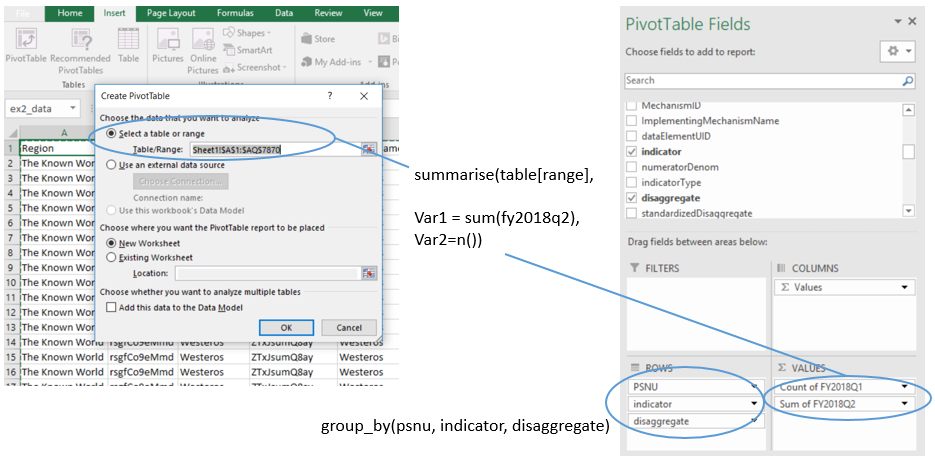
summarise is analogous to a Pivot Table



In Excel, open the training dataset “ex2\_data.xlsx”. Insert a pivot table. Change the PivotTable > Design > Layout > Report Layout = “Tabular Form,” Report Layout = “Repeat All Item Labels” and Subtotals = “Do not show subtotals.” Add different categorical variables to the “ROWS” field, such as PSNU. Add Count of region to the VALUES field.

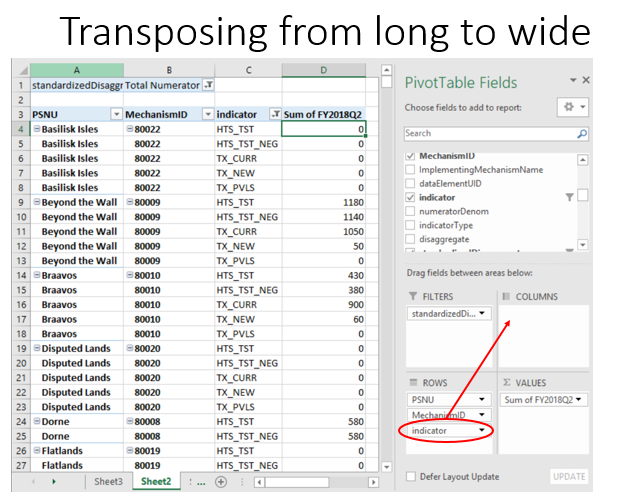
* What do the numbers mean in the “Count of region” column? Why do some PSNUs have higher numbers than others? Add indicator to the ROWS field. Why do some indicators have higher numbers than others?
* Add “fy2018q1” and “fy2018q2” to the VALUES field. Then change Count of fy2018q2 > Value field settings > summarise value field by > Sum. What happens?

Perform the analogous steps in RStudio by running lines 23-25 and lines 27-30 in the script “Session 3.R”.

* Based on the RStudio syntax: “function(input, arguments)” what is the function, what is the input, and what are the arguments we are using?

Use the filter and select commands to create a subset dataset that only contains three indicators: TX\_CURR, TX\_NEW, and HTS\_TST\_POS. Then filter for standardized disaggregate == “Total Numerator” and ismcad = N. See code lines 40-43 for help. Use code lines 45-47 to summarise\_at. You may alternatively try lines 49-52 to summarise\_all.

* Based on the RStudio syntax: “function(input, arguments)” what is the function, what is the input, and what are the arguments we are using?
* **Did it work?** Compare the results of the summarise\_at with the equivalent Excel pivot table. Are the results the same?



In Excel, open the training dataset. Create a pivot table with PSNU, MechanismID and indicator in the rows; and Sum of Fy2018Q2 in the values (see screenshot).

* Is this pivot data table wide or long?

Move a variable from the “ROWS” field to the “COLUMNS” field in your PivotTable fields panel. For example, move indicator from ROWS to COLUMNS (see screenshot).

* What happens to the indicators?
* This pivot data table is now wide with respect to \_\_\_\_\_\_\_\_\_\_\_ and long with respect to \_\_\_\_\_\_\_\_\_\_\_.
* Try moving different variables into the COLUMNS pivot field. How does the table behave?

Using your clinical cascade subset dataframe (not the whole raw training dataset), use the gather function to transpose the numeric values in the dataset into long format. See the R training session 3 script line 59.

* Based on the RStudio syntax: “function(input, arguments)” what is the function, what is the input, and what are the arguments we are using?

Transpose the long format dataset to be wide with respect to indicator using the tidyr “spread” function. See line 74.

* **Did it work?** Check the result of the R transpose in an Excel pivot table that has pivoted indicator into the COLUMN field. Make sure the same filters are applied in Excel for indicator and disaggregate.

Calculating Indicators

Small Group Exercise: what calculation?

* What sorts of calculations do we need to make on raw datasets to get the following calculated indicators?
  1. Cumulative totals
  2. Testing yields
  3. Proxy linkage
  4. Index tests done as a proportion of tests done among all modality types
  5. Treatment net new
  6. Silent transfers

Small Group Exercise 2:

* Using the native MER structured dataset, use Excel formulas (where possible) to calculate the following:
  + The S/APR cumulative totals for each indicator for each PSNU (fy2018q1 + fy2018q2)
  + Year-over-year treatment NET NEW comparing fy2018q2 to fy2017q2.
  + The HTS yield for fy2018q2.

Use mutate to create a proxy linkage variable that is equal to newly initiated patients on treatment divided by total HIV positive tests received. (Line 81 of code)

* Based on the RStudio syntax: “function(input, arguments)” what is the function, what is the input, and what are the arguments we are using?
* What happens if TX\_NEW = NA or HTS\_TST\_POS = 0? What can we do about that?
* Try using summarise to explore and describe the new linkage indicator (e.g., line 83-85 of code).

Merging Datasets

Import the text file got\_IMPATT.txt with tab as the delimiter. (See line 91 in the code).

* When considering implementation attributes such as subnational PLHIV or subnational HIV prevalence, at what level do these attributes exist? At what level should we therefore attempt to merge the datasets?

Use some basic exploratory commands such as nrow() or unique() to look at the PSNUs between the different data frames – see line 96-99 in the code.

* Why are the PSNUs different between the two datasets?

Try joining the datasets using left\_join with the got\_calcwide dataframe and the got\_IMPATT dataframe. (line 103 in the code)

* Based on the RStudio syntax: “function(input, arguments)” what is the function, what is the input, and what are the arguments we are using?

Use anti\_join to look at records that do not successfully merge based on your joining variable (lines 108-109 in the code).

* What could we do to improve the merge?