How do we **process and analyse** the dataset?

Table of contents

Why processing the raw data is so important

After household survey data has been collected by enumerators and handed over to the Profiling Working Group, the profiling partners must now process the dataset and prepare it for analysis. In many cases, this will be the responsibility of the Profiling Coordinator, working alongside Information Management Officers or National Statistics Office staff.

Although every dataset has its peculiarities, there are several data processing practices that are almost always recommended to use. This guidance gives a list of good practices for processing and preparing **quantitative data** for analysis, as well as suggestions for how to resolve common issues that come up when processing data.

Having a thorough pilot that leaves time for reviewing and processing the data gives you the opportunity to spot errors.

Important to note is that while a lot of steps in data processing can be automated by using statistical software such as SPSS, R, and in some cases Excel, to detect errors, we cannot overstate that these software are not a replacement for basic analytical thinking and an understanding of the context.

For starters, to know what errors to look for in a dataset requires a thorough understanding of the questionnaire and the data collection process (i.e. which questions applied to which target populations, and which questions had the potential to be misinterpreted by respondents). Of course, good practice throughout the data collection process helps here: having a questionnaire coded on a mobile data collection device with controls to prevent basic errors in advance eases the burden on the people processing the data, and clear trainings and debriefs with enumerators ensures that issues during the interviews are kept at a minimum or are documented after.

This is one of the reasons why piloting, or doing a field test of the data collection, is so important. Having a thorough pilot that leaves time for reviewing and processing the data gives you the opportunity to spot these errors, fix or add more controls to the questionnaires, and retrain enumerators.

But even after all this careful work, there are always a few things you still need to do before you can dig into the data and start thinking about the results. **Even if you are not the one processing the data, it is important to know these main steps, as issues with the data typically cannot be resolved by the data processer alone.**

Coming in 2018: A guide for using **SPSS** and **R** to process the data

CHECKLIST

How do I process the data?

1. Before you start: making order out of chaos

- ☑ Prepare your process for documenting all changes to the data
- ☑ Be clear and consistent with folder structures and naming conventions
- ☑ Make sure you have all expected files containing all questions and respondents

2. Preparing the database: it's got to look good

- ☑ Make database more readable by adding labels where needed
- ☑ Merge household data with individual-level data
- ☑ Create tables with the frequency distributions of each variable to have overview and to start detecting errors

3. Cleaning common errors: the easy stuff

- ☑ Check if respondents answered all questions intended for them to answer
- ☑ Review open-ended answers and categorize them
- ☑ Resolve typos in answers

4. Checking for unexpected values: any weird results?

- ☑ Check for unexpected results, including outliers and oddities
- ☑ Check for contradictions in respondents' answers to different questions
- ☑ Check for missing data that you cannot confidently explain
- ☑ Consult with field partners to decide how to resolve unexpected values
- ☑ Removing sensitive or identifying data

5. Preparing variables for analysis: finally, the fun part!

- ☑ Group distinct answers into categories or ranges
- ☑ Split up or extract components from an answer
- ☑ Make new variables out of the variables you have according to the Analysis Plan and tabulation plans

1. Before you start: making order out of chaos

Document all changes to the data

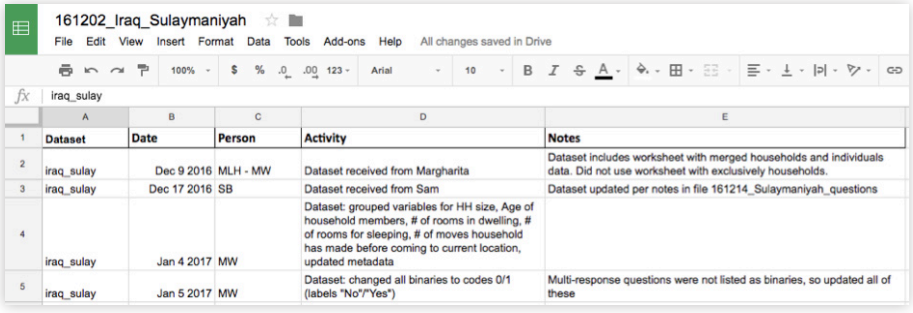
Whether you are using Excel or other statistical software, it is imperative that you record all changes you make to the dataset and that you describe how you made those changes. If you're working in a group and multiple people make changes to the dataset, the record of changes could also specify who made each change. This allows you and others to review and replicate the work.

