

Cleaning Data

We start checking the data for any obvious errors and attempt to correct those errors as part of cleaning the data. This step occurs **PRIOR** to the actual data analysis (as much as possible).

1. Cleaning the data and preparing the dataset for analysis can be cumbersome if we have not spent enough time on the preceding events
 - a. Specifying a SMART research question
 - b. Specifying the variables of interest, their structure and interrelationships
 - c. Specifying the analysis we intend to perform, which helps in ensuring that we are collecting data pertaining to variables in the appropriate manner (example, if we want to present the mean age, we cannot collect data as 20-35 years, 36-50 years etc. We will have to collect data on age as a continuous variable)
 - d. Specifying the unit of measurement for each variable- height is marked in cms, weight is marked as Kg etc
 - e. Creating an appropriate data entry sheet
2. Please remember- the analysis should not take more than 5% of the total time pertaining to the project if the above steps are done well. If the above steps are not done well, the analysis may take even 90% of the allotted time for the project.

How do you clean the data?

1. Save the data using a different name- “XYZ data to clean”
2. Always keep the raw data entry separate so that you have a backup in case you make any errors while cleaning or analysing

3. Check to see that only the top row contains headers
4. There are no other header rows inserted in the excel sheet
5. There are no missing rows in the excel sheet
6. Check each variable based on the variable structure defined and within the plausible limits defined
7. Do not mark missing data as “missing”, “not done”, “not available” “forgot” etc. You can leave the cell blank, but better still, use a unique consistent numeric value to denote such categories. This becomes important for continuous variables, if we enter text, we cannot look at the distribution unless we clean the data such that those cells are left blank. Caveat- if we leave it blank, we do not know why the cell is blank. If we have structured the data entry sheet properly, this problem should not arise!
8. Check to see that the data is in a biologically accepted range. If we have structured the data entry sheet properly with predefined ranges, this should not be a problem.
9. Use the sort function in excel very carefully as mistakes in sorting can completely jumble up your data.
10. I usually prefer sorting data using the stats software....but if you want to use excel, then click data in the top tabs in excel, choose the cell of the variable to be sorted, then click sort, check my data has headers, then choose the variable from the drop down box, click ok. You can add multiple levels by clicking add level, thus I can sort by age and sex, etc.

for workshop - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Developer

From Access From Web From Text From Other Sources Existing Connections Refresh All Connections Properties Edit Links Sort Filter Advanced Text to Columns Remove Duplicates Data Validation Consolidate What-If Analysis Group Ungroup Subtotal Outline

D2 0

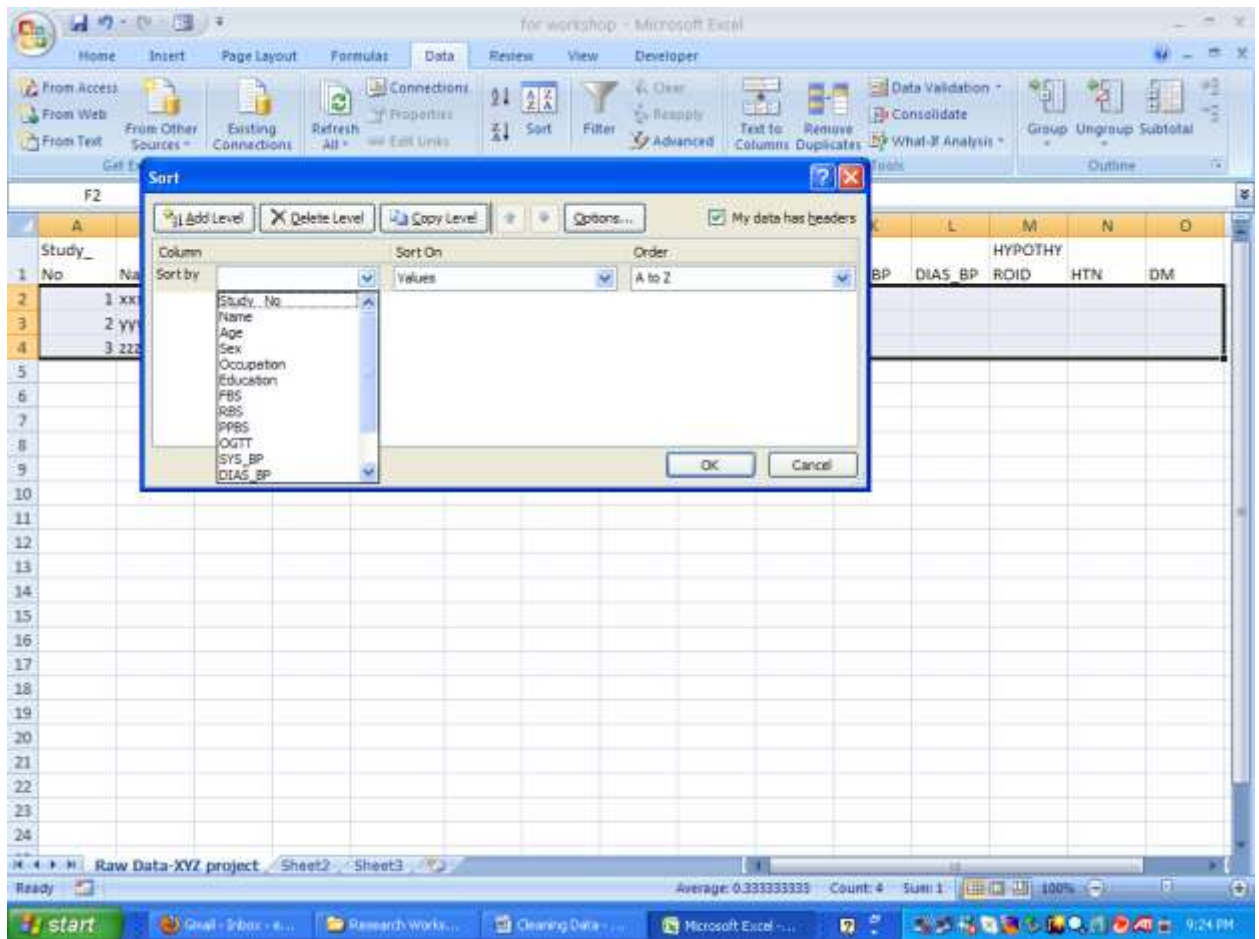
Study_	No	Name	Age	Sex	Occupation	Education	FBS	RBS	PPBS	OGTT	SYS_BP	DIAS_BP	HYPOTHY	ROID	HTN	DM
1	1	xxx	23	0	0	0										
2	2	yyy	24	1	1	1										
3	3	zzz	25	0	0	0										
4																
5																
6																
7																
8																
9																
10																
11																
12																
13																
14																
15																
16																
17																
18																
19																
20																
21																
22																
23																
24																

Raw Data-XYZ project Sheet2 Sheet3

Ready 100%

start Gmail - Inbox - s... Research Works... Clearing Data... Microsoft Excel... 9:27 PM

Clinical Research Unit



There is little cleaning that will be necessary if you have prepared your data sheet well.

The major cleaning will then be focused on determining values that are outliers- or lying outside a reasonable range of biologically plausible values (for instance a bmi of 200 etc- could be data entry error- 200 instead of 20 but even these reduce if you structure the data entry sheet well)

Cleaning data thus starts, **much before data entry** into the sheet- from the time you think of the variable.

Additionally, it is always better to recheck each row as soon as you complete entry to minimize data entry errors. The best is to do double data entry and check for errors...however, this may not always be possible.