# End User Documentation – MSNA Analysis Script

*Draft: 22 August 2018*

Contents

[End User Documentation – MSNA Analysis Script 1](#_Toc522817227)

[Brief Description – Content and Purpose of This Document 2](#_Toc522817228)

[Toolkit Overview 2](#_Toc522817229)

[Toolkit Capability and Limitations 3](#_Toc522817230)

[Step 0: Finalizing all prerequisite steps 3](#_Toc522817231)

[Data 3](#_Toc522817232)

[Step 1: Downloading and setting up the toolkit 3](#_Toc522817233)

[Downloading the Toolkit 3](#_Toc522817234)

[Install R and R Studio 4](#_Toc522817235)

[Step 2: Preparing your inputs 4](#_Toc522817236)

[Prepare your analysis input file 4](#_Toc522817237)

[Read me tab 5](#_Toc522817238)

[Data tab 5](#_Toc522817239)

[Kobo questions tab 5](#_Toc522817240)

[Kobo choices tab 5](#_Toc522817241)

[Choice ranks tab 5](#_Toc522817242)

[Parameters tab 5](#_Toc522817243)

[Stratification\_samplingframe tab 6](#_Toc522817244)

[Cluster\_samplingframe tab 6](#_Toc522817245)

[Composite\_indicators tab 6](#_Toc522817246)

[Analysis plan tab 9](#_Toc522817247)

[Finalizing analysis input file 9](#_Toc522817248)

[Step 3: Running Your R Script 10](#_Toc522817249)

[Step 4: Using Your Outputs 10](#_Toc522817250)

[Review of Outputs 11](#_Toc522817251)

[Using the Disaggregations Master Table 11](#_Toc522817252)

## Brief Description – Content and Purpose of This Document

### Toolkit Overview

If you are reading this, then you have been tasked with analyzing complex survey data for a Multi-Sectoral Needs Assessment (MSNA) or other assessment. REACH has developed an R-based toolkit to facilitate this analysis, and this handy guide will walk you the user through the steps needed to setup, run and use this toolkit.

The toolkit is designed for ease of use. The user only needs to prepare their inputs and describe their key composite and disaggregation indicators, and the program will automatically generate the desired results ready for joint analysis workshops and report writing. Key inputs needed include:

* Cleaned dataset in Kobo format (using xml names and headers)
* Survey tool xls form, choices list, and choices ranking (if applicable)
* Survey design indicators (strata, clusters and their names)
* Sampling information for population weighting purposes
* Composite indicator definitions with desired thresholds
* Analysis plan

The toolkit is comprised of a script in R and a toolbox in ARC, with most of the functions running behind the scenes (some files listed in ‘internal’ folder, which you do not need to pay attention to). Outputs from this script will include:

* Tables
* Charts
* Maps
* Heatmap

There are five overarching steps in the toolkit’s use that will be explained throughout this guide:

Step 0: Finalize all prerequisite steps

Step 1: Downloading and setting up the toolkit

Step 2: Preparing your inputs

Step 3: Running the R Script

Step 4: Using Your Outputs

### Toolkit Capability and Limitations

This toolkit will take a lot of the repetitive work out of preparing your survey results, but there are limitations to what the user should expect.

|  |  |
| --- | --- |
| Toolkit capabilities | Toolkit limitations |
| * Produces composite indicators from weighted sums of variables recoded to numerical values * Produce summary statistics, confidence intervals and hypothesis tests for numerical and categorical variables for simple random, stratified and cluster sampling. Weights are calculated by the tool (complete sampling frames required) * Repetition for multiple subsets based on a variable * Disaggregation by one variable (+1 variable for repetition) * Produces raw numbers prepared for indesign datamerge, barcharts with error bars for factsheets and reports and heatmaps | * This program will not clean your data for you. Inputs must be cleaned beforehand, and the quality of the output depends on the quality of the inputs. * This program will not define your indicators and composite indicators for you. You must correctly construct the composite indicators. What you ask from it is what you will get * You will need to provide complete inputs including the data in the right format, the kobo tool and detailed sampling frames matching the dataset |

