Get every one in the picture

Data Quality 1

Workshop on data analysis and report writing for civil registration based vital statistics

Nadi, Fiji
30 January – 03 February 2023

Bloomberg Philanthropies
DATA FOR HEALTH INITIATIVE
Pacific Community
Communauté du Pacifique
Australian Bureau of Statistics
UNITED NATIONS ESCAP
Economic and Social Commission for Asia and the Pacific

Importance of data quality

- Poor data can lead to misleading analysis and subsequently, the potential for poorly informed decisions and policy-making

- Poor data costs money

- Need to establish **TRUST** in our data
  - This doesn’t mean it needs to be perfect
  - It does mean that it should be the best of what we have available and that we need to be honest about its limitations

Reviewing data quality should be continual

1. During data collection:
   - ➢ Review the systems to ensure data is collected using the correct methods and tools, and conduct routine quality control checks

2. At analysis stage:
   - ➢ Review individual records (unit record data)
   - ➢ Review tabulated data before further analysis
   - ➢ Review the plausibility of computed measures; including comparing the measures to other sources of information

Data sources

For the production of vital statistics during this workshop, we will be using:

- Birth and death data collected by the **civil registration office** and/or the **Ministry of Health**

- Population data derived from your most recent Census or projections (developed nationally or internationally e.g. by SPC or UNWPP) as denominator data to compute various indicators

Checking for errors

- Checking against other time periods and sources of vital statistics
  - Consistency checks should always be carried out, both on the data and the key indicators (e.g. birth and death rates) before they are used or made more widely available
  - Comparison of data can be done by checking against corresponding data from previous years.
    - E.g. significant changes from one year to another may require further investigation

Data cleaning: overview

Data cleaning steps:

1. **Setting up unit record data**: all required data fields and records have been carried over into the working spreadsheet
2. **Removing duplicate records**
3. **Excluding irrelevant data for our analysis:** inappropriate records have been excluded (for example, still births have been removed from live births data)
4. **Consistent variable names and data labels:** records use variables which are consistent and can therefore be readily aggregated
5. **Dealing with missing values**: using other data sources to fill the gaps

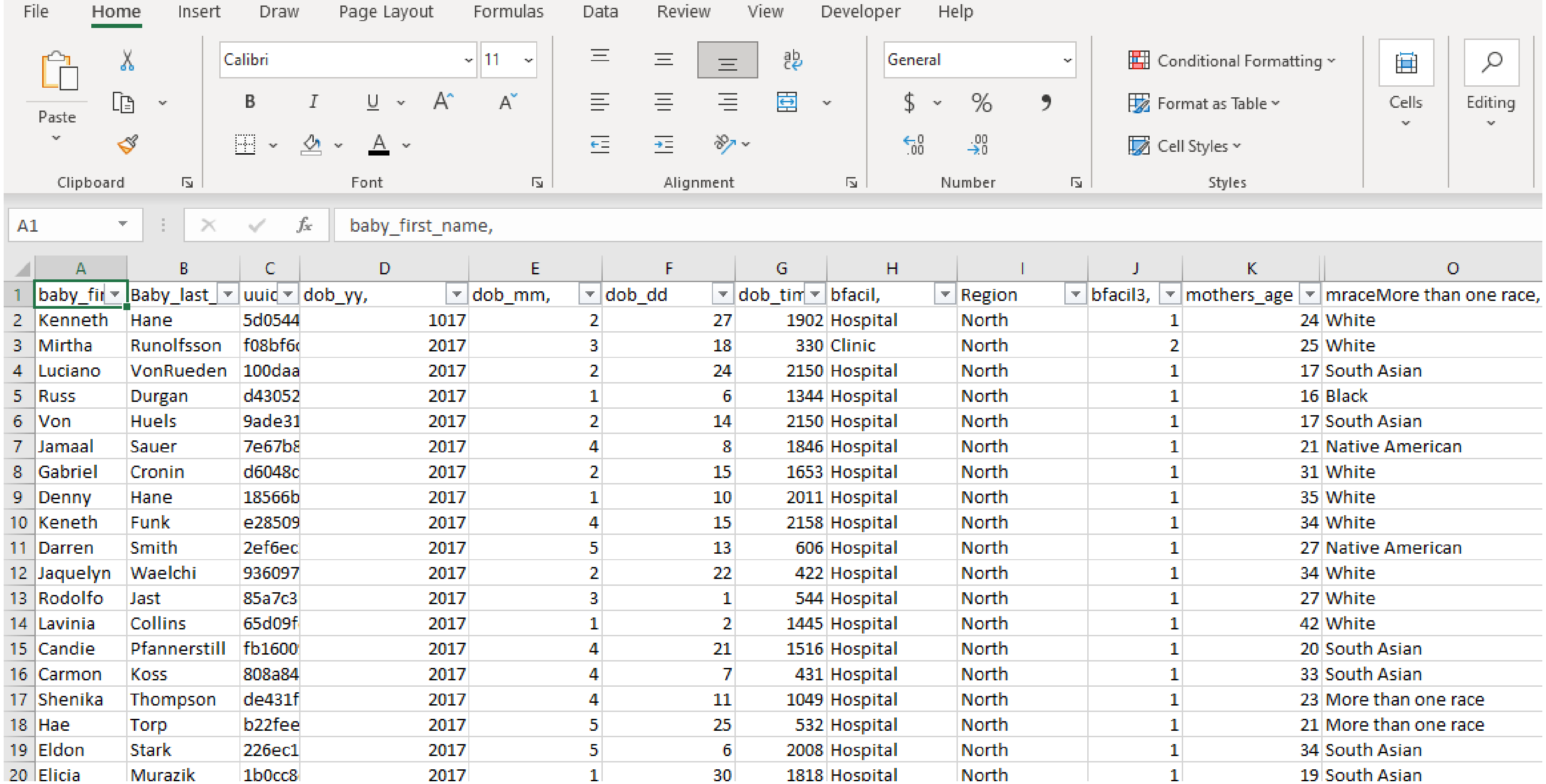
Important tip: Make sure to keep a note of every change you make to your dataset as you go along!!