Keeping track of changes made to datasets will be much easier if, instead of manipulating the datasets manually, they are altered automatically by scripts. This allows one to: combine processing and documentation; easily find and correct mistakes; save time on repetitive tasks; and avoid duplicating work if new or updated datasets are provided. Though there is no easy way of keeping track of changes in Excel, SPSS, Stata and other statistical software do have this option (for instance programming the commands for processing in a syntax file in SPSS or a "do file" in Stata), which makes them the gold standard in data processing.

tip

If not using SPSS or Stata, one option is to make a new column next to the column of the variable you make changes to, call it "flag" or something similar. Whenever you make changes to the main variable on a case, you also give the flag variable of that case a value, say 1. This can be done in Excel, too, so it is not something you need a statistical software for.

A workaround for those that do not use statistical software is to keep a manual log. Here is an example using Google Sheets for a profiling exercise in Sulaymaniyah, Iraq (2016):



161202_Iraq_Sulaymaniyah

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

iraq_sulay

	A	B	C	D	E
1	Dataset	Date	Person	Activity	Notes
2	iraq_sulay	Dec 9 2016	MLH - MW	Dataset received from Margharita	Dataset includes worksheet with merged households and individuals data. Did not use worksheet with exclusively households.
3	iraq_sulay	Dec 17 2016	SB	Dataset received from Sam	Dataset updated per notes in file 161214_Sulaymaniyah_questions
4	iraq_sulay	Jan 4 2017	MW	Dataset: grouped variables for HH size, Age of household members, # of rooms in dwelling, # of rooms for sleeping, # of moves household has made before coming to current location, updated metadata	
5	iraq_sulay	Jan 5 2017	MW	Dataset: changed all binaries to codes 0/1 (labels "No"/"Yes")	Multi-response questions were not listed as binaries, so updated all of these

Folder structures and naming conventions

Establish and maintain a filing and naming system to keep different versions of datasets distinct from one another.

"Version chaos" is a common problem plaguing data analysts, who often must store several copies of a similar dataset. The following points explain good folder structuring and file naming practices, which can help keep versions under control.

Folder Structure

- Make an 'Original ('raw') File' to keep original files;
- Make a 'Working Files' folder that contains (1) the latest files and (2) a folder named '_old' to keep previous working files. Separating the latest working file from older versions in different folders can help to distinguish versions.

File Naming

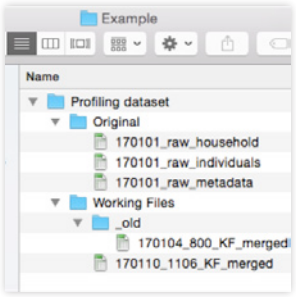
- File names might contain the following pieces of information: date (in yymmdd format), time (in hhmm format), the initials of the person who processed the file, and a substantive description of its contents.

ex.

A working file may be stored and named as follows:

~/Working Files/_old/170104_800_KF_merged.xlsx

This indicates that the file was processed on January 4, 2017, at 8:00, by a person with the initials KF, and that it contains data merging households and individuals.



Having all data on hand

Make sure you have all necessary datasets and that each dataset contains all the variables and observations that it is expected to contain.

To do this, check the content of your data file(s) against your methods and tools (e.g. questionnaire, sampling design, summaries from implementing partners in the field with the totals for households visited, etc.). Do you have all the raw data you need to head towards analysis?

ex.