## Step 0: Finalizing all prerequisite steps

### Data

* Before using the tool, you must have all data cleaned and in Kobo format
  + Data must have xml names and headers (rather than label names) – select option on Kobo
  + Data must not have group names attached to headers – deselect option on Kobo
  + For select multiple questions: the first column must show all selected options listed, then subsequent columns (each listing an individual option – usually titled as [question name. variable name] should indicate if each individual option was selected (with TRUE if no, FALSE if yes)
  + Column names must match exactly to original kobo question names
  + All choices in the dataset must be listed in the kobo questionnaire/choices list (new choices created during data cleaning must be added)
* You must have your sampling frames prepared and be sure that stratum/cluster names match the corresponding values in the data exactly
* You must have your indicators and composite indicators defined (likely in coordination with the clusters/sectors during the joint design/analysis process).

## Step 1: Downloading and setting up the toolkit

### Downloading the Toolkit

* Request the most recent version of the toolkit from Martin Barner ([martin.barner@impact-initiatives.org](mailto:martin.barner@impact-initiatives.org)) or Eliora Henzler ([eliora.henzler@impact-initiatives.org](mailto:eliora.henzler@impact-initiatives.org)).
* Create a folder on your computer titled “XXX\_MSNA\_Analysis”. XXX should be the three letters for your country (e.g UGA = Uganda, NGA = Nigeria). Where this folder is located on your computer does not matter.
* The toolkit will be in a self-contained zip file. Unzip the contents of the zip file into the folder you created. The toolkit will not work if you try to work off of a not properly unzipped folder.
* The zip file will contain the following content:
  + input folder: everything that you need to provide as input goes here
  + internal folder: ignore the contents of this folder, do not edit any files
  + output folder: this is where all outputs will be saved
  + Manual.docx: the file you are currently reading, right now!
  + run.R file: The file with the R code that you will need to execute.
  + .gitignore, LICENSE, README files: you can ignore these files

### Install R and R Studio

In order to run this script you need to download and install the following:

R: For Windows, go to <https://cran.r-project.org/bin/windows/base/> and click on the link to download the latest version of R.

R Studio for Windows, from <https://www.rstudio.com/products/rstudio/download/#download>

## Step 2: Preparing your inputs

### Prepare your analysis input file

* The analysis input file encompasses all of the information needed for R to produce your analysis. The xls file consists of several components (in separate tabs):
  + **Read me:** This gives you instructions as how to fill out the analysis input file. When you are finished, you will need to click the “update analysis definition” button on this sheet as a final step before feeding into R.
  + **Data:** In this sheet, you will copy and paste your cleaned data set in Kobo format
  + **Kobo questions:** You will insert your Kobo tool in xls format in this sheet
  + **Kobo choices:** You will insert your Kobo choices from your tool in this sheet
  + **Choice ranks:** if you need your choices to show up in a particular order in your analysis outputs, you designate the order of your choices in this tab
  + **Parameters:** In this sheet, you identify the sampling strategy used, cluster variables, and which labels should be used in your outputs
  + **Stratification sampling frame**: if probability stratified or probability stratified cluster sampling was used, fill out this tab detailing strata variables, and the population figures used in the sampling frame
  + **Cluster sampling frame:** if probability cluster or probability stratified cluster sampling was used, fill out this tab detailing the first (and second/third if applicable) cluster variable, and the population figures used in the sampling frame
  + **Composite indicators:** this tab includes all of the information used to define composite indicators, including your PIN calculations and thresholds.
  + **Analysis Plan:** this tab details your analysis plan and which variables at what levels should be analysed by the scripts. You can also note which types of outputs (tables, charts, heatmaps, etc.) you need created for each variable analyzed.
* The following sections will cover the components of the analysis input file in detail so that it is clear what is required for each.

### Read me tab

* The read me tab gives an overview of the analysis input file and is self-explanatory
* **After filling out the analysis input file, you must return to the read me tab, click the “update analysis definition” button on the left**, which prepares the file for R. This should only be done as a final step when the analysis input file is completed.