Steps for setting up and cleaning data

Birth registration data	Death registration data
Date of Birth	Date of Birth
Date of registration of birth (if available)	Sex
Sex	Date of death
Place of Birth (Hospital, Health Facility, Home)	Date of death registration (if available)
Place of Residence (Village, Province, Island)	Age (use separate fields for days, months and years)
Mother’s date of birth	Place of death (Hospital, Health Facility, Community)
Mother’s age	Place of residence (Village, Province, Island)
Live or still birth (or all live births)	Causes of death (by line of death certificate – 1 variable per line) (if available) *(optional)*
Birth weight *(optional)*	Underlying cause of death (if available) *(optional)*
Length *(optional)*	Ethnicity *(optional)*
Weeks gestation *(optional)*	
Ethnicity *(optional)*	

Step 1: Setting up unit record data

One record (person) per row and one field (variable) per column



	A	B	C	D	E	F	G	H	I	J	K	O
1	baby_fir	Baby_last_	uuic	dob_yy,	dob_mm,	dob_dd	dob_tim	bfacil,	Region	bfacil3,	mothers_age	mraceMore than one race,
2	Kenneth	Hane	5d0544	1017	2	27	1902	Hospital	North	1	24	White
3	Mirtha	Runolfsson	f08bf6	2017	3	18	330	Clinic	North	2	25	White
4	Luciano	VonRueden	100daa	2017	2	24	2150	Hospital	North	1	17	South Asian
5	Russ	Durgan	d43052	2017	1	6	1344	Hospital	North	1	16	Black
6	Von	Huels	9ade31	2017	2	14	2150	Hospital	North	1	17	South Asian
7	Jamaal	Sauer	7e67b8	2017	4	8	1846	Hospital	North	1	21	Native American
8	Gabriel	Cronin	d6048c	2017	2	15	1653	Hospital	North	1	31	White
9	Denny	Hane	18566b	2017	1	10	2011	Hospital	North	1	35	White
10	Keneth	Funk	e28509	2017	4	15	2158	Hospital	North	1	34	White
11	Darren	Smith	2ef6ec	2017	5	13	606	Hospital	North	1	27	Native American
12	Jaquelyn	Waelchi	936097	2017	2	22	422	Hospital	North	1	34	White
13	Rodolfo	Jast	85a7c3	2017	3	1	544	Hospital	North	1	27	White
14	Lavinia	Collins	65d09f	2017	1	2	1445	Hospital	North	1	42	White
15	Candie	Pfannerstill	fb1600	2017	4	21	1516	Hospital	North	1	20	South Asian
16	Carmon	Koss	808a84	2017	4	7	431	Hospital	North	1	33	South Asian
17	Shenika	Thompson	de431f	2017	4	11	1049	Hospital	North	1	23	More than one race
18	Hae	Torp	b22fee	2017	5	25	532	Hospital	North	1	21	More than one race
19	Eldon	Stark	226ec1	2017	5	6	2008	Hospital	North	1	34	South Asian
20	Elicia	Murazik	1b0cc8	2017	1	30	1818	Hospital	North	1	19	South Asian

Step 1 (cont): setting up unit record data

1. Do not work on the original data set > copy the data into a working spreadsheet

1. When extracting data, ensure that all records are transferred:

    - Check the totals against the original source (such as the database)

    - Look at the total number of records and make sure that it is within an expected range

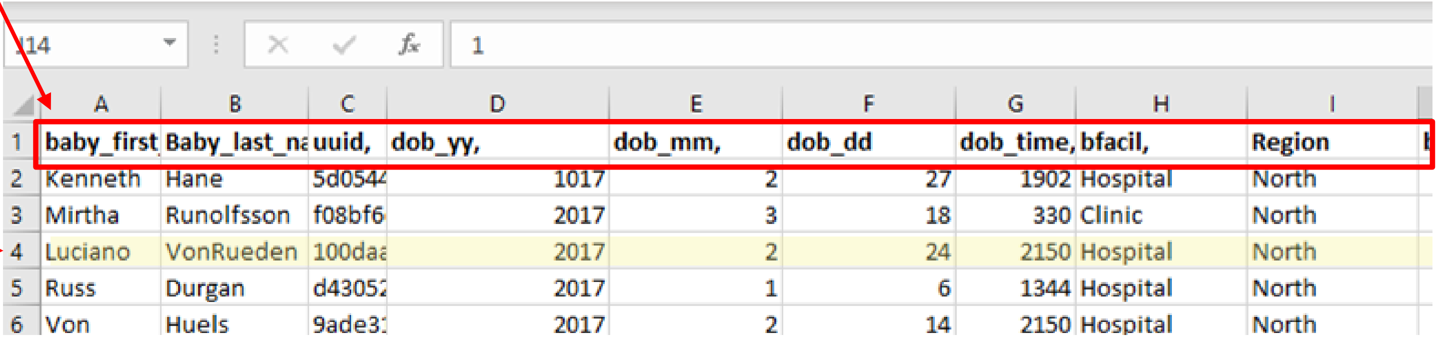
    - Make sure to remove any blank rows or columns

STEP 1 DEMONSTRATION: SETTING UP UNIT RECORD DATA

Note: It’s important that you don’t work on the original dataset. Begin by copying your data into a new Excel spreadsheet.

1. The first row in your spreadsheet will be the fields (variable) names e.g. name, ID number, DOB. There should be **one field per column**.

2. Make sure you only have **one record (person) per row** in your Excel spreadsheet



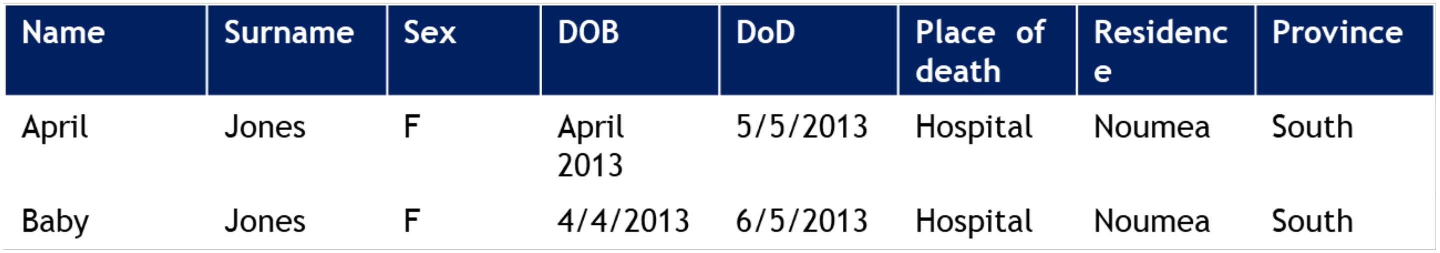
J14 | fx | 1

	A	B	C	D	E	F	G	H	I	J
1	baby_first	Baby_last_na	uuid,	dob_yy,	dob_mm,	dob_dd	dob_time,	bfacil,	Region	b
2	Kenneth	Hane	5d0544	1017	2	27	1902	Hospital	North	
3	Mirtha	Runolfsson	f08bf6	2017	3	18	330	Clinic	North	
4	Luciano	VonRueden	100daa	2017	2	24	2150	Hospital	North	
5	Russ	Durgan	d43052	2017	1	6	1344	Hospital	North	
6	Von	Huels	9ade31	2017	2	14	2150	Hospital	North	

Step 2: removing duplicate records

- Need to find and flag duplicate records before removing

- Questions to ask - How do we know if it is a duplicate? Do all fields have to be an exact match?



Name	Surname	Sex	DOB	DoD	Place of death	Residence	Province
April	Jones	F	April 2013	5/5/2013	Hospital	Noumea	South
Baby	Jones	F	4/4/2013	6/5/2013	Hospital	Noumea	South

Is this the same person?

- Which record will we use if the data is not exactly the same
- Be careful when checking for duplicates that you don’t remove twins.

Data Matching

- Data matching helps us to identify duplicate records, so that we can remove them.

- For deaths to be considered “matched” they must match on 3 of the following criteria (if surname included) or 4 if not.

