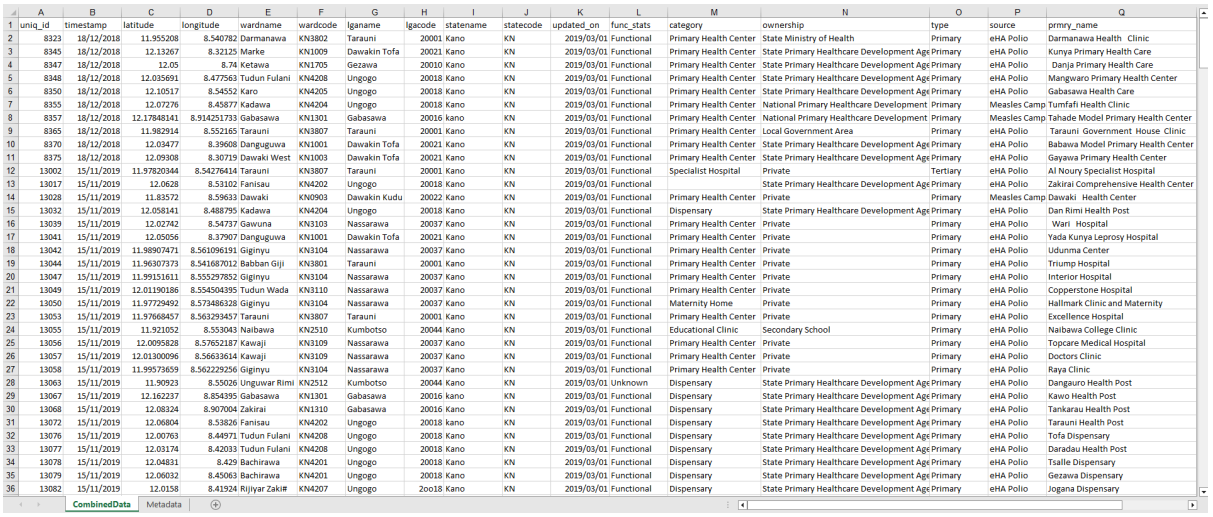
In Part 1 you will learn how to standardize and clean categorical data, looking specifically at correcting common errors, and prepare it for GIS use.

You may either continue working on the Excel file from Task 2 (saved in the 3_IntermediateData folder) or open the following file, which has been prepared specifically for this Practical Task 3:

…\GRID3DataMng\BackupFiles\Kano_HealthFacilities_TASK3.xlsx

Steps to follow:

1. **Open** and visually inspect the spreadsheet
2. Check the information in the **metadata** tab
3. Look at the **data** tab to scan for obvious issues



