SSA Data Management, Analysis & Report Training

Data Base Management

DEN	ResSe	ResAg	State	Count
1	M	38	1	1
2	M	29	1	2
3	M	46	1	3
4	M		2	4
5	M	83	2	5
6	M	28	2	6
7	M	36	3	7
8	M	44	3	8
9	M	40	3	9
10	M	43	3	9
11	M	35	4	10
12	M	47	4	11
13	M	53	4	12

Objectives

1. Be able to minimize errors during data processing
2. Be able to clean errors and/or outliers during processing.
3. Store data safely

Data Processing – Stages

- There are four stages in the processing of the data where errors may occur: data grooming, data capture, editing (cleaning) and estimation.

a) Data grooming involves preliminary checking before entering the data onto the processing system in the capture stage

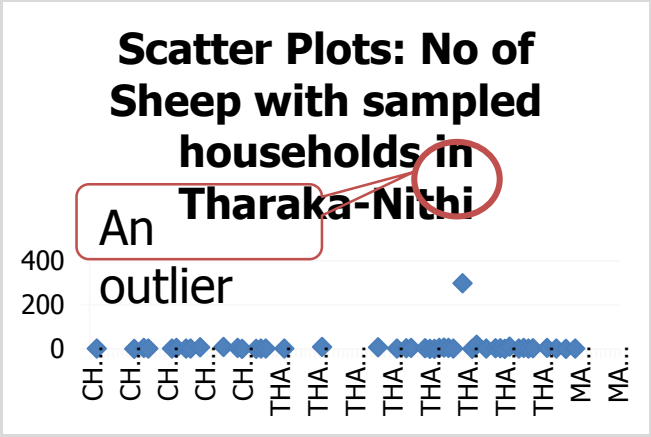
b) Data capture – input of data into a (computer) database . This may lead to introduction of errors into the data base

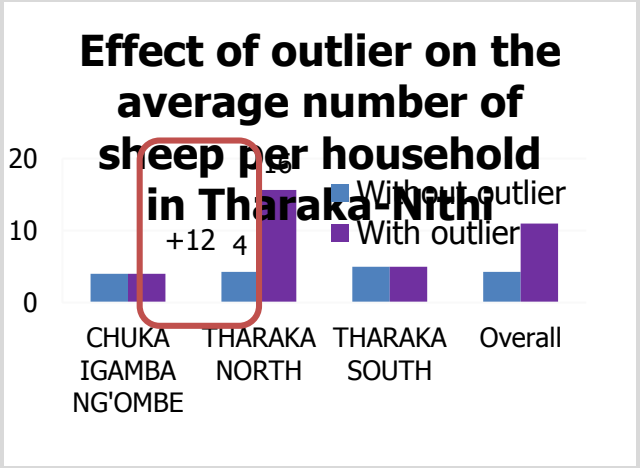
c) Editing/cleaning – the purpose is to eliminate and/or reduce errors

d) Estimations/Derivations of variable

Data Processing – Outliers

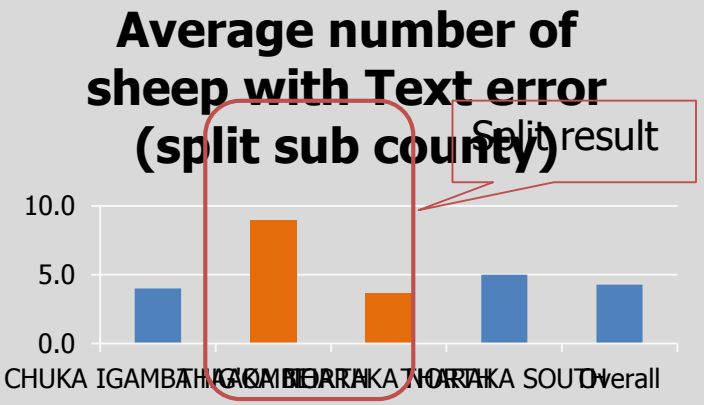
- Outlier. A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.
  - ➢ It may be a correct figure e.g. data from a very rich household (chief).
  - ➢ It may also be a data collection/entry error;
  - ➢ Better to exclude them from the data set as they could influence the outcome of data analysis (see graph example)





Data Processing – Errors & Outliers

- Error: a deviation from accuracy or correctness; a mistake. Errors can be introduced during data collection as well as data entry.



  - **Text errors** normally lead to of splitting of data during analysis. It is common with names of places, variety etc.