    - Surname (similar spelling or sound OK)
    - Date or Death/Month of Report (same month)
    - First name (similar spelling or sound OK)
    - Island (place of death or report or residence)
    - Age at Death (within 1 year)

- Some possibility of under-matching when data quality poor (i.e. Insufficient data to match criteria)

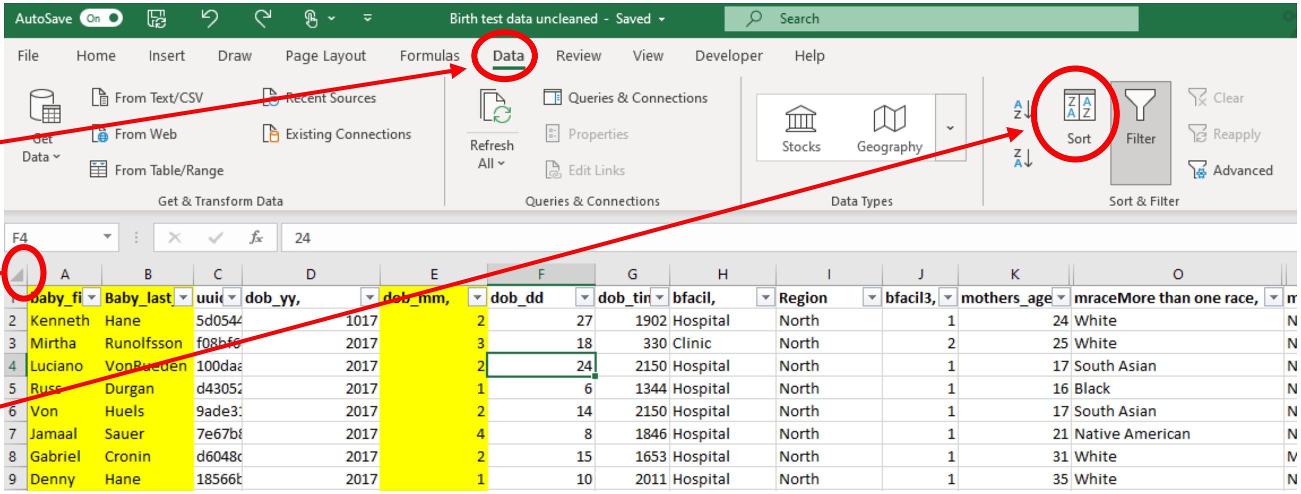
STEP 2 DEMONSTRATION: REMOVING DUPLICATE RECORDS

There are multiple ways to remove duplicate records in Excel. The most straightforward method of checking for duplicates is to use the **Sort** function in Excel.

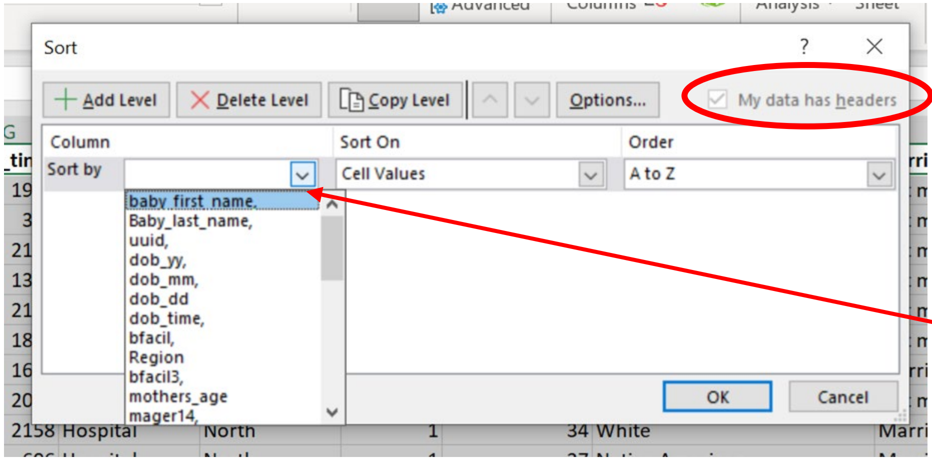
As mentioned in the previous slide, it is important that you sort on 3 or more fields (e.g. first name, surname and month of birth). This makes it easier to go through your data, line by line and identify records that match on all 3 fields.

Here is an example, using first name, surname and month of birth to check for duplicate records:

1. Click on the **Data** tab. Then, make sure that all of your data is selected by clicking on the arrow in the top left corner of you spreadsheet, and then click on **Sort**.



	A	B	C	D	E	F	G	H	I	J	K	O	
1	baby_fi	Baby_last	uuic	dob_yy,	dob_mm,	dob_dd	dob_tin	bfacil,	Region	bfacil3,	mothers_age	mraceMore than one race,	m
2	Kenneth	Hane	5d0544	1017	2	27	1902	Hospital	North	1	24	White	N
3	Mirtha	Runolfsson	f08bf6	2017	3	18	330	Clinic	North	2	25	White	N
4	Luciano	VonRueden	100daa	2017	2	24	2150	Hospital	North	1	17	South Asian	N
5	Russ	Durgan	d43052	2017	1	6	1344	Hospital	North	1	16	Black	N
6	Von	Huels	9ade31	2017	2	14	2150	Hospital	North	1	17	South Asian	N
7	Jamaal	Sauer	7e67b8	2017	4	8	1846	Hospital	North	1	21	Native American	N
8	Gabriel	Cronin	d6048c	2017	2	15	1653	Hospital	North	1	31	White	M
9	Denny	Hane	18566b	2017	1	10	2011	Hospital	North	1	35	White	N



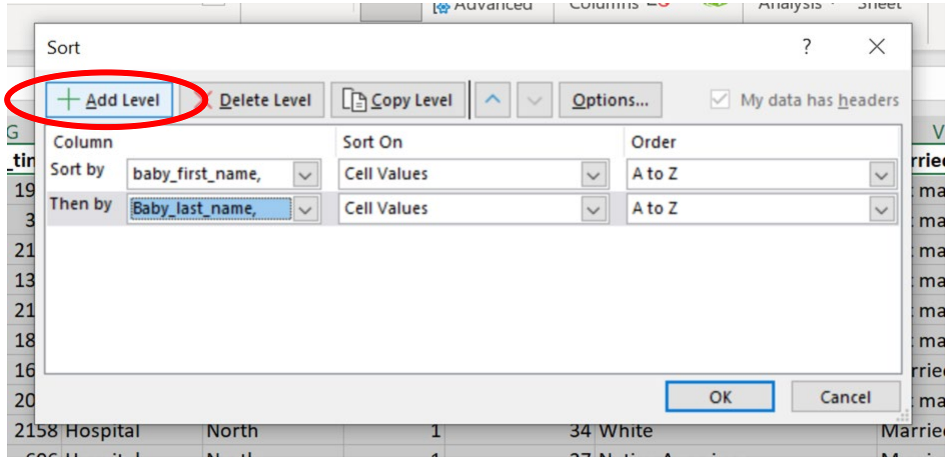
Column	Sort On	Order
Sort by	Cell Values	A to Z

baby_first_name,
Baby_last_name,
uuid,
dob_yy,
dob_mm,
dob_dd
dob_time,
bfacil,
Region
bfacil3,
mothers_age
mager14,