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	uniq_id	timestamp	latitude	longitude	wardname	wardcode	lganame	lgacode	statename	statecode	updated_on	func_stats	category	ownership	type	source	prmry_name
2	8323	18/12/2018	11.955208	8.540782	Darmanawa	KN3802	Tarauni	20001	Kano	KN	2019/03/01	Functional	Primary Health Center	State Ministry of Health	Primary	eHA Polio	Darmanawa Health Clinic
3	8345	18/12/2018	12.13267	8.32125	Marke	KN1009	Dawakin Tofa	20021	Kano	KN	2019/03/01	Functional	Primary Health Center	State Primary Healthcare Development Age	Primary	eHA Polio	Kunya Primary Health Care
4	8347	18/12/2018	12.05	8.74	Ketawa	KN1705	Gezawa	20010	Kano	KN	2019/03/01	Functional	Primary Health Center	State Primary Healthcare Development Age	Primary	eHA Polio	Danja Primary Health Care
5	8348	18/12/2018	12.035691	8.477563	Tudun Fulani	KN4208	Ungogo	20018	Kano	KN	2019/03/01	Functional	Primary Health Center	State Primary Healthcare Development Age	Primary	eHA Polio	Mangwaro Primary Health Center
6	8350	18/12/2018	12.10517	8.54552	Karo	KN4205	Ungogo	20018	Kano	KN	2019/03/01	Functional	Primary Health Center	State Primary Healthcare Development Age	Primary	eHA Polio	Gabasawa Health Care
7	8355	18/12/2018	12.07276	8.45877	Kadawa	KN4204	Ungogo	20018	Kano	KN	2019/03/01	Functional	Primary Health Center	National Primary Healthcare Development	Primary	Measles Camp	Tumfafi Health Clinic
8	8357	18/12/2018	12.17848141	8.914251733	Gabasawa	KN1301	Gabasawa	20016	kano	KN	2019/03/01	Functional	Primary Health Center	National Primary Healthcare Development	Primary	Measles Camp	Tahade Model Primary Health Center
9	8365	18/12/2018	11.982914	8.552165	Tarauni	KN3807	Tarauni	20001	Kano	KN	2019/03/01	Functional	Primary Health Center	Local Government Area	Primary	eHA Polio	Tarauni Government House Clinic
10	8370	18/12/2018	12.03477	8.39608	Danguguwa	KN1001	Dawakin Tofa	20021	Kano	KN	2019/03/01	Functional	Primary Health Center	State Primary Healthcare Development Age	Primary	eHA Polio	Babawa Model Primary Health Center
11	8375	18/12/2018	12.09308	8.30719	Dawaki West	KN1003	Dawakin Tofa	20021	Kano	KN	2019/03/01	Functional	Primary Health Center	State Primary Healthcare Development Age	Primary	eHA Polio	Gayawa Primary Health Center
12	13002	15/11/2019	11.97820344	8.54276414	Tarauni	KN3807	Tarauni	20001	Kano	KN	2019/03/01	Functional	Specialist Hospital	Private	Tertiary	eHA Polio	Al Noury Specialist Hospital
13	13017	15/11/2019	12.0628	8.53102	Fanisau	KN4202	Ungogo	20018	Kano	KN	2019/03/01	Functional		State Primary Healthcare Development Age	Primary	eHA Polio	Zakirai Comprehensive Health Center
14	13028	15/11/2019	11.83572	8.59633	Dawaki	KN0903	Dawakin Kudu	20022	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	Measles Camp	Dawaki Health Center
15	13032	15/11/2019	12.058141	8.488795	Kadawa	KN4204	Ungogo	20018	Kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Dan Rimi Health Post
16	13039	15/11/2019	12.02742	8.54737	Gawuna	KN3103	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Wari Hospital
17	13041	15/11/2019	12.05056	8.37907	Danguguwa	KN1001	Dawakin Tofa	20021	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Yada Kunya Leprosy Hospital
18	13042	15/11/2019	11.98907471	8.561096191	Giginyu	KN3104	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Udunma Center
19	13044	15/11/2019	11.96307373	8.541687012	Babban Giji	KN3801	Tarauni	20001	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Triump Hospital
20	13047	15/11/2019	11.99151611	8.555297852	Giginyu	KN3104	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Interior Hospital
21	13049	15/11/2019	12.01190186	8.554504395	Tudun Wada	KN3110	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Copperstone Hospital
22	13050	15/11/2019	11.97729492	8.57486328	Giginyu	KN3104	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Maternity Home	Private	Primary	eHA Polio	Hallmark Clinic and Maternity
23	13053	15/11/2019	11.97668457	8.563293457	Tarauni	KN3807	Tarauni	20001	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Excellence Hospital
24	13055	15/11/2019	11.921052	8.553043	Naibawa	KN2510	Kumbotso	20044	Kano	KN	2019/03/01	Functional	Educational Clinic	Secondary School	Primary	eHA Polio	Naibawa College Clinic
25	13056	15/11/2019	12.0095828	8.57652187	Kawaji	KN3109	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Topcare Medical Hospital
26	13057	15/11/2019	12.01300096	8.56633614	Kawaji	KN3109	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Doctors Clinic
27	13058	15/11/2019	11.99573659	8.562229256	Giginyu	KN3104	Nassarawa	20037	Kano	KN	2019/03/01	Functional	Primary Health Center	Private	Primary	eHA Polio	Raya Clinic
28	13063	15/11/2019	11.90923	8.55026	Unguwar Rimi	KN2512	Kumbotso	20044	Kano	KN	2019/03/01	Unknown	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Dangauro Health Post
29	13067	15/11/2019	12.162237	8.854395	Gabasawa	KN1301	Gabasawa	20016	kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Kawo Health Post
30	13068	15/11/2019	12.08324	8.907004	Zakirai	KN1310	Gabasawa	20016	kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Tankarau Health Post
31	13072	15/11/2019	12.06804	8.53826	Fanisau	KN4202	Ungogo	20018	kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Tarauni Health Post
32	13076	15/11/2019	12.00763	8.44971	Tudun Fulani	KN4208	Ungogo	20018	Kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Tofa Dispensary
33	13077	15/11/2019	12.03174	8.42033	Tudun Fulani	KN4208	Ungogo	20018	Kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Daradau Health Post
34	13078	15/11/2019	12.04831	8.429	Bachirawa	KN4201	Ungogo	20018	Kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Tsalle Dispensary
35	13079	15/11/2019	12.06032	8.45063	Bachirawa	KN4201	Ungogo	20018	Kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Gezawa Dispensary
36	13082	15/11/2019	12.0158	8.41924	Rijiyar Zaki#	KN4207	Ungogo	2oo18	Kano	KN	2019/03/01	Functional	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Jogana Dispensary