### Data tab

* Refer to the prerequisites for the data listed above:
  + Data must have **xml names and headers** (rather than label names) – this is an option you can select when downloading the data from Kobo.
  + Data must **not have group names attached to headers** – deselect that option when downloading the data from Kobo.
  + **Select multiple questions must be formatted according to kobo standard**: the first column must show all selected options listed, then subsequent columns (each listing an individual option – usually titled as [question name. variable name] should indicate if each individual option was selected (with a 0 if no, 1 if yes)
  + **Column names must match exactly to original kobo question names**
  + **All choices appearing in the dataset must be listed in the kobo questionnaire/choices list**
* When you’ve finished cleaning data, make sure it is in the standard Kobo format
  + If you’ve changed or deleted columns during data cleaning, or **if your data does not fulfill any of the conditions above, please check with the Data Unit, as they may be able to help you convert it to the correct format.**
* **Paste your cleaned data set in standard Kobo format into the data tab in the analysis input file**

### Kobo questions tab

* Copy your tool directly from the xls form and paste it in the Kobo questions tab
* If you added new variables through defining composite indicators, be sure to add them to the kobo tool (both questions and choices tab), define the data type (integer, select\_one or select\_multiple) according to how you want them to be analysed and appear in your outputs. If you don’t add them, you can still analyse them with the tool; It will not label them and will try to guess from the data whether it is numeric or select\_one. The [[ALL]] option in the data analysis plan (see below) will only consider variables that appear in the kobo tool.

### Kobo choices tab

* Copy the choices in your tool directly from the xls form choices tab and paste it in the Kobo choices tab
* As mentioned above, if you added new variables through defining composite indicators, be sure to add them to the tool (both questions and choices tab), define the data type (integer, text, etc.), and label them as you want them to appear in your outputs.
* If you created new choices during your data cleaning (i.e. categorizing “other” responses as a specific option), be sure you add them to the choices tab
* If you use the kobo feature that loads choices from an external file, add those choices here as well.

### Choice ranks tab

* This tab only applies if you want ordinal variables to be interpreted in a certain order in your outputs. The bar charts will automatically be sorted from largest to smallest, so if you want a different order, indicate it here.
* For each choice requiring a specific order, copy the list\_name, name, and label from your choices tab, and indicate the rank (ordering) by adding integer numbers in the specified order in column E of the choice ranks tab.

### Parameters tab

* In this tab, you will outline the basic parameters to identify which variables are related to your survey methodology. You will identify the sampling strategy and cluster variables used, as well as the labels to be used in your outputs.
* Sampling strategy: select the sampling methodology used in your assessment
* Cluster variables: select the variable(s) used to identify which cluster the record belongs to. In most cases, cluster sampling is based on settlement, village, or another level of location. Choose which variables were used to determine the cluster. Note that the value must match exactly with the data column names.
* Questionnaire choices label column: In this column, put the exact header of the column (e.g. “label::English”, etc.) from your kobo choices list you want used as labels in your analysis outputs. If you want to create new label options to be used in your outputs, you can create another column in your kobo choices tab to identify the desired labels to be used.

### Stratification\_samplingframe tab

* You need to fill out this tab if you used probability stratified or probability stratified cluster sampling. In this tab, you outline your strata variables and the population figures used to determine your sampling so that the script can apply the appropriate weighting in the analysis. The variables will almost always be demographic or geographical information, such as district, state, respondent type, etc.
* In columns B, C, and D, row 3, identify the strata variables used in your sampling frame. If less than three variables are needed to identify the strata, you can use only columns B or B and C. These values must match exactly column headers in the data. From row 4 down, list all of your choices (with exact spelling from your Kobo choices list and your dataset) and all unique combinations of the values that appear in the data. In column E, insert the population figures of each respective stratum.
* This information should be directly from your TOR or sampling frame, with some formatting adjustments. It is important to note that all spelling of your variables must be exactly as it appears in the data. Every combination (first strata variable, second strata variable) that appears in the dataset must be included here, or else the script will not include it in the analysis.