  - **Numerical error** may influence statistical analysis such sum, averages etc.

Data Processing Errors

- Common error during data capture

a) Entering data into a wrong record (row)

- ✓ This error can be reduced by hiding the already entered record – but this may bring about duplication of record

b) Duplication of records – entering the same record more than once

- ✓ Label each questionnaire with a single number and avoid having two or more questionnaire with same number

c) Entering data into a wrong field (Column) –

- ✓ This can be minimized by validating the fields or hiding other column or using different colors for the Column

Data Processing-Errors

- Common error during data capture

d) Mixing figures for two or more units of measurements eg. 200gm and 20kg – common with vegetables seeds and crop seed. Use 200gm and 20000gm or 0.2kg and 20kg

- ✓ Convert of all units into one unit and help data clerk understand this.

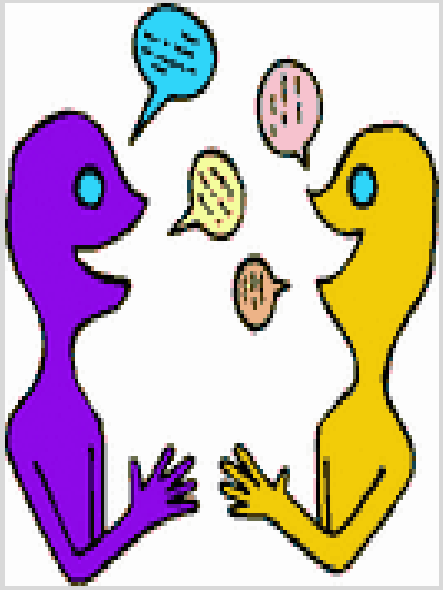
e) Mixing numeric (20) data with text (kg) in entering quantitative data – Alpha-numeric (20kg)

- ✓ Use only numeric - 20 but NOT 20kg.

f) Text and number errors - typing mistakes by the data clerk. These can be minimized through Cell validation, Copy and pasting, an filtering

Data Errors

- **Discussion:** A data clerk entered the following area (acre) planted with maize by 10 households; 2, 3.5, 45, 2, 4, 1, 1.5, 10, and 4.5. From KII famers in the area normally plant between 1-7 area of land:



a) In pairs: Identify the error, and discuss the possible source(s) of error data.

b) How would you handle this error?

Data cleaning, editing and verification

Right tool kit and skills

Definitions

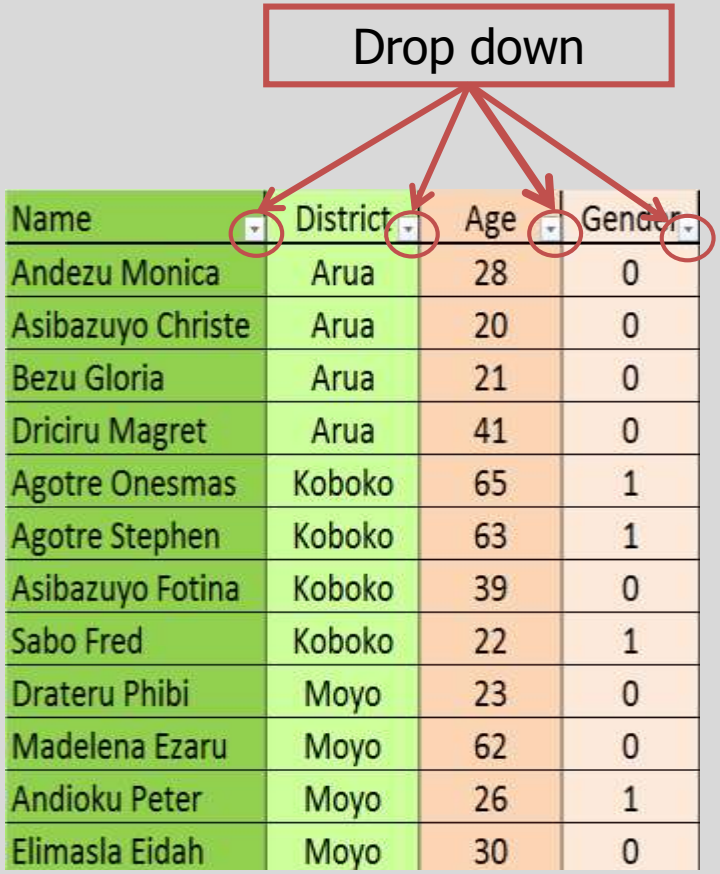
- Data cleaning - To ensure that there's is no inconsistent data, errors or outlier in the variables or duplicate in the records.
- Data Editing: - Prepare for analysis by correcting, condensing or otherwise modifying the data.
- Data Varification: - To establish the soundness, accuracy, or legitimacy of a data set. It can be through corroboration or support with other evidence (variable).
- In Microsoft Excel, data filters provides a very useful tool for data cleaning and editing.