The survey used contained 50 household-level variables, and 14 additional variables were entered by the enumerator – making 64 variables in all. However, the dataset pertaining to this survey contains only 55 variables.

This could be the result of a colleague accidentally leaving out a portion of the survey that contained nine variables.

ex.

The survey sampled 1,889 households and 4,367 individuals. Check to make sure that the data file indeed has 1,889 household and 4,367 individual observations.

If not, it's possible that a set of observations from a certain village or camp is missing.

2. Preparing the dataset: it's got to look good

What is a good-looking dataset? It is ***clean***, as in it does not have any special formatting and there is a clear hierarchy between variables. This means it also generally looks a bit boring to the untrained eye. Some people prefer to add colours and headings to show which variables belonged to, for example, the specific modules of the questionnaire, but generally speaking it is easiest to run analysis on the dataset if it does not have this. One way of organising to one person may also not make as much sense to a profiling partner also working on the same dataset.

Making a dataset readable by clarifying labels

It is always useful to have variables that clearly reference the question in the questionnaire. This means giving variables a short name, for instance the number of the question from the questionnaire, and then having a separate metadata worksheet that explains what the variable represents.

ex. Extract from a profiling dataset from Erbil, Iraq (2016)

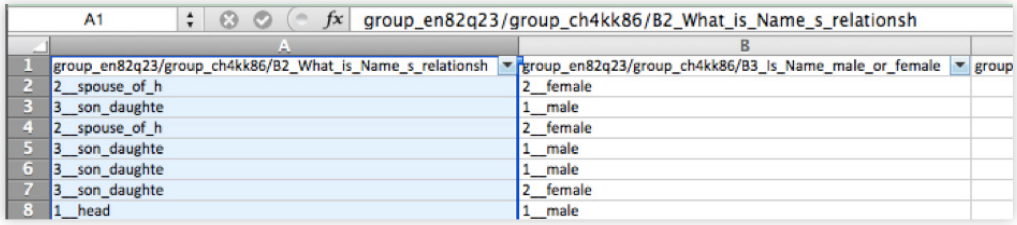
	A	B	C	D	E	F	G	H	I	J
1	group	strata	HH_SIZE	HH_SIZE_G	HEAD_SEX	HEAD_AGE	HEAD_AGE_(	B3	B4	B4_1
2	Refugee	Baharka, Bna	6	5 to 8 meml	Male	35	31 - 40	Male	35	31 - 40
3	Refugee	Baharka, Bna	6	5 to 8 meml	Male	35	31 - 40	Female	30	21 - 30
4	Refugee	Baharka, Bna	6	5 to 8 meml	Male	35	31 - 40	Female	7	0 - 12
5	Refugee	Baharka, Bna	6	5 to 8 meml	Male	35	31 - 40	Male	10	0 - 12
6	Refugee	Baharka, Bna	6	5 to 8 meml	Male	35	31 - 40	Female	14	13 - 20
7	Refugee	Baharka, Bna	6	5 to 8 meml	Male	35	31 - 40	Male	15	13 - 20
8	Refugee	Baharka, Bna	5	5 to 8 meml	Male	30	21 - 30	Male	30	21 - 30

ex. Extract from the associated metadata document, which gives information on each variable:

	A	B	C	D	E	F
1	Dataset	Name	Label from questionnaire	New label	DataType	Level
2	iraq_erbil	group	group	Population Group	text	household
3	iraq_erbil	strata	strata	Geographical Strata	text	household
4	iraq_erbil	HH_SIZE	HH_SIZE	Size of household	numeric	household
5	iraq_erbil	HH_SIZE_G	HH_SIZE	Size of household (grouped)	text	household
6	iraq_erbil	HEAD_SEX	HEAD_SEX	Sex of household head	text	household
7	iraq_erbil	HEAD_AGE	HEAD_AGE	Age of household head	numeric	household
8	iraq_erbil	HEAD_AGE_GROUP		Age group of household head	text	household
9	iraq_erbil	B3	Is [Name] male or female?	Sex of household members	text	individual
10	iraq_erbil	B4	How old is [Name]?	Age of household members	numeric	individual
11	iraq_erbil	B4_1		Age group of household members (grouped)	text	individual