### Cluster\_samplingframe tab

* You need to fill out this tab if you used probability cluster or probability stratified cluster sampling. Depending on your method, you may need to weight by sampling probability. In this tab, you outline your cluster variables and the population figures used to determine your sampling so that the script can apply the appropriate weighting in the analysis. The cluster variables used are typically by settlement, but may be other levels of location
* In columns B (and C if a second cluster, and D if a third cluster variable was used), identify the cluster variables used in your sampling frame. In column E, insert the population figures used.
* This information should be directly from your TOR or sampling frame, with some formatting adjustments. It is important to note that all spelling of your variables must be exactly as it appears in the data. Every combination (first cluster variable, second cluster variable) that appears in the dataset must be included here, or else the script will not include it in the analysis.
* If your sample Is proportional to cluster population size (not just the planned, but the actual sample!), it is technically possible to not weight by cluster size. To skip the cluster level weighting but still adjust certainty estimates to cluster sampling, in the parameters sheet select the non-clustered sampling strategy that matches your case, but still list the relevant variables in the cluster variables column of the “paramters” sheet.

### Composite\_indicators tab

Most of you have developed composite indicators for your MSNA, likely for the PIN calculations. The composite indicators are the sum of the input variables after they have been weighted based on a set of conditions. Different variables may be grouped together, and as may be the case with PIN composite indicators, the grouping of variables must meet a certain threshold to be categorized a certain way.

For example, a composite indicator could be created to categorize a household as vulnerable. Four input variables are considered (single, female headed household; child headed household; presence of vulnerable member including disabled, chronically ill, or UASC; no working aged members). In your composite indicator sheet, you will specify that these input variables should be grouped together and considered as one to determine vulnerability. You will determine the specific conditions, values, and weighting for each of the variables, and what threshold should be met (i.e. how many of the individual input variables values must be met) for a household to be categorized as vulnerable.

In this tab, you will define all of the composite indicators, or indicators that are constructed from multiple questions. You will need to determine the following for each composite indicator:

* The variable names of all questions that will contribute to the composite indicator (input variables)
* The condition for the value (if the value should be equal, smaller equal, larger, etc. to what is listed)
* The values of each of those input variables which will contribute to the composite indicator
* The weight of each value that will contribute to the composite indicator
* Note: these attributes (variable, value, weight) will typically be developed in coordination with the clusters/sectors through the joint analysis process. This is where PIN definitions and thresholds come into play.

First, let’s understand the structure of the table you see under the **composite\_indicators** tab.

**value**

This column should contain the values of the input variable (exactly as appears in data), or the required “value” for a given condition.

**condition**

This column sets a condition (any, equal, less, greater, all, skipped, etc.) for the value in the next column. If setting multiple conditions for the same variable, this goes in order and is overwritten by the last fulfilled condition listed. More details below.

**weight**

This column indicates the weight applied to the composite variable, if this condition is met.

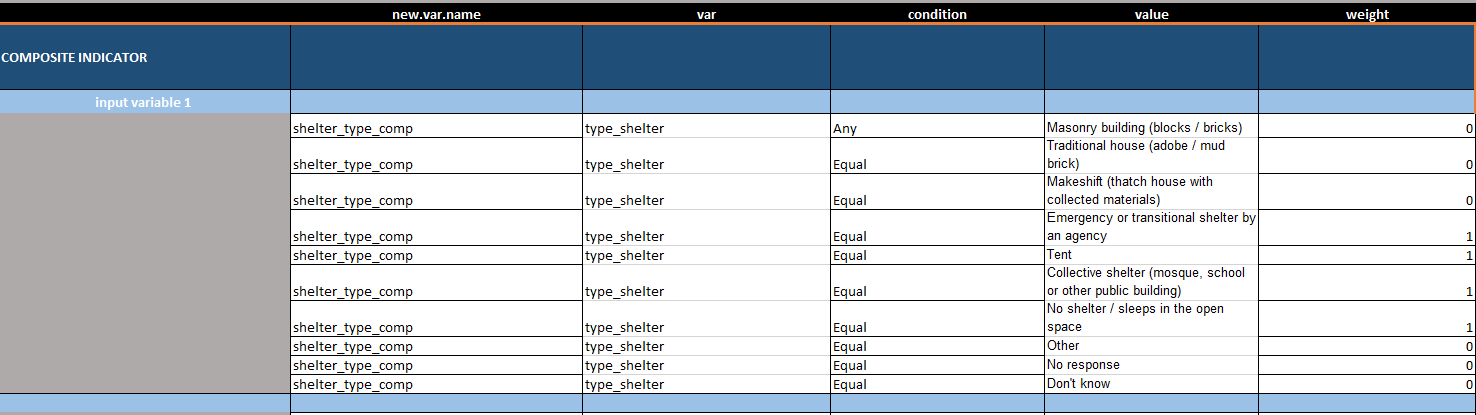
**New.var.name**

This column should contain a name for your new composite variable. Each white cell under each blue section should have the **EXACT SAME NAME.**

**var**

This column should contain the name the input variable (exactly as it appears in data) for this light blue section.