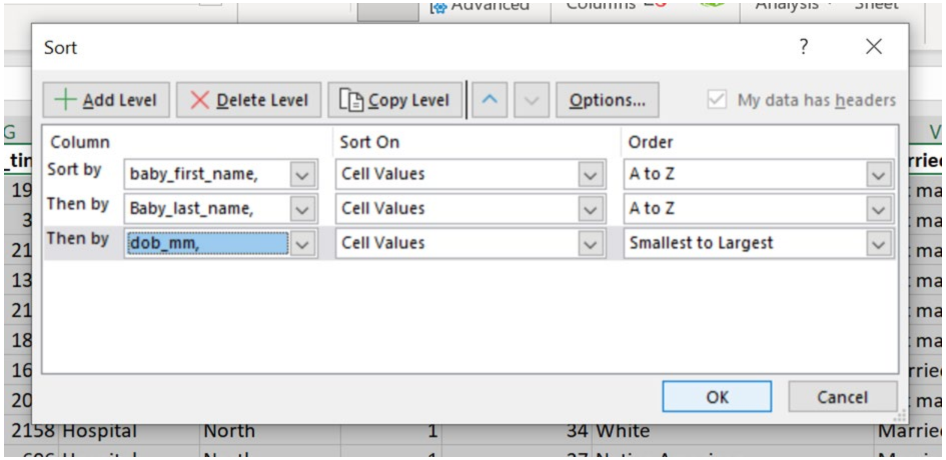
2. A pop-up box will appear. Make sure that the ‘**My data has headers**’ box is checked.

Then click on the drop down menu next to ‘**Sort by**’. This is where you will select the first field you want to sort on (first name).

3. Then click on ‘**+ Add level**’ to add the next field you want to sort by. In our case, this is surname.



Column	Sort On	Order
Sort by baby_first_name,	Cell Values	A to Z
Then by Baby_last_name,	Cell Values	A to Z

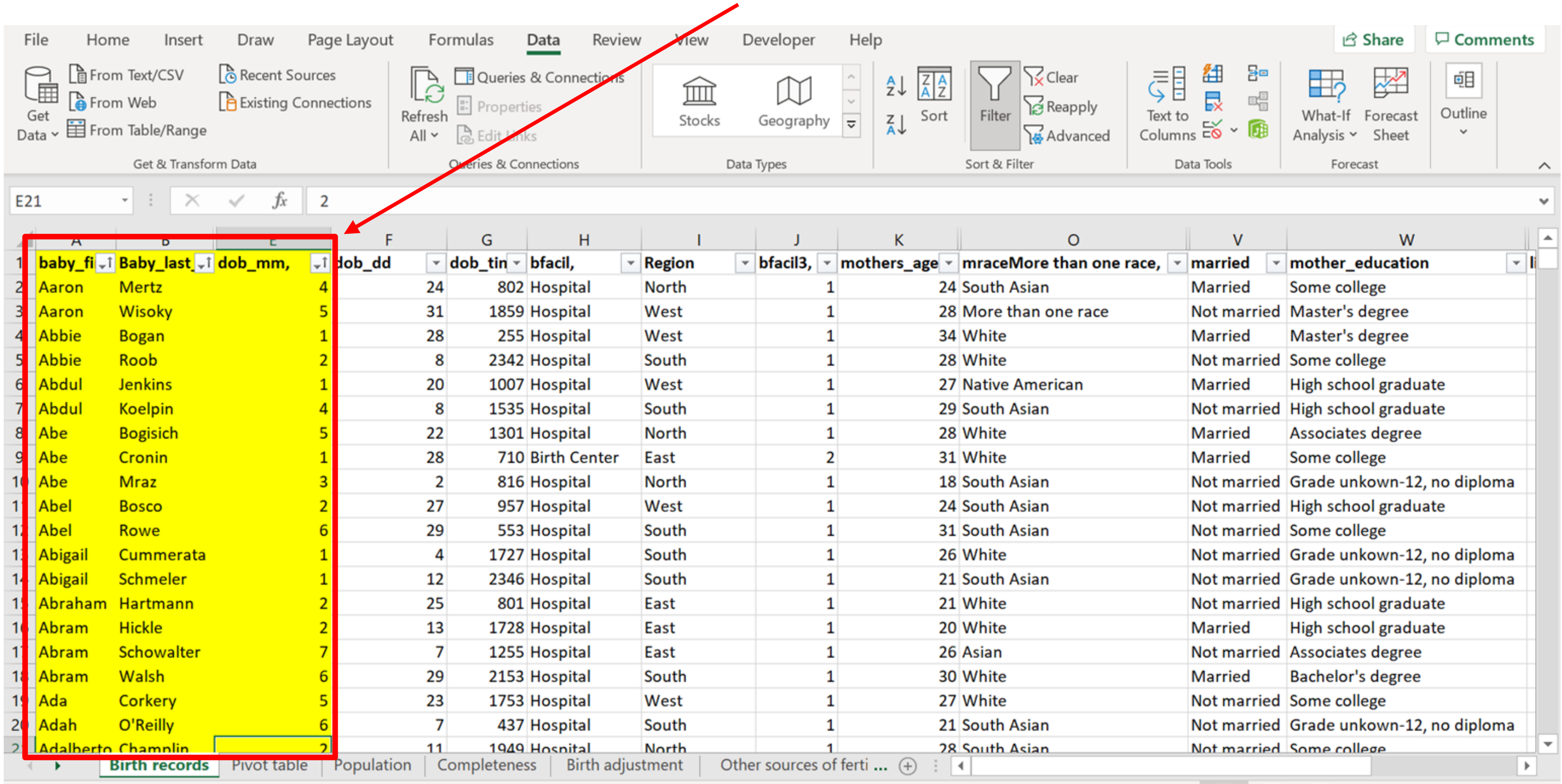


Column	Sort On	Order
Sort by baby_first_name,	Cell Values	A to Z
Then by Baby_last_name,	Cell Values	A to Z
Then by dob_mm,	Cell Values	Smallest to Largest

4. Repeat the previous step to add the third field you want to sort by. In our case, this is month of birth. Then click OK.

5. You'll notice that your data has now been sorted first by first name, then by surname and then by month of birth.

You can now scroll line by line and check for any matching records which have identical values for each of these 3 fields and remove records manually.