When a dataset is downloaded from mobile data collection software such as KoBo Toolbox, the variable and even the answer options may have odd names that are automatically created. If using KoBo Toolbox, there are now ways to make sure the variables and answer options retain their original labels when downloading. If this is not possible, there are various ways that these variables and answer options can be renamed in Excel depending on what changes are needed by using commands such as LEFT or RIGHT. In SPSS, this requires the use of syntax to program the software to replace all of the new labels with the labels of your choice.

ex. Result after downloading Erbil, Iraq (2016) profiling dataset from KoBo



A1 | fx | group_en82q23/group_ch4kk86/B2_What_is_Name_s_relationsh

	A	B	
1	group_en82q23/group_ch4kk86/B2_What_is_Name_s_relationsh	group_en82q23/group_ch4kk86/B3_Is_Name_male_or_female	group
2	2__spouse_of_h	2__female	
3	3__son_daughte	1__male	
4	2__spouse_of_h	2__female	
5	3__son_daughte	1__male	
6	3__son_daughte	1__male	
7	3__son_daughte	2__female	
8	1__head	1__male	

Merging datasets

Sometimes different units of measurement (ex. community, household, individual) may be stored in separate files or worksheets if using Excel. However, to analyse the data collected in a unified manner, it is necessary to merge the files.

ex.

You have a file of data at the household-level and another file containing data on the individuals that constitute those households. There should be unique identifying codes that match households and individuals.

Using spreadsheet or statistical software, you will need to systematically combine the files so that household data is associated with the correct individuals and different individuals are correctly grouped into households.

Creating tables

In order to get to know the dataset and to start identifying possible errors, a good first step is to create a series of tables of each variable included. These tables can just show the frequency of each response, in other words how many respondents answered each question. This can be shown as the number itself or as the percentage of the total. This can already show whether there were any unexpected missing information, and this is a good way to start detecting outliers (more on outliers in section 4).

ex.

A table showing the frequency distribution of the variable HH_SIZE, representing the number of members in the household, from Erbil, Iraq (2016).

Note that this includes both the percentage of the total sample and the frequency (the number or count). This is a default format - no fancy formatting needed at this stage.

No. of household members	Percentage	Frequency
1	1 %	48
2	3 %	208
3	8 %	480
4	12 %	732
5	18 %	1055
6	18 %	1056
7	13 %	770
8	10 %	608
9	5 %	288

3. Cleaning common issues and errors: the easy stuff

Check if respondents answered all the questions intended for them to answer

In other words, make sure that the "skip logic" was followed. Skip logic or skip patterns is not a technical term, but refers to the pattern used in the questionnaire to require certain respondents to only answer the questions that are relevant for them and to skip the others that are not. For instance, did a person who was never displaced answer displacement-related questions? Were employment-related answers provided for a toddler? Respondents answering questions that were not intended for them to answer is a more common issue for paper data collection, but it still comes up from time to time as a result of mobile data collection as well.

To check that this is not the case, you need to know the questionnaire well. Then, by applying a filter to certain variables in Excel, or writing conditional arguments in statistical software, this can reveal if skip logic was followed consistently.

ex.

A respondent answered that she did not work in the last 30 days, which should mean that questions regarding employment should be skipped for this respondent. However, data shows that she reported an income from a job from that period.

The way to resolve this may be to delete the reported income if not sure where the error was and leave this cell blank. If it appears, however, that this respondent answered all other questions about employment with reasonable answers matching that reported income, then it is likely preferable to revisit the answer to the initial question, maybe even changing this to yes, she did work in the last 30 days. This is because there is a convincing argument that the error happened there and it was actually appropriate that she did not skip the next few questions.