Each **DARK BLUE ROW** is for 1 composite indicator. No information will be included on this row, but all the rows underneath it up for will contain the information needed for 1 composite indicator.



Each **LIGHT BLUE ROW** is for 1 “input variable” that will be used to create that composite indicator.

Each **WHITE ROW** is for 1 condition of the “input variable” that will be used to create the composite variable.

More details:

* **Condition** options:
  + **Equal**: if the value is equal to what is listed in the “value” column, then the weight is applied
  + **Smaller equal**: if the value is smaller than or equal to the numeric threshold listed in the “value” column, then the weight is applied
  + **Larger**: if the value is larger than the numeric threshold listed in the “value” column, then the weight is applied
  + **Any**: If the value is any of the values listed in the “value” column, then the weight is applied (options separated by exactly and only a single space). Only for select\_one and select\_multiple questions.
  + **All**: If the value is all of the values listed in the “value” column, then the weight is applied (options separated by exactly a single space). Only makes sense for select\_multiple questions
  + **None**: If the value is anything but one or more of the listed values (options separated by exactly a single space) then the weight is applied
  + **Else**: If none of the other listed conditions are met, then this weight is applied. By default, if no condition is applied and no “else” is specified, a weight of 0 is used
  + **Skipped**: If the record has been skipped due to kobo tool skip logic, then the weight is applied
* **Value**:
  + This column should contain the numeric threshold for the input variable if the condition you’ve set is **equal** (and data type is numeric)**, smaller equal** or **larger**. The values may be a numeric threshold or the text value as it appears in the data.
  + This column should be empty if the condition you’ve set is **else** or **skipped**.
  + This column should list choices (separated by a single space, no comma etc.) if the condition you’ve set is **any**, **all** or **none**.
  + This column should name a single choice if the variable is a select\_one or select\_multiple, and the condition you’ve set is **equal**
* **Execution hierarchy of conditions**: If there is an overlap of conditions listed for the same variable, the tool will take the weight of the last listed condition that is fulfilled. There will only be one weight assigned per input variable: that of the last condition that is fulfilled.

For each composite indicator you want to define, you will need to **COPY and PASTE** a new block of rows to the bottom of the table, and redefine the composite indicator, input variables, values and weights.

You can make a composite indicator based on previous composite indicators but they are produced in order from top to bottom. For example, if you want to code food\_PIN based on food consumption score and self\_assessment you need to build those two composite indicator blocks **above** the food\_PIN.

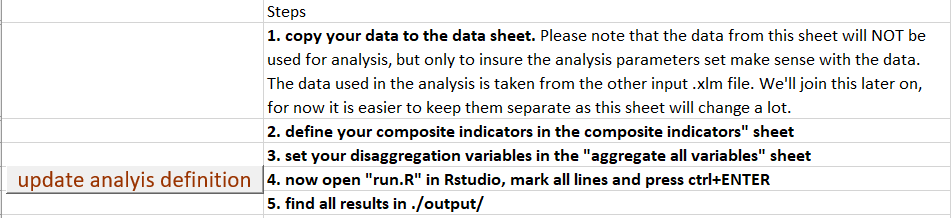
Note: If you have created new composite indicators in the composite\_indicator tab that should be included in your outputs, you must add the new variables (composite indicator) to the Kobo questionnaire and Kobo choices tabs. By default, the composite indicators are treated as numeric variables in the analysis and outputs. If you want them to be treated as categorical (select one) instead, you need to add them to the questionnaire, and if you want them to be added to the choices, you need to add them to both the questionnaire and the choices. To do this, copy the new.var.name into column B of the Kobo questions tab. Fill in the data type in column A and the label in column C. Fill in the rest of the information as you would for any Kobo questionnaire tool, and do the same in the Kobo choices list.