	A	B	E	F	G	H	I	J	K	O	V	W
1	baby_fi	Baby_last_	dob_mm,	dob_dd	dob_tin	bfacil,	Region	bfacil3,	mothers_age	mraceMore than one race,	married	mother_education
2	Aaron	Mertz	4	24	802	Hospital	North	1	24	South Asian	Married	Some college
3	Aaron	Wisoky	5	31	1859	Hospital	West	1	28	More than one race	Not married	Master's degree
4	Abbie	Bogan	1	28	255	Hospital	West	1	34	White	Married	Master's degree
5	Abbie	Roob	2	8	2342	Hospital	South	1	28	White	Not married	Some college
6	Abdul	Jenkins	1	20	1007	Hospital	West	1	27	Native American	Married	High school graduate
7	Abdul	Koelpin	4	8	1535	Hospital	South	1	29	South Asian	Not married	High school graduate
8	Abe	Bogisich	5	22	1301	Hospital	North	1	28	White	Married	Associates degree
9	Abe	Cronin	1	28	710	Birth Center	East	2	31	White	Married	Some college
10	Abe	Mraz	3	2	816	Hospital	North	1	18	South Asian	Not married	Grade unkown-12, no diploma
11	Abel	Bosco	2	27	957	Hospital	West	1	24	South Asian	Not married	High school graduate
12	Abel	Rowe	6	29	553	Hospital	South	1	31	South Asian	Not married	Some college
13	Abigail	Cummerata	1	4	1727	Hospital	South	1	26	White	Not married	Grade unkown-12, no diploma
14	Abigail	Schmeler	1	12	2346	Hospital	South	1	21	South Asian	Not married	Grade unkown-12, no diploma
15	Abraham	Hartmann	2	25	801	Hospital	East	1	21	White	Not married	High school graduate
16	Abram	Hickle	2	13	1728	Hospital	East	1	20	White	Married	High school graduate
17	Abram	Schowalter	7	7	1255	Hospital	East	1	26	Asian	Not married	Associates degree
18	Abram	Walsh	6	29	2153	Hospital	South	1	30	White	Married	Bachelor's degree
19	Ada	Corkery	5	23	1753	Hospital	West	1	27	White	Not married	Some college
20	Adah	O'Reilly	6	7	437	Hospital	South	1	21	South Asian	Not married	Grade unkown-12, no diploma
21	Adalberto	Champlin	2	11	1949	Hospital	North	1	28	South Asian	Not married	Some college

Note: You'll notice that columns 'C' and 'D' have been hidden. This isn't essential but it just makes it easier to view the fields that we are interested in (first name, surname, month of birth). You can do this by highlighting a column, right-clicking with the mouse and then selecting 'Hide'.

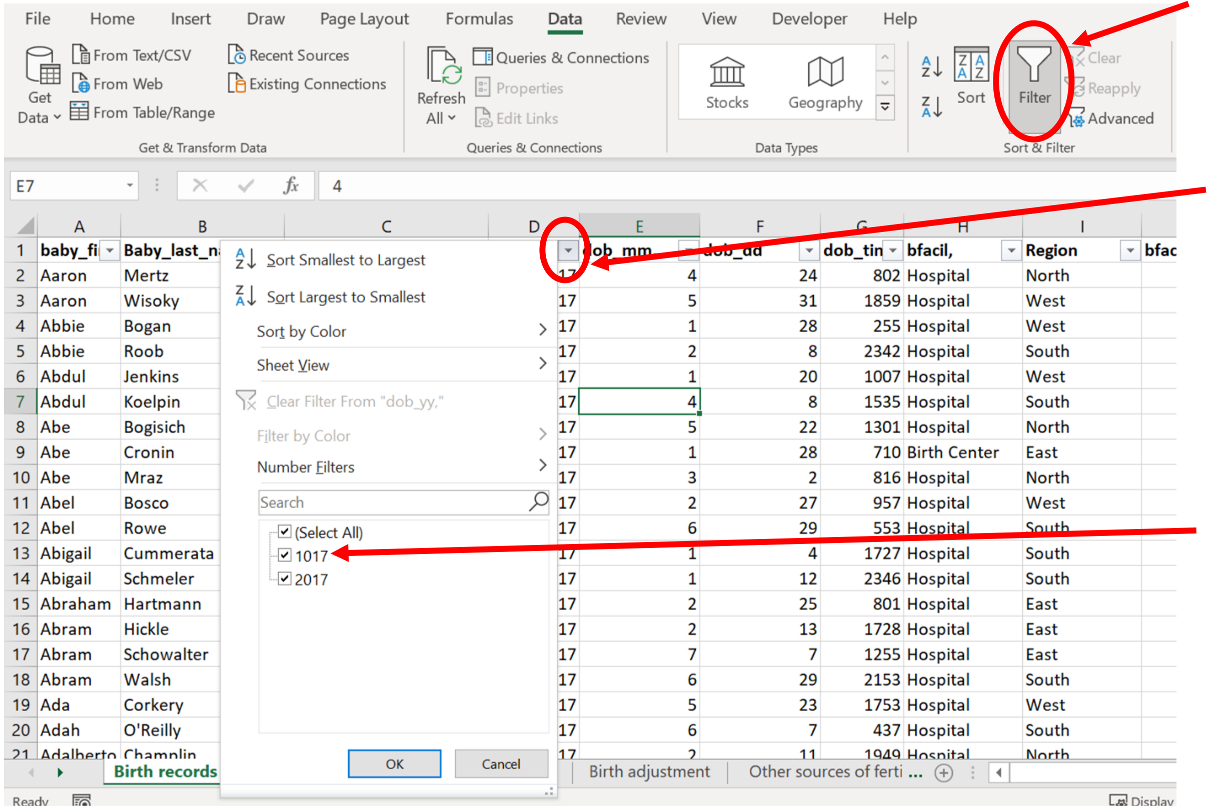
Step 3: excluding irrelevant data for our analysis

❖ Stillbirths should be in a different file (not part of live births or deaths)

❖ These are important events, but should be analyzed separately

STEP 3 DEMONSTRATION: EXCLUDING IRRELEVANT DATA

One of the easiest ways of checking for erroneous or irrelevant data is to use the ‘**Filter**’ function



	A	B	D	E	F	G	H	I	J
1	baby_fi	Baby_last_n		dob_mm	dob_dd	dob_tin	bfacil,	Region	bfac
2	Aaron	Mertz	17	4	24	802	Hospital	North	
3	Aaron	Wisoky	17	5	31	1859	Hospital	West	
4	Abbie	Bogan	17	1	28	255	Hospital	West	
5	Abbie	Roob	17	2	8	2342	Hospital	South	
6	Abdul	Jenkins	17	1	20	1007	Hospital	West	
7	Abdul	Koelpin	17	4	8	1535	Hospital	South	
8	Abe	Bogisich	17	5	22	1301	Hospital	North	
9	Abe	Cronin	17	1	28	710	Birth Center	East	
10	Abe	Mraz	17	3	2	816	Hospital	North	
11	Abel	Bosco	17	2	27	957	Hospital	West	
12	Abel	Rowe	17	6	29	553	Hospital	South	
13	Abigail	Cummerata	17	1	4	1727	Hospital	South	
14	Abigail	Schmeler	17	1	12	2346	Hospital	South	
15	Abraham	Hartmann	17	2	25	801	Hospital	East	
16	Abram	Hickle	17	2	13	1728	Hospital	East	
17	Abram	Schowalter	17	7	7	1255	Hospital	East	
18	Abram	Walsh	17	6	29	2153	Hospital	South	
19	Ada	Corkery	17	5	23	1753	Hospital	West	
20	Adah	O'Reilly	17	6	7	437	Hospital	South	
21	Adalberto	Champlin	17	2	11	1949	Hospital	North	

1. Under the ‘**Data**’ tab, click on ‘**Filter**’

2. When you click on the drop down arrow on a field header, a pop-up box will appear. In our case, we clicked on the arrow next to the year of birth. This box shows you all of the data values for this field (variable) within your dataset.

