

Sankey Diagrams

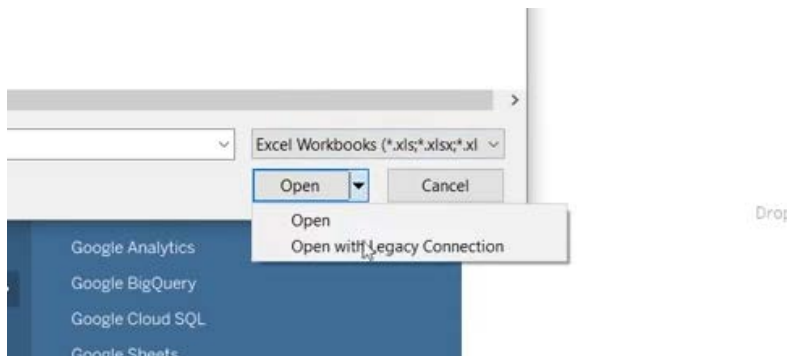
Based on a Super Data Science video, here:

<https://www.youtube.com/watch?v=1HwCzIA9hI4>

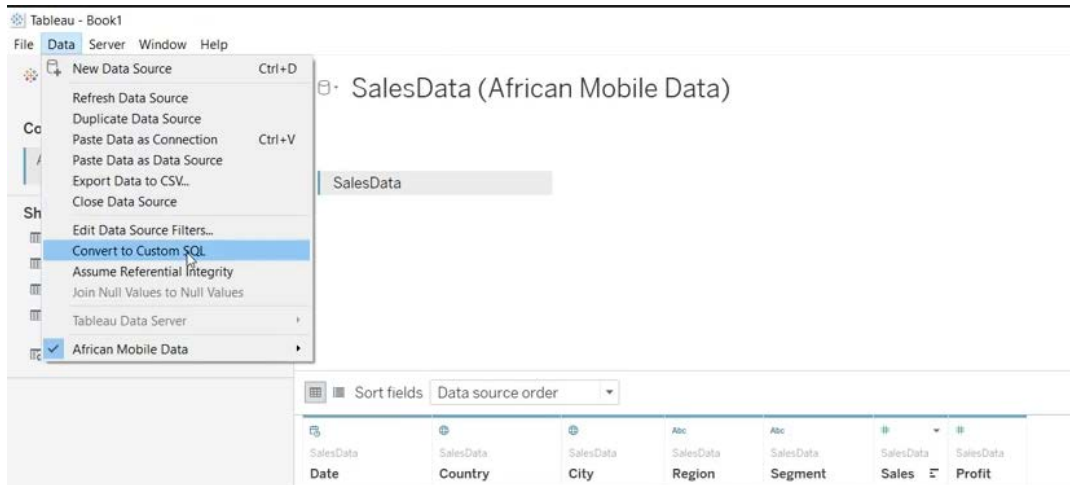
This document assumes that you are making a Sankey diagram that shows the flow of [Numbers] from categories in the [Start] variable to categories in the [End] Variable.

Getting and Formatting Data

- Find the data that you want to use in Tableau as you normally would. When opening the data, however, instead of clicking “Open”, click the arrow next to it and select “Open with Legacy Connection”.



- When you have selected your data and selected it as your data source, click on the Data tab and select “Convert to Custom SQL”.



- In the custom SQL page, make a copy of the SQL code that exists. Use UNION ALL and paste a copy of the SQL code to make a second copy of your data in your data. Add a column called VizSide for both copies of the data. One will be “Side 1” the other “Side 2” or something similar.

Creating the Sankey Part of the Sankey Diagram

- Create a new variable [ToPad]
if [VizSide] = "Side 1" then 1 else 49 end
- Right click on [ToPad] and click "create bins". Make the bin size = 1 and call the bins [Padded].
- Create a calculated field [t]
 $(INDEX()-25)/4$
- Move [t] to the Columns and [Padded] to Detail. Right click [t] and "Compute Using" [Padded]. Change Marks to Circles.
- Create a calculated field [Rank 1]
 $RUNNING_SUM(SUM([Numbers]))/TOTAL(SUM([Numbers]))$
- Duplicate [Rank 1], call it [Rank 2] and leave it entirely the same.
- Create a calculated field called [Sigmoid]
 $1/(1+EXP(1)^{-[t]})$
- Create a calculated field called [Curve]
 $[Rank\ 1] + (([Rank\ 2] - [Rank\ 1]) * [Sigmoid])$
- Move [Start] and [End] to Detail
- Move [Curve] to the Rows. Edit the Table Calculations for [Curve]. Under [Rank 1], Compute Using Specific Dimensions and select [Start], [End], and [Padded], and drag them so that they are in that order. Under [Rank 2] in the same Table Calculations window for [Curve], do the same thing, but make the order [End], [Start], [Padded]. For [t], make sure that it is only using the Specific Dimension [Padded].
- Move [Padded] to Path.
- Create a calculated field [sizing]
 $WINDOW_AVG(SUM([Numbers]))$
- Put [sizing] in Size, right click and Compute Using [Padded].

Everything else is cosmetic. You can remove background lines, trim the axes and remove the headers. Add color to the Sankey diagram and create bar charts as end caps on either side to provide context in a full dashboard. You may want to turn the tooltips off, as they can be distracting. You can also go to "Dashboard" → "Actions" to create interactive highlights.