CombinedData | Metadata

Data tables containing a large amount of information are very difficult to inspect visually for mistakes. This task will take you through some simple processes in Excel that will help to identify potential errors in your data.

Potential errors to identify:

- Extra spaces
- Special/unusual characters
- Capitalization inconsistencies

Note: The following processes are not definitive and will not identify all potential errors in your data. These examples are designed to highlight the range of errors you should be looking for; the specific errors you find will depend heavily on the type of data you are using.

1.1 Remove extra spaces

Excel's **TRIM** function removes spaces from the beginning and end of text fields and changes any multiple spaces within text fields to single spaces.

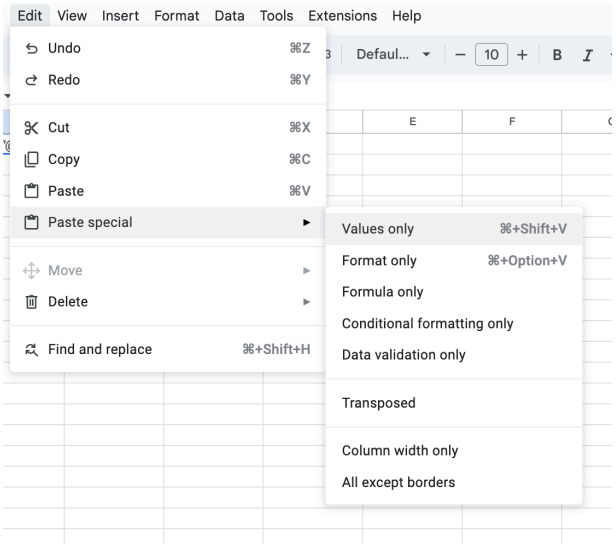
Steps to follow:

1. Sort data by **uniq_id** (smallest to largest)
2. Navigate to column Q (prmry_name)
3. Look at the top 15 health facility names–what do you notice?
4. Use the **TRIM** function–e.g. *=TRIM(Q2)*–in column R to remove leading and trailing spaces
5. Apply the function to all rows

