

TABLEAU

DATA

PREPARATION

A DATA DOZEN WORKSHOP

BUILT FOR THE OLIN BUSINESS SCHOOL AT WASHU

DATA DOZEN

a visual analytics studio

LET'S GET STARTED

This workshop dives into the details of how to transform unruly data into Tableau-ready files. Designed for the technical novice, this workshop will give all business analysts a substantial leg up when it comes to data analytics. This content serves as the bedrock for all Tableau projects in the real world, the nitty gritty of preparing your data set for analysis. From creating the proper data structure with pivots, joins, and unions, to cleaning up the data's inconsistencies with calculated fields, data type changes, groups, and sets, you'll learn how to maximize the possibilities your data set has to offer. Don't forget to [download Tableau Public](#) and download the data set to build your own version! As always, all of the resources related to this workshop, including the introductory slides and complete Tableau workbook, can be found on our [workshop series website](#).

PART 1: LECTURE

INTRODUCTORY SLIDES

- ▶ All the slides related to the introductory lecture can be found on the <https://datadozen.com/olin-business-school/>.

PART 2: DEMO

WORLD BANK: WORLD DEVELOPMENT INDICATORS

- ▶ This type of [data bank of the world development indicators](#) from the World Bank is the epitome of an ideal data prep experience.
- ▶ When creating the data set is this fluid of an experience, we spend more time focused on the content of the data instead of its structure.

UNITED NATIONS: LIFE EXPECTANCY DATA

- ▶ Download the datasets
 - ▶ <https://population.un.org/wpp/> > Download Data Files > Mortality > Life Expectancy > download
 - ▶ Documentation > Definition of Regions > [First Link](#) to Methods > Full view > Excel, CSV download
- ▶ Life Expectancy and Predictions Sheets
 - ▶ Header
 - Copy and paste into a new sheet called "Source"
 - Added URL to Source worksheet
 - Added download date
 - Deleted UN symbol image
 - Checked bottom of worksheet for any additional extraneous information

- Deleted row 1 (checking to be sure “life expectancy at birth...” is documented already in source tab)
- ▷ Country / Area
 - Filter Type (column F) to just country/area
 - Delete unnecessary columns
 - A Index
 - B Variant
- ▷ Create Country Life Expectancy Sheet
 - Copy all country/area values
 - Create new sheet
 - Paste special > just values (this resets wonky formatting and indentation)
 - Rename column A to Country of area (United Nations)
 - Delete Type (This leaves 18 columns)
- ▷ Life Expectancy Predictions
 - Rinse and repeat all steps for the future predictions
- ▶ World Average
 - ▷ Rinse and repeat for world averages
 - ▷ Transpose the decade columns by copy and paste special
- ▶ Add Metadata into the same spreadsheet

PART III: DATA PREP IN TABLEAU

DATA STRUCTURE: PIVOTS, UNIONS, JOINS

- ▶ Pivot (Adding Country Life Expectancy (1950-2020) table)
 - ▷ Life Expectancy definition from UN: “The average number of years of life expected by a hypothetical cohort of individuals who would be subject during all their lives to the mortality rates of a given period. It is expressed as years.”
 - ▷ My Hypothesis
 - I have 201 original records, one record for each of the 201 countries or areas included.
 - Pivot decade halves (14 x 201 records = 2,814 total records)
 - ▷ Actual Results
 - Yes! 2,814 records.
 - Change variable names to be useful
 - Pivot Names = Decade Halves
 - Pivot Values = Life Expectancy at Birth
 - ▷ Build Sheets to Test Data
 - Count
 - Life Expectancy Heatmap
- ▶ Union (Adding 2 - Life Predict (2100) table)
 - ▷ My Hypothesis

- 201 countries * (14+16 decade halves) = 6,030
 - Select new Decade Halves headers and add data to pivot
- ▷ Actual Results
 - Wound up with double the predicted number of records: 12,060
 - Add the new table or sheet variable to your heat map on columns.
 - We've got null records for every decade that doesn't appear in the opposite sheet
- ▷ Data