Data Cleaning and editing using Filter

1. Highlight all the **VARIABLE HEADINGS**

2. Go to **DATA** menu and click on **SORT & FILTER** icon. Drop down menu will appear on the right side of every variable heading.

3. Click on the DROP DOWN ICON and scan for any inconsistent data or outlier within the list you see.



Name	District	Age	Gender
Andezu Monica	Arua	28	0
Asibazuyo Christe	Arua	20	0
Bezu Gloria	Arua	21	0
Driciru Magret	Arua	41	0
Agotre Onesmas	Koboko	65	1
Agotre Stephen	Koboko	63	1
Asibazuyo Fotina	Koboko	39	0
Sabo Fred	Koboko	22	1
Drateru Phibi	Moyo	23	0
Madelena Ezaru	Moyo	62	0
Andioku Peter	Moyo	26	1
Elimasla Eidah	Moyo	30	0

Data Cleaning using Filter

4. Once you have identified inconsistent data or outlier, first De-select **ALL**, then **SELECT** the inconsistent or outlier data. Click OK. Only selected one(s) will appear on the screen

5. Check the Data ENTRY NUMBER(S) corresponding to inconsistent or outlier data identified, Go back to the HARD Copy of the questionnaire and CORRECT.

6. Where the inconsistent or the outlier is existing in the hard copy, CONSULT the enumerator or team leader for correction.

Data cleaning: Procedure

7. In the event that neither the hard copy nor the enumerator/team leader can not help, the data manager will have to make judgment to OMIT (DELETE) the inconsistent our outlier data if it will affect the final analysis.

Number filter

- To apply a number filter, execute the following steps.

1. Click on the DROP DOWN ICON for a numeric data
2. Click Number Filters and select appropriate command e.g Greater than from the list.

Note

- You can also display records equal to a value, less than a value, between two values, the top x records, records that are above average, etc. The sky is the limit!
- This could be useful for picking outliers, grouping data etc.

Text filter

To apply a text filter, execute the following steps.

1. Click on the DROP DOWN ICON for a text data
2. Click Text Filters and select appropriate command e.g equals, begin with etc.

OR

1. Enter the text character or Name in the search area and the result will be displayed and pess OK

- Note: you can also display records that begin with a specific character, end with a specific character, contain or do not contain a specific character, etc. The sky is the limit!

Data Form

- The **data form** allows you to add, edit and delete records (rows) and display only those records that meet certain criteria. Especially when you have wide rows and you want to avoid repeated scrolling to the right and left, the data form can be useful.

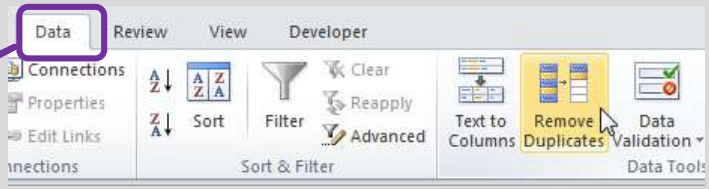
1. Click the Form command on the Quick Access Toolbar.
2. Use the Find Prev and Find Next buttons to easily switch from one record (row) to another.

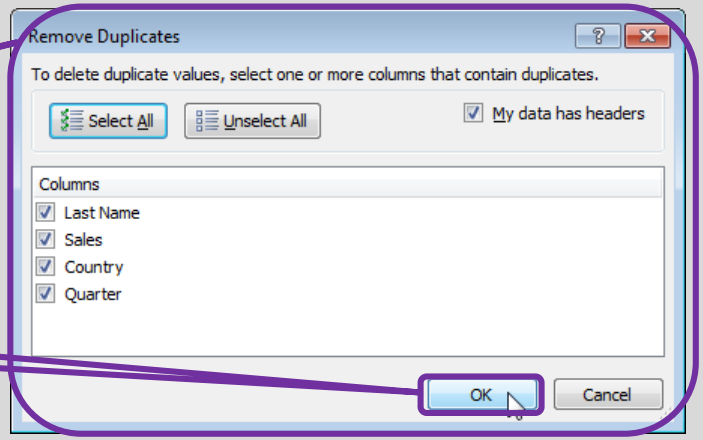
Cleaning Duplicates

- Duplicates are records entered more than once in the data-base. They tend to influence statistical analysis (sum, average, counts etc).