Review open-ended answers and categorise them

Make sure that you have coded open-ended responses. For example, if the questionnaire included questions with the answer option "If other, please specify", these responses need to be reviewed, and if any of the options listed correspond with another existing category then this option should be recoded to that existing category. If enough "other" options are similar, then the partners can consider creating a new category to include in the analysis.

For more complex open-ended responses, you will need to construct a coding scheme and agree upon the scheme with partners in advance. To limit the risk of two people coding the same response in different ways, keep your coding scheme as detailed as possible.

ex.

Respondents were asked an open-ended question about the main reason they were displaced. Because this is a sensitive question, this was left open-ended to allow for a broad range of possibilities. The partners developed a coding scheme that corresponded with the penal code for that region so that the responses could be given an appropriate code for that context.

For most responses this worked fine, but not for all. It turned out that the coding scheme was not as clear cut as initially expected, since respondents often gave multiple intersecting reasons including some reasons that were outside the scope of the penal code.

The partners therefore decided to consult with displacement experts working on protection issues to clarify whether a specific response did or did not constitute certain vague concepts such as a "direct threat of violence", or a "usurpation of housing". Though this required time to discuss and review individual answers, the partners felt strongly that this was important to prioritise, as having these responses properly categorised was central to the analysis of the displacement situation.

Resolve typos in answers

Make sure that every categorical value is spelled consistently, and if you are using codes (ex. "1" for yes; "5" for Camp A; etc.) they are being consistently applied to the answers.

ex.

The city of Sulaymaniyah has several different names and English spellings, including Sulaimaniya, Sulaymaniyah, and Slemani. Make sure your dataset sticks to one spelling variant throughout. Using programmed digital surveys typically means you will not see this issue, but you may encounter it if you have an "If other, specify" option. In this case, because it is a geographic variable, if it is relevant in your setting, you may want to think about using P-Code (place codes).

4. Checking for unexpected values: any weird results?

Unexpected results, including outliers and oddities

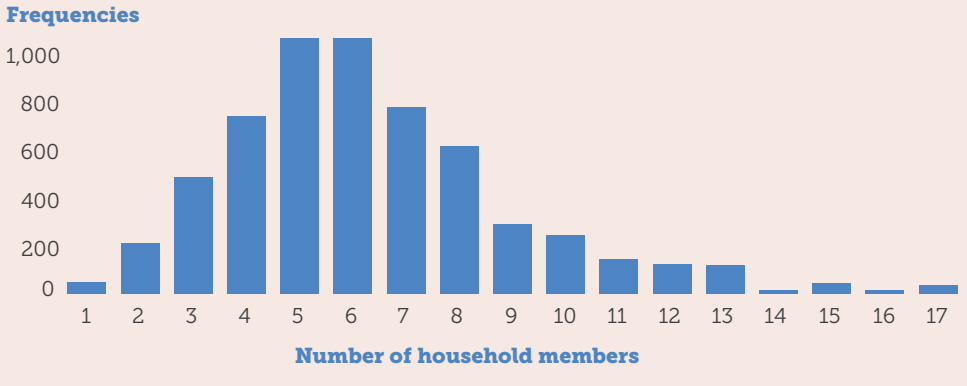
Make sure that the ranges and averages of your variables correspond to what you know about the context, or to what you can infer from social science reasoning. An outlier is an observation that is very different from the other observations in a dataset. Though this can occur by chance, it can also be an indication of an unintended error somewhere in the data collection process.

This can be checked by producing descriptive statistics, which are the basic features of the data, in this case of a variable applied to the observations in your sample, such as the total number, the average, the higher end of the range and the lower end of the range. After reviewing the basic frequency tables already created, you can take a closer look at the descriptive statistics. If certain variables have ranges or averages that appear unlikely, flag these to address with your team, and check whether this may be the result of errors during the data collection or tool creation process.

ex.