3. We can see here that under our year of birth, there is a value 1017 which is not a plausible value for year of birth, indicating an erroneous value. We can then sort the data by year of birth to find this record and decide what decision to take next. Perhaps we can use an alternative data source to find the correct year of birth for these records.

4. It’s a good idea to check all of your fields (variables) this way, to check for erroneous or irrelevant data. For example, if you have a field for ‘live birth’, you can make sure that there are no still births included in your dataset.

Step 4: consistent variable names and data labels

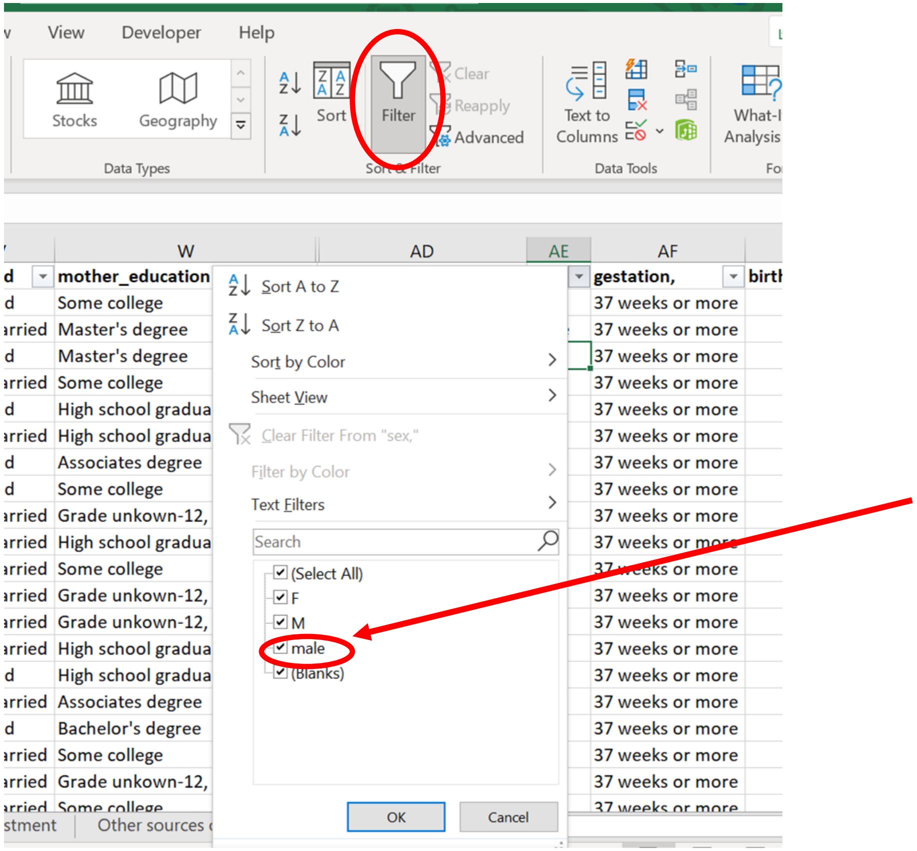
- Ensuring that variable names and data labels are consistent makes aggregation easier

- Variables should have been entered in a consistent manner – but this is not always the case, especially when using older data
  - In best practice – these should be controlled by your metadata standards

- Common problems
  - Sex: if we are using M/F for sex, then all records should have one of these values in the field, rather than some having recorded as male, Male, 1, etc.
  - Dates: Inconsistent data formats

STEP 4 DEMONSTRATION: CONSISTENT VARIABLE NAMES AND DATA LABELS

Again, we can use the ‘**Filter**’ function to check for consistent data labels [You can refer to the previous demonstration slides for detailed instructions].



d	mother_education	gestation,	birt
d	Some college	37 weeks or more	
arried	Master's degree	37 weeks or more	
d	Master's degree	37 weeks or more	
arried	Some college	37 weeks or more	
d	High school gradua	37 weeks or more	
arried	High school gradua	37 weeks or more	
d	Associates degree	37 weeks or more	
d	Some college	37 weeks or more	
arried	Grade unkown-12,	37 weeks or more	
arried	High school gradua	37 weeks or more	
arried	Some college	37 weeks or more	
arried	Grade unkown-12,	37 weeks or more	
arried	Grade unkown-12,	37 weeks or more	
arried	High school gradua	37 weeks or more	
d	High school gradua	37 weeks or more	
arried	Associates degree	37 weeks or more	
d	Bachelor's degree	37 weeks or more	
arried	Some college	37 weeks or more	
arried	Grade unkown-12,	37 weeks or more	
arried	Some college	37 weeks or more	

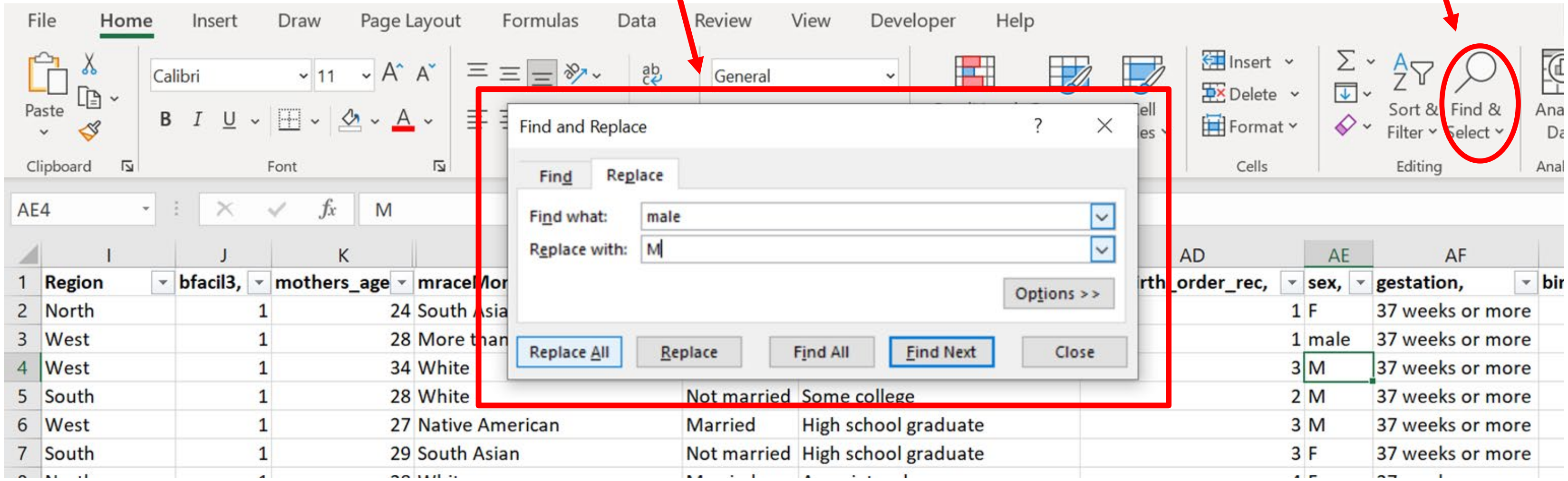
1. In our case, we wanted to check the data values for ‘sex’, to make sure that they are all labeled either *M* or *F*.