- Removing duplicates

1. Click any single cell inside the data set.
2. On the Data tab, click Remove Duplicates.
   - Dialog box appears with field levels will appear.
3. Leave all check boxes checked and click OK.





Data verification

1. Certain variables are related or linked to others. For example
   i. Amount of seed planted or production is normally linked to the area of land cultivated and crop planted .
      - A derived variables such as seed rate (seed planted/area) or yield may provide a useful hint to the validity of the data set collected.

   ii. Variety names are normally linked to crop names. Looking at the two variables (fields) at the same time helps in data cleaning.

Derived Variables

- **Derived variables-** are those that are generated from two or more set of variables – examples
  a) Seed rate = amount of seed planted ÷ area planted
  b) Yield = amount of crop produced ÷ area harvested
  c) Animal units = equivalent of all animals to one standard e.g. cattle

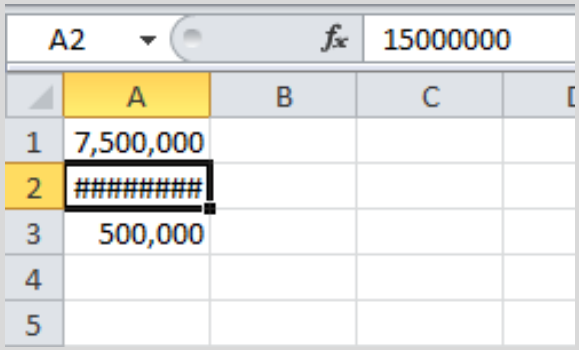
Steps

1. Insert a new Colum next to one of the variables
2. Name the variable field (column heading)
3. Apply appropriate formula in the first cell below the name
4. Extend the formulae within the cells of the column (field)

Derived Variables (Formula) Errors

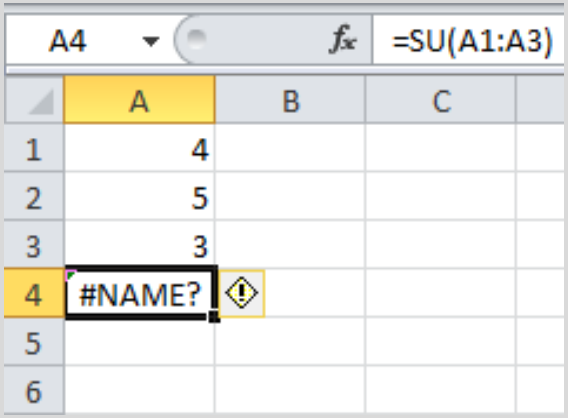
a) ########? Error When your cell contains this **error code**, the column isn't wide enough to display the value.

- ✓ Widen or double click column (Field) header



b) #NAME? error occurs when Excel does not recognize text in a formula.

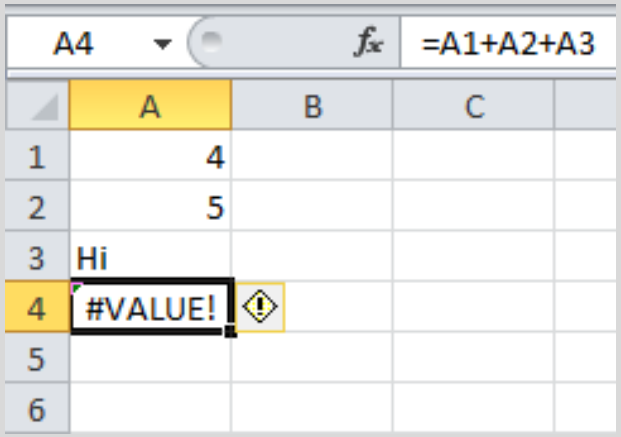
- ✓ Here correct the formula from **SU to SUM**



Derived Variable (Formula) Errors

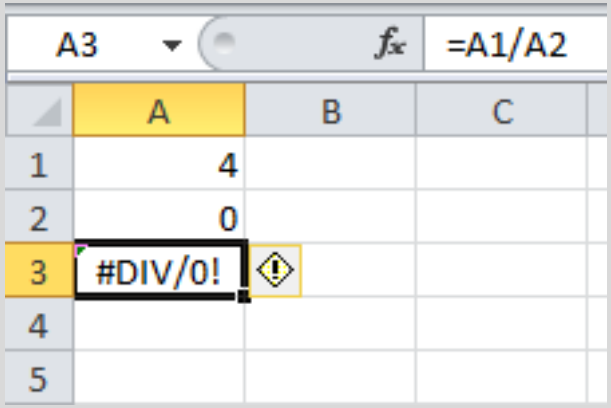
c) #VALUE! Error. Occurs when a formula has the wrong type of argument.