A graph showing the frequency distribution of the variable HH_SIZE, representing the number of members in the household, from Erbil, Iraq (2016).

Note that this does not show concerning outliers, as a household with 17 members is, though rare, still within the scope of what is expected for the context.



ex.

Descriptive statistics for Variable B4_A ('q26,'Do you own any of the following assets: Cars'), showed that nearly 80% of respondents owned cars, although you know that this area has very few car owners given the poor quality of the roads.

ex.

The average value of Variable C3 ('q23, Household expenditure estimate for last 30 days, in USD') is 12.34 USD, but the range of the variable runs from 0 to 77,234.01 USD. The upper bound may signify an outlier that should be looked into.

ex.

Making a bar chart for Variable C9_B ('q45' 'Highest-level education attained') you visually identify that the answer 'Above University' had the highest bar followed by 'Elementary School'.

This cannot be right given your knowledge that most people in this context have not attended university and that most have at least some high school education.

Flag this as a variable for you to look into to see if mistakes occured during data collection, when the data was entered, or in earlier steps of processing.

Contradictions in respondents' answers to different questions

Check for logical inconsistencies between variables. Sometimes, we might come across contradictory data on the same persons or households.

Using your analysis framework or questionnaire, identify variables that could have mutually exclusive, or contradictory, values. Then, look through your dataset for persons or households that have contradictory values on these variables. If contradictory data is found, consult with your team about how to correct them.

ex.

In a household roster, an individual's age is recorded as '0-5 years,' but this individual's education level is listed as '4-year university or more.'

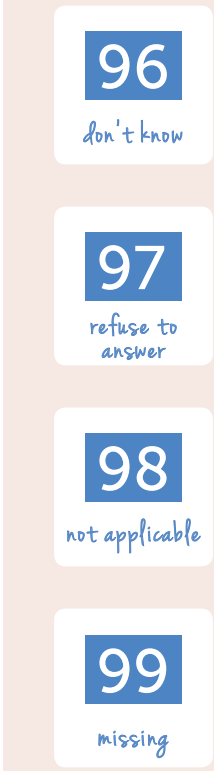
ex.

A household reported that it had never been displaced, but provided an answer for 'date of initial displacement.'

Check for missing data that you cannot confidently explain

Make sure that you can explain and account for every missing, empty, or null value. There are three main reasons why a variable may be blank.

- **True Missing** – This refers to cases in which a respondent should have given an answer, but for whatever reason no answer is recorded. This can occur during data collection, data entry, or data processing. In contrast to meaningful omissions, we do not know what to make of these 'True Missing' values. Too many of these and it makes the results really hard to interpret. These results should be identified using a unique 'missing data code', which is typically coded as 99 to avoid this getting confused with the codes for any other response options.
- **Processing Error** – This could occur when the respondent did provide a response, but for some reason the response is not present in the data file, meaning that there was a mistake during the entry of the data (if using paper and pen). In rare cases the originalquestionnaire can be found to resolve this,



but usually not, making this issue indistinguishable from a True Missing. These results should also be coded as 99 to avoid this getting confused with the codes for any other response options.

- **Not Applicable** - This occurs when the question is simply not relevant for the respondent. Normally skip logic is used in the design of the questionnaire to avoid this, but in some cases this is not done or was done incorrectly. If a question is not applicable to a respondent but they are still asked, it is critical to have "N/A" as an answer option, which is typically coded as 98 to avoid this getting confused with the codes for any other response options. If the questionnaire has a skip logic and all rules were followed, then these would appear as blank cells in the dataset, or if using SPSS the cells could have a small dot in them.

Of course it is always important to include two other options for every question: respondent does not know the answer and respondent refuses to answer. In most cases, this is grouped together as "DK/RA" and is coded as 97 because it takes work for people to decide which to list. In some cases it is helpful to distinguish (for instance this could be useful in a pilot) meaning that RA could be 97 and DK would be 96.