Using the filter function in Excel, we can see that there are some records which have sex coded as ‘male’, instead of *M* or *F*. We need to find these records and replace them with the correct data label.

We can then use the '**Find and Replace**' function, to replace the labels of data which have using wrong values.

2. Under the '**Home**' tab, select '**Find and Select**' button, and then click on '**Replace**'. A pop-up box will appear.

3. Type the label of the data that you want to replace inside the "**Find what**" box. In our case this is *male*.
Type the data label that you would like to replace it with. In our case, this is *M*.



AE4 | M

Find what: male

Replace with: M

	Region	bfacil3,	mothers_age	mrace/Mor			birth_order_rec,	sex,	gestation,
2	North	1	24	South Asia			1	F	37 weeks or more
3	West	1	28	More than			1	male	37 weeks or more
4	West	1	34	White			3	M	37 weeks or more
5	South	1	28	White	Not married	Some college	2	M	37 weeks or more
6	West	1	27	Native American	Married	High school graduate	3	M	37 weeks or more
7	South	1	29	South Asian	Not married	High school graduate	3	F	37 weeks or more

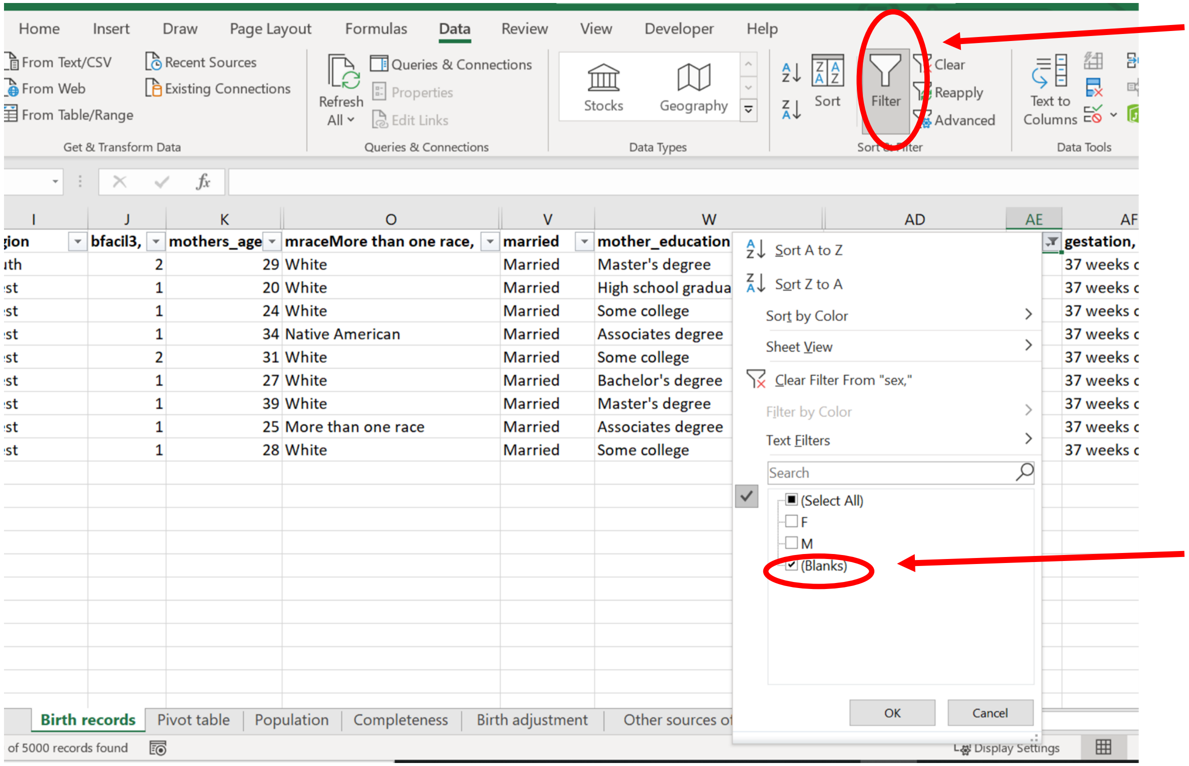
Step 5: dealing with missing values

- Is there original data missing?

- Can we obtain this information by using an alternative data source?
  - The first step is to see if we can obtain this information from a different data source. For example, you may need to combine data sources such as civil registration and health data.
  - Alternatively, methods are available for estimating this information e.g. if there is no age recorded for the mother in a birth record, but we have her date of birth, we can calculate this ourselves.
  - Finally, missing values can be redistributed following the distribution of recorded values. For example, the age distribution of deaths for which age at death was recorded (or the age of the mother, for births) can be applied to the missing values.

STEP 5 DEMONSTRATION: DEALING WITH MISSING VALUES

To check for missing values, we can use the ‘**Filter**’ function again.



I	J	K	O	V	W	AF
gion	bfacil3,	mothers_age	mraceMore than one race,	married	mother_education	gestation,
ith	2	29	White	Married	Master's degree	37 weeks c
st	1	20	White	Married	High school gradua	37 weeks c
st	1	24	White	Married	Some college	37 weeks c
st	1	34	Native American	Married	Associates degree	37 weeks c
st	2	31	White	Married	Some college	37 weeks c
st	1	27	White	Married	Bachelor's degree	37 weeks c
st	1	39	White	Married	Master's degree	37 weeks c
st	1	25	More than one race	Married	Associates degree	37 weeks c
st	1	28	White	Married	Some college	37 weeks c