### Analysis plan tab

* This tab details your analysis plan and which variables at what levels should be analyzed by the scripts. You can also note which types of outputs (tables, charts, heatmaps, etc.) you need created for each variable analyzed.
* **Repeat for column**: choose the level you want the analysis independently repeated for (i.e. district). If you want standalone products, such as factsheets per district, you indicate here the level.
  + This cuts the dataset into pieces and does the whole analysis and outputs for each part of the data, independent from other parts
  + For example, you want the same analysis run for the listed variables (or all) for each district (separate from each other, not to compare)
* **Disaggregate by**: choose the variable you want to compare between different groups, regions, etc. within the dataset.
  + This is used within the analysis as the independent variable, so you can compare, run hypothesis tests between them, and view the results together in a chart.
  + For example, you want to compare variables by levels of disaggregation within the dataset (rather than independent from other parts)
* **Variable**: choose the variable you want analyzed
  + There is a possibility to run analyze all variables by writing [[ALL]]. Note: this only includes variables from the Kobo questionnaire that are select one, select multiple, or numerical types questions. Be aware that this usually produces thousands of outputs depending on the disaggregation. It is better to identify which specific indicators to analyze rather than including all in the analysis plan.
  + Note: If you want your composite indicators (PINs) shown in the outputs, you need to add the new variable (composite indicator) into your kobo questionnaire in the respective tab.

### Finalizing analysis input file

* Once you have inserted all of the required information in each of these components, you are almost done preparing your inputs – congratulations!
* After filling out the analysis input file, you must return to the read me tab, click the “update analysis definition” button on the left, which prepares the file for R. This should only be done as a final step when the analysis input file is completed. Be sure that the analysis input file is saved in the “input” folder, or else this will fail and R will not be able to find the file.