	O	P	Q	R	S
1	**type**	**source**	**prmry_name**		
2	Primary	eHA Polio	Darmanawa Health Clinic	=TRIM(Q2)	
3	Primary	eHA Polio	Kunya Primary Health Care		
4	Primary	eHA Polio	Danja Primary Health Care		
5	Primary	eHA Polio	Mangwaro Primary Health Center		
6	Primary	eHA Polio	Gabasawa Health Care		
7	Primary	Measles Camp	Tumfafi Health Clinic		
8	Primary	Measles Camp	Tahade Model Primary Health Center		
9	Primary	eHA Polio	Tarauni Government House Clinic		
10	Primary	eHA Polio	Babawa Model Primary Health Center		
11	Primary	eHA Polio	Gayawa Primary Health Center		
12	Tertiary	eHA Polio	Al Noury Specialist Hospital		
13	Primary	eHA Polio	Zakirai Comprehensive Health Center		
14	Primary	Measles Camp	Dawaki Health Center		
15	Primary	eHA Polio	Dan Rimi Health Post		
16	Primary	eHA Polio	Wari Hospital		

6. Replace the data in column Q with the corrected data in column R

   Copy > **Paste special** >**Values Only** to replace column Q with the corrected data



7. Delete the duplicated column (column R)

Q	R
prmry_name	
Darmanawa Health Clinic	Darmanawa Health Clinic
Kunya Primary Health Care	Kunya Primary Health Care
Danja Primary Health Care	Danja Primary Health Care
Mangwaro Primary Health Center	Mangwaro Primary Health Center
Gabasawa Health Care	Gabasawa Health Care
Tumfafi Health Clinic	Tumfafi Health Clinic
Tahade Model Primary Health Center	Tahade Model Primary Health Center
Tarauni Government House Clinic	Tarauni Government House Clinic
Babawa Model Primary Health Center	Babawa Model Primary Health Center
Gayawa Primary Health Center	Gayawa Primary Health Center
Al Noury Specialist Hospital	Al Noury Specialist Hospital
Zakirai Comprehensive Health Center	Zakirai Comprehensive Health Center
Dawaki Health Center	Dawaki Health Center
Dan Rimi Health Post	Dan Rimi Health Post
Wari Hospital	Wari Hospital

You can repeat this process for any other columns that are likely to have spacing errors (any text data).

1.2 Correct special characters

Typos or unusual characters may affect how the data are displayed in GIS and prevent you from running some analysis tools. These errors should always be corrected prior to any geospatial analysis. These steps will show you how to identify and correct any errors related to special characters:

Steps to follow:

1. Navigate to the **wardname** column and scroll down to cell E36
2. Note the obvious "#" typo in the cell–this would not cause any problems in a string field, but as it is indeed a typo, it needs to be resolved
3. Look at cell H36, where you will notice another error
4. The letters in this cell would cause an error in the integer field and must be replaced with zeros (see Appendix 1 for more information on date types)

Now let's check both columns for repeats of the same errors:

Steps to follow:

1. Select the whole of column E (wardname) and press *Ctrl + H* to open the **Find and Replace** dialog box
2. Enter the special character (#) to find and replace it with the desired value (you can leave the **Replace** box empty to delete the special character rather than replace it with something)
3. Repeat this process for column H (lgacode) to check if there are any more cells that contain the letter "O"'instead of the number "0"

*Note: When using the **Find and Replace** tool for this example, find and replace individual selections. Using the **replace all** function, while useful in some cases, may result in changing cells across the entire table and is not recommended.*

1.3 Correct place name capitalization inconsistencies

Name inconsistencies can cause issues when carrying out statistical or spatial analysis. A common error is capitalization errors. In the **statename** column, you can see that some entries have a capital "K" and others have a lowercase "k." Let's standardize these to all be capitalized using the **PROPER** function.

Steps to follow:

1. Create a new column next to Column I (statename)
2. Use the **PROPER** function–e.g. *=PROPER(I2)*–to ensure that the first letter in each cell is capitalized
3. Apply this function to all rows
4. Copy and paste as values-only back into column I to replace the original data
5. Delete the new duplicated column

Part 2:Duplicated or missing records

In this section you'll learn how to identify and best manage missing or duplicate records.