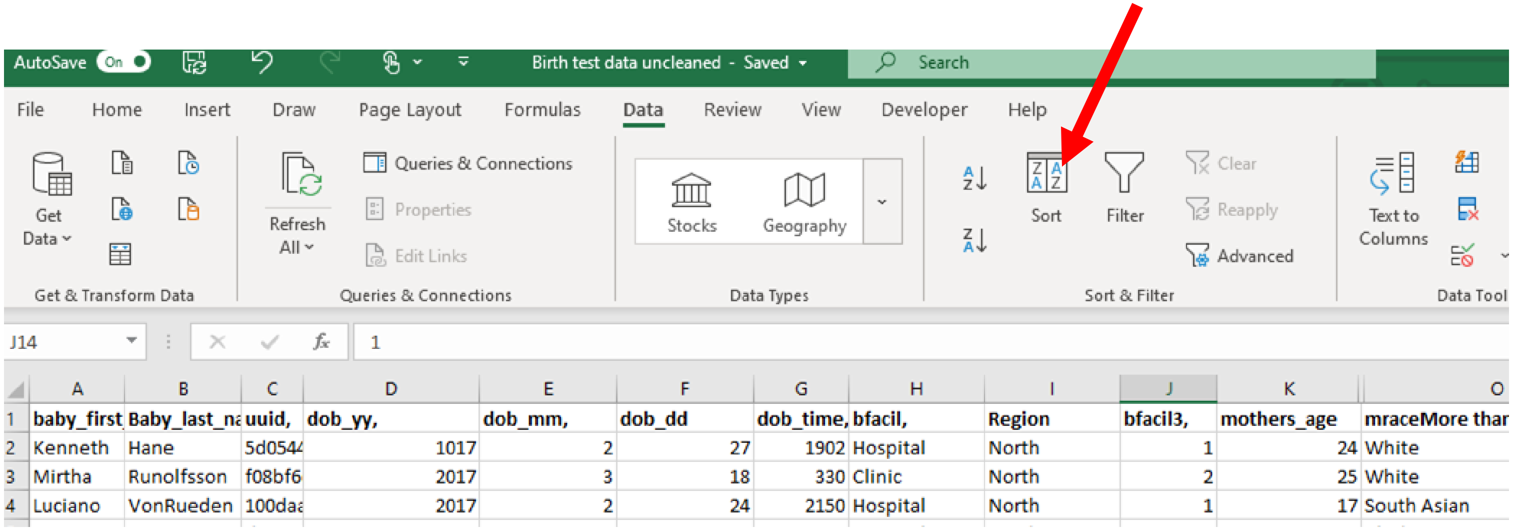
1. We want to check if there are any missing values for the field ‘*sex*’. Using the ‘**Filter**’ function as described in the previous demo slides, we can check for missing values. Click on ‘**Filter**’ under the ‘**Data**’ tab and then the drop down arrow on the field heading you want to explore.

2. We can see here that there are some missing values for ‘*sex*’, which are represented by the term **(Blanks).** If we select only the data that are blank we can go back to our dataset and see which records have missing data for this field and decide what action to take.

Other useful functions in excel for preparing data

SORT FUNCTION IN EXCEL

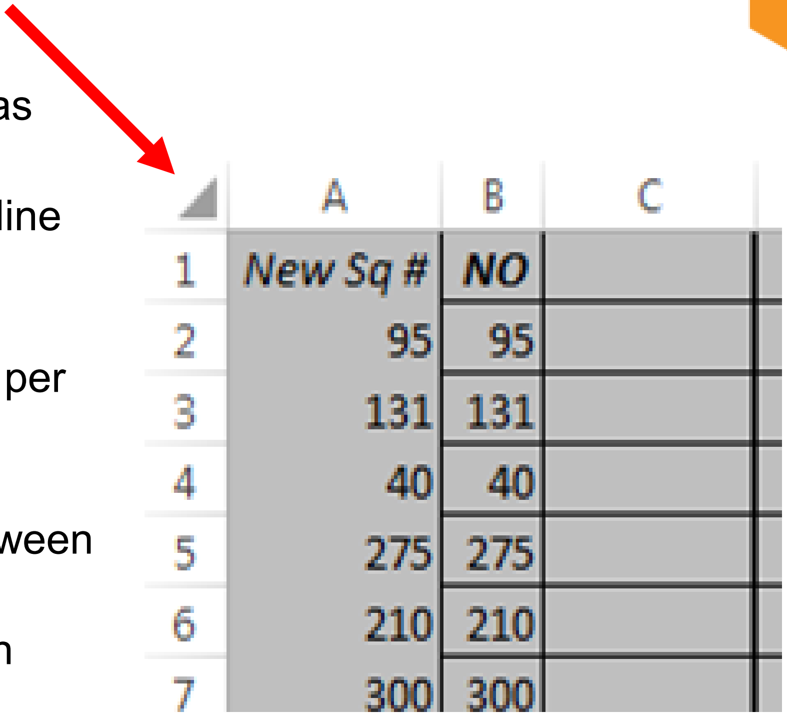
The most important tool when using excel to clean the data is the **sort function** which appears under the **data tab** in Excel. By clicking on the button marked, you can sort highlighted text by any of the fields in your data set.



	A	B	C	D	E	F	G	H	I	J	K	O
1	baby_first	Baby_last_na	uuid,	dob_yy,	dob_mm,	dob_dd	dob_time,	bfacil,	Region	bfacil3,	mothers_age	mraceMore thar
2	Kenneth	Hane	5d0544	1017	2	27	1902	Hospital	North	1	24	White
3	Mirtha	Runolfsson	f08bf6	2017	3	18	330	Clinic	North	2	25	White
4	Luciano	VonRueden	100daa	2017	2	24	2150	Hospital	North	1	17	South Asian

How to sort data in excel

1. Ensure that you are not using your original data as sometimes things go wrong!
2. Ensure that there is one record per line and one line per record — if this isn't the case, re-format your data.
3. Similarly, ensure that there's one field (variable) per column
4. When selecting data to sort –select ALL data by clicking on the arrow in the left upper corner (between the A and 1).
5. Ensure there are no blank columns or rows which may interrupt the sort function
6. Label your fields in the top row and make sure these are not repeated later in the data set.



	A	B	C
1	*New Sq #*	*NO*	
2	95	95	
3	131	131	
4	40	40	
5	275	275	
6	210	210	
7	300	300	

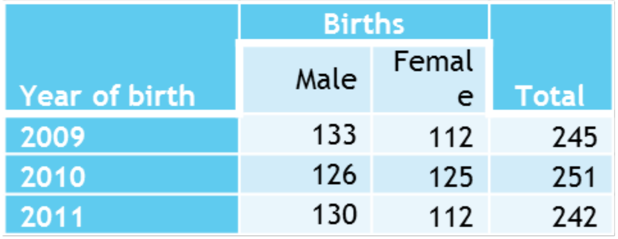
Pivot tables

- A pivot table is a special Excel tool that allows you to summarize and explore data interactively.

- Our worksheets contain a large set of population data;

  > In its current form, this data is hard to understand, because there's too much detail.

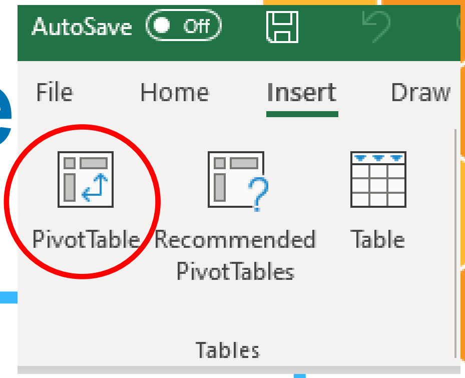
- To make sense of the information, we need to summarize it, and a pivot table is the perfect tool.



Year of birth	Births		Total
	Male	Female	
2009	133	112	245
2010	126	125	251
2011	130	112	242

Period of birth	Births		
	Male	Female	Total
2009-2011	389	349	738

How to build a pivot table



Basic steps to build a pivot table:

1. Highlight the data sheet by clicking in the top left corner (between A and 1)

2. On the Insert tab of the ribbon, click the PivotTable button

3. In the Create PivotTable dialog box, check/ select the data and click OK

4. Specify which variables to use as columns and rows to tabulate your data by moving them into the appropriate place

5. Use the count of function, and a variable which has no blanks to populate your table

6. Once a table is set up the way you want, copy it and paste it into a new worksheet, as pivot tables cannot be locked

Get every one in the picture

Q&A