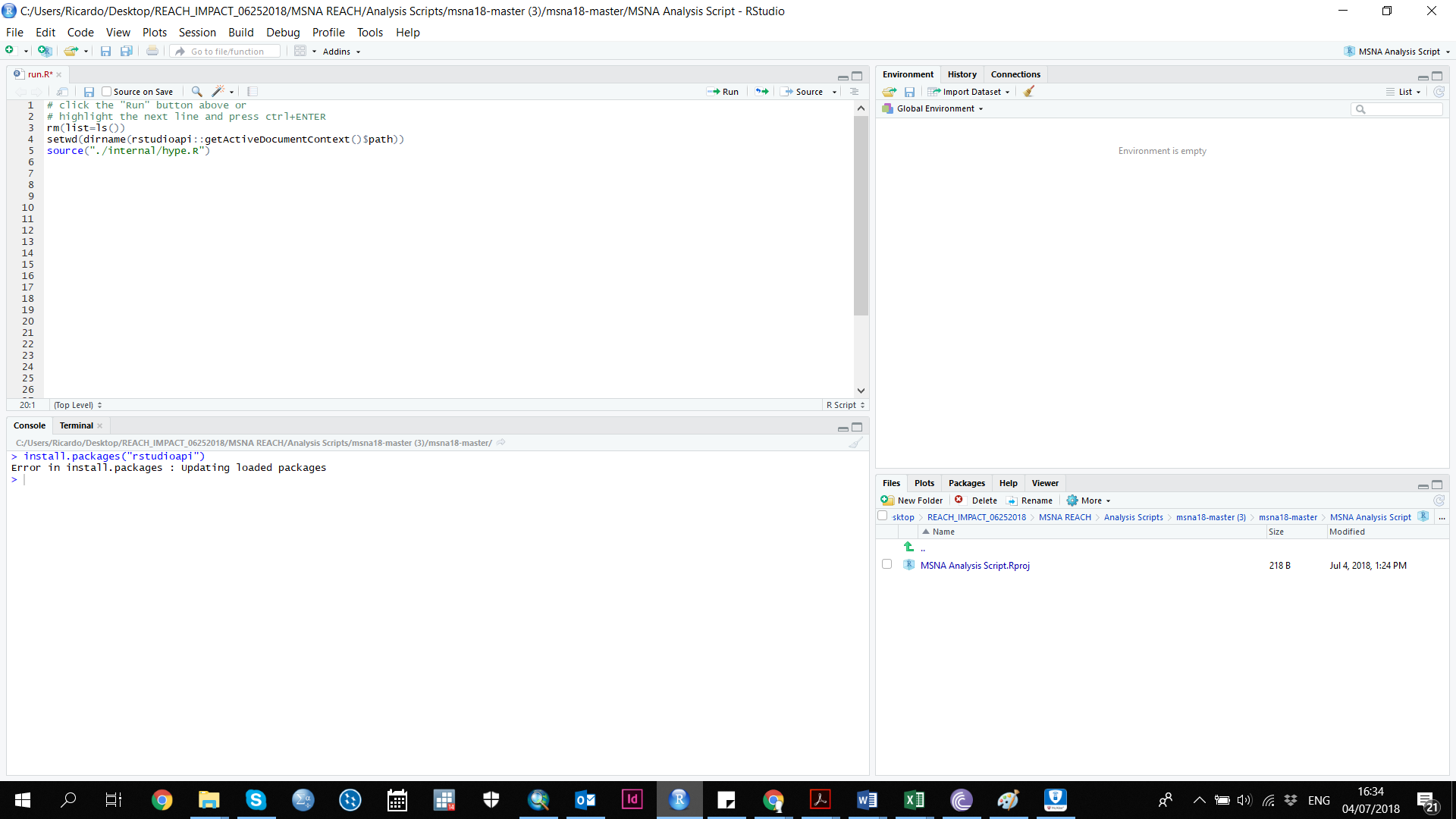


## Step 3: Running Your R Script

Running the R script is quick and easy as compared to all of the work you’ve done to prepare your analysis input components. Within the folder you created “XXX\_MSNA\_Analysis” which contains the toolkit, you will see a file titled run.R. This is the master script of code which will run the entire analysis script.

Open this file with R Studio (you must also have R installed on your computer, but it is not necessary to have R open at this step). You will see a few lines of code under a window titled “run.R”.

Select all the lines of code. In the top right of that window, click on the button “Run,” or with all text highlighted, press command enter.



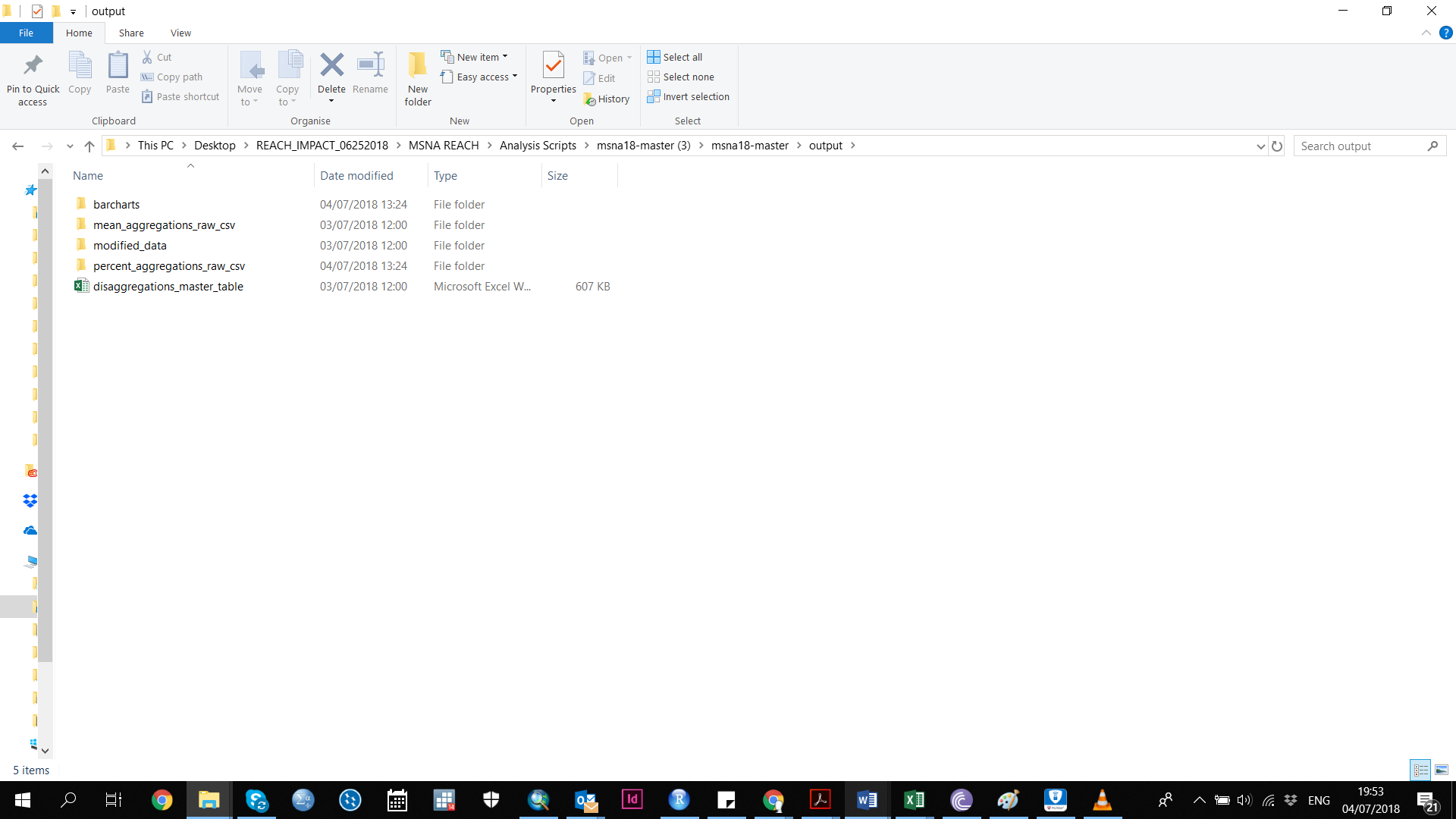
The first time you run the script in R studio, it will take longer than subsequent times. The program needs time to download all software and packages, so be patient. **For the first time running the scripts, make sure you are connected to the internet**.