Removing sensitive information

While the topic of sensitive data should have been discussed earlier in the profiling process and decisions may have been made about whether or not to collect certain kinds of information, you may find that other variables you collected or derived variables you produced can highlight similarly sensitive information. There should be a deliberate and agreed-upon decision as to whether to keep the variable, whether to share that data, and if so with whom it should be shared with.

This should be done at the end of the data processing phase and may result in having two versions of the dataset, one for sharing within a small group and another for sharing more broadly.

Anonymising the data according to a data sharing protocol agreed in the Profiling Working Group. Even anonymous data may contain identifiable information or sensitive information. Before sharing the dataset with others, especially outside the profiling partners, you may need to:

- Remove columns of data that is not suitable for sharing with all partners (i.e. names)
- Change precise figures, such as small numbers, to ranges (i.e. "3" might be changed to "<5")

Reducing Geographical Precision

When GPS coordinates of are collected alongside sensitive data consider removing the precision of the GPS points by several decimal places. Mobile data collection through tablets will collect GPS points to 6-7 decimal places identifying households, and if the data is sensitive we should remove this precision to 2-3 decimal places so not to geo-locate the household, community centre, school, etc.

5. Preparing variables for analysis: finally, the fun stuff!

Group distinct answers into categories or ranges

Categorise respondents' answers into categories that are more relevant or easier to interpret.

When the data was collected, the survey may have asked the respondent to provide a specific number in a range, a specific location, or a specific type of profession. While it can be useful to have more specific data, it can also be unnecessary or make it harder to interpret the information you collected. It may be necessary to group some of these answers into more meaningful categories.

ex.

A common category is age groups. Your questionnaire gathered household members' ages in years. However, for your context you are more interested in or only really need to know about ranges of age not specific number of years they have been alive.

Does your research objectives really require you to know if a respondent was 15 as opposed to 14? Or do you need to know, for instance, if a respondent was a minor, working-age, school-age, or other meaningful range of ages?

ex.

This can also be the case for geographic locations. For instance, a survey collected what town, camp, or settlement the respondent lived in.

Let's say that two of the camps, Camp A and Camp B, are technically distinct camps, but because they are so close together (i.e. separated by a tiny road) they share many similar conditions. For the sake of this profiling exercise, it may be helpful to combine Camp A and B into Camp Z.

Split or extract components from an answer

Sometimes answers are collected in a format that is difficult or impossible to run calculations. It may be necessary to split parts of an answer into different variables or cut out parts of an answer.

ex.

Your survey collected information on the date that the respondent arrived in the current location. The format of collection was MM/DD/YYYY. Let's say that because of the characteristics of the situation and the objectives of the exercises, only the month and year are required.

First, you would need to systematically remove the "DD" from every single answer. Second, you would need to separate the date variable into two variables: "Month" and "Year."

Prepare all other variables for analysis

Data processing is not complete until key variables are created for the analysis. Common variables created here include the ratio of populations outside the working age to the working age population (dependency ratio), whether a dwelling fits within local standards for adequate living conditions, the situation of overcrowding, a simple binary variable for whether a person or a household has been displaced, etc. These new variables are also called derived variables because they need to be created based on other variables or a combination of variables in the dataset. Good thing you took the time to plan this out in your analysis plan and tabulation plan!

Where can I look for more information?

Assistance with Excel:

- Exceljet - great repository of training videos and common functions and formulas
- Microsoft's Excel Help Center - useful explanations for basic functions from the makers of Excel
- ExtendOffice - some additional tips and tricks with diagrams and screenshots
- Random tips documents resulting from a google search such as this one can sometimes also help!

Assistance with SPSS and R:

- SPSS Tutorials - excellent repository of advice from the basics to more advanced commands using syntax
- Stats Make Me Cry - informal and fun companion for all those processing woes using R or SPSS, including videos
- Random tips documents resulting from a google search such as this one can sometimes also help!