- ✓ Change the value of cell (A3) to a number or use function {=SUM(A1:A3)} to ignore cells that contain text



d) # **DIV/0! Error. Occurs** when a formula tries to divide a number by 0 or an empty cell.

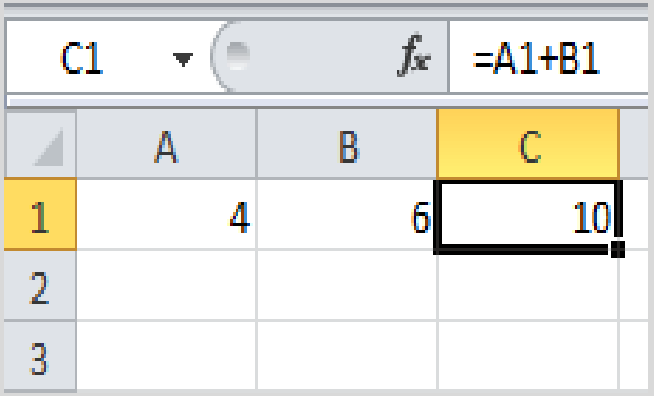
- ✓ Prevent the error from being displayed by using the logical function IF {=IF(A2=0," ",A1/A2)}
- ✓ Ignore and delete cells with #DIV/0!
- ✓ Change the value of the cell (A2) to a value that is not equal to 0.

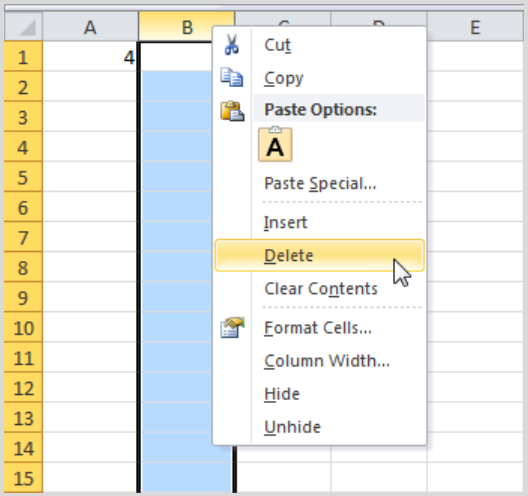


Derived Variable (Formula) Errors

c) #REF! error. Occurs when a formula refers to a cell that is not valid. Eg. Cell C1 references cell A1 and cell B1.

- This error occurs when one of the reference cell column or row is deleted
  - To fix this error, you can either delete +#REF! in the formula of cell B1 or you can undo your action by clicking Undo in the Quick Access Toolbar (or press CTRL + z).





Exercise 3.1: Data Cleaning and Editing

- You have been provided with the following data set for SSA conducted in 4 districts (Arua, Moyo, Adjumani and Koboko) in West Nile -Uganda.

1) Identify outliers, text and/or numerical errors by fields and records. What would be the effect of these outliers/errors on the fields? What would you do to minimize the effects.

2) Are there any duplicate in the data set? If so, how many duplicates? How would you remove the duplicate from the data set?

Exercise 3.2: Data Derivation and Verifications

- You have been provided with the following data set for SSA conducted 4 districts (Arua, Moyo, Adjumani and Koboko) in West Nile -Uganda.

1. Derive seed rates and yields (to one decimal point) of crops planted by famers in 2014 in West Nile districts. Identify and explain the type of formula errors, and suggest possible ways of fixing the errors.
2. Identify any outlier in the calculated seed rates and/or crop yields. Explain this could have come about and how you would deal with the outliers?
3. Verify whether records on the change of crop field between 2014 and 2015 are correct. Verify whether the reasons for change of crops match the change records.
   - **Note:** Change in crop was recorded as follows: 0= No change; 1= Change
4. Calculate seed rate change in 2015 compared to 2014, and categories these change as positive, negative and no change in another field.