Receiving an error

* If the script fails, you will most likely receive an error message that tells you exactly why it failed and what the issues are. There will also be information included on how to solve the error. The error is likely due to an issue with a component of your analysis input file.
* The script has been developed to anticipate errors and give you a message with useful information. If you aren’t familiar with R, the message may look like gibberish. Ignore the gibberish-looking part of the text, and there will likely be a few sentences explaining the issue.
* If there is an error that contains no clear information how to resolve it– congratulations, you’ve found a bug! Please send you input xlsm file to the data unit immediately and they will resolve the issue as quickly as possible.

## Step 4: Using Your Outputs

### Review of Outputs



After running the run.R script, your outputs should now appear in the XXX\_MSNA\_Analysis/output folder. Other folders will be created as well with intermediary files for other outputs only. Please find below a description of each of the different outputs:

Contents of output folder after run.R script is run

|  |  |
| --- | --- |
| **Output** | **Description** |
| Charts folder | Barcharts with confidence intervals and heatmaps as requested in the analysis plan. |
| Modified data | Your dataset returned with new columns added for the composite indicators. CSV format. |
| Disaggregations\_master\_table | Spreadsheet where you can view results of single indicators and disaggregation variables. XLSX format. |
| Datamerge | The datamerge file is produced for InDesign. This is directly linked to your analysis plan, so any analyzed variables needed for factsheets must be listed in the analysis plan. |
| Error files (optional) | If you’ve received an error message in R as to why running the script failed, the error might point to a file containing more details. You will find these in the output folder as well. |
| Log.txt file | This file has internal information about how things went when you ran the script. |

### Using the Disaggregations Master Table

The disaggregations master table will allow you to select a question/variable and then disaggregate the results by another indicator of your choice. Simply select your question/variable under “show variable” and select your disaggregation variable under “disaggregate by”. You can only select one disaggregation group, one repeat.for value, and one variable to display at a time. If you select more than one, you will see wrong numbers. Results for categorical variables are presented as proportions and numeric variables as means. Note: each time you open the file, make sure you go to the data tab, click refresh all, and link to the relevant cvs file (“output/master\_table\_long.csv”). Right click the pivot table and click refresh.