2.1 Duplicated records

Duplicated records can sometimes be a mistake or data entry error, but they should never be deleted without careful consideration or further investigation.

Steps to follow:

1. Sort the data in alphabetical order for **prmry_name**
2. Navigate to rows 168 and 169 and note that there are two (2) entries with the same health facility name (Jemomi Health Post)
3. Review the rest of the information for these entries:
    - Are the location coordinates the same?
    - Are they located in the same ward?
    - Does the attribute data match?

Note: They are located in different wards and the location coordinates are different, meaning they are different facilities–not a duplicate

You may want to flag these health facilities for later review to double-check that one name has not been entered incorrectly.

4. Navigate to rows 317 and 318, where there are another two (2) health facilities with the same name (Wasai Health Post) and investigate

	A	B	C	D	E	G	M	N	O	P	Q
1	uniq_id	timestamp	latitude	longitude	wardname	lganame	category	ownership	type	source	prmry_name
316	35813	15/11/2019	12.021183	8.58309	Dakata	Nassarawa	Primary Health Center	National Primary Healthcare Development	Primary	eHA Polio	Warshu Hospital
317	13156	15/11/2019	12.09229	8.40487	Dawanau	Dawakin Tofa	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Wasai Health Post
318	13156	15/11/2019	12.09229	8.40487	Dawanau	Dawakin Tofa	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Wasai Health Post
319	35285	15/11/2019	12.036685	8.865334	Joda	Gabasawa	Dispensary	National Primary Healthcare Development	Primary	eHA Polio	Wasarde Health Post
320	13041	15/11/2019	12.05056	8.37907	Danguguwa	Dawakin Tofa	Primary Health Center	Private	Primary	eHA Polio	Yada Kunya Leprosy Hospital
321	35190	15/11/2019	12.083914	8.63162	Yada Kunya	Ungogo	Specialist Hospital	National Primary Healthcare Development	Tertiary	eHA Polio	Yadakunya Leprosy Center
322	35233	15/11/2019	11.985135	8.73479	Gawo	Gezawa	Dispensary	National Primary Healthcare Development	Primary	eHA Polio	Yafata Health Post
323	35035	15/11/2019	12.15648667	8.48274213	Tumfafi	Dawakin Tofa	Dispensary	National Primary Healthcare Development	Primary	eHA Polio	Yakasai Dandalama Health Post
324	13239	15/11/2019	12.270885	8.811603	Yumbu	Gabasawa	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Yama Health Post
325	13108	15/11/2019	12.16253	8.487	Tumfafi	Dawakin Tofa	Dispensary	State Primary Healthcare Development Age	Primary	eHA Polio	Yama Kanawa Dispensary

Note: The attributes and latitude/longitude coordinates are the same for these entries, meaning they are likely a duplicate.

5. Flag for later review or delete the duplicate (depending on the defined guidelines for your organization or team). For this training exercise, delete one of the duplicated rows

2.2 Missing data

What you do with missing data depends heavily on the data themselves and should always be looked at on a case-by-case basis. Sort the data again by column A (uniq_id). Between rows 71 and 101 there are five (5) rows with some missing data. Assess whether you should leave each cell with missing data blank, or if you have enough information to populate a given cell with data.

Review the following examples and recommendations:

- Missing state code or name (columns I and J)
  - Because all the data are from Kano State, you can fill in any missing state codes (KN) and names (Kano)
- Missing ward name (column E)
  - Use the ward codes (column F) to identify the missing ward name and fill in the cell
- Missing health facility category (column M)
  - Information such as health facility category, type, or ownership may be used to categorize the data in spatial or statistical analysis. Rather than leaving the cell blank, mark the cell with the word “unknown” (use the Find and Replace tool as you did earlier in the task). This will now show up as a separate category in GIS, but can still be represented on the maps.

Missing coordinate data (latitude and longitude) will result in data entries not displaying in GIS. Without coordinate data, the software doesn’t know where to place the data point. It is important to flag any entries that have missing coordinate data.

Ways to identify missing data

Conditional Formatting
Select the range of data you want to analyze and then click the Conditional Formatting button on the Home tab. Then choose the type of formatting you want to apply. Any cells that do not meet the criteria will be identified. This can be a great way to quickly identify any missing values in your data.

VLOOKUP Function
Select the range of data you want to analyze and then enter the VLOOKUP formula. Specify the column you want to search and then the column you want to return a value for. Once the formula is entered, any cells that do not meet the criteria will be identified.

Pivot Tables
Select the data range you want to analyze and then click the Pivot Table button on the Insert tab. From there, you can select the columns you want to analyze and choose the type of analysis you want to view; then you can quickly scan the data to identify any missing values.

Make sure you add a comment under the metadata tab giving a brief description of the data cleaning processes you have used; be sure to include the date and editor's name.

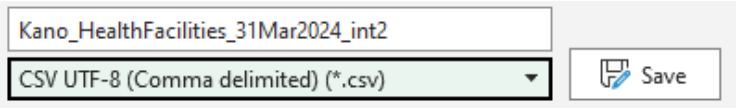
Date created	31-Mar-24
Created by	A.Learner, Junior data scientist, GRID3
Details	Health facility data for 8 LGAs in Kano state, Nigeria (Dawakin Tofa, Dawakin Kuda, Gabasawa, Gezawa, Kumbotso, Nassarawa, Tarauni, Ungogo). Data collected between 6 January and 10 March 2024.
Date updated	31-Mar-24
Edits made	General data cleaning and pre-processing. Duplicated row deleted (Wasai Health Post) by A.Learner on 31 March 2024
Version	1
Data source	Collected by LGA survey teams during 2024 measles campaign
CRS	Unknown
Terms of use	For internal use only
Known errors	Data expected from 10 LGAs but only received for 8
Additional notes	GPS and ODK forms were used during data collection. Metadata tab created by A.Learner on 31 March 2024

Part 3: Save data as CSV file

Now that you have prepared and cleaned your data, they are ready to be imported into QGIS, visualized, and checked for any further data issues. First, ensure you have saved an up-to-date version of your data as an Excel file (XLSX) in the 3_IntermediateData folder. There is already an intermediate file in this folder, so use a number system to differentiate between the two files.

…\GRID3DataMng\3_IntermediateData\Kano_HealthFacilities_31Mar2024_int2.xlsx

It is important to keep a copy of the data in .xlsx format, as this ensures that the metadata are stored within the same file as the data. However, Excel files are not compatible with QGIS. To successfully import the data into GIS, you must also save a copy of the data as a CSV file.



Comma-separated values (CSV) is a simple text file format that uses commas to separate values, and uses new lines to separate records. A CSV file stores tabular data (numbers and text) in plain text, where each line of the file represents one data record.

Only one sheet can be stored in a CSV file, meaning that the metadata will no longer be stored with the data themselves. This is why it is important to save a copy of the data in both formats.
Once you have saved the *CombinedData* sheet as a CSV UTF-8 file, you will see the tab name change to represent the sheet's new name.

Summary

Learning checklist

You have now completed the exercises in Task 3, demonstrating good practice in data preparation and management for GIS. You can confidently say you have:

- Saved a raw data file in a readily accessible, fixed directory with a consistent naming convention
- Navigated a pre-defined data management folder structure
- Created detailed metadata by using previous knowledge of the data

In the next task, you will address anomalies with the table structure that would generate errors when you import the data into GIS, and then will create an intermediate-level data file.

Appendix 1 – Data formats and fields in QGIS

Tabular data in QGIS can take any of the following formats:

Data Type	Name	Example
Text	String	Health facility
Whole number	Integer	321
Decimal number	Real	3.456
Date	Date (YYYY-MM-DD)	2016-07-28
Time	Time (HH:MM:SS+nn)	18:33:12+00
Date & Time	DateTime (YYYY-MM-DD HH:MM:SS+nn)	2016-07-28 18:33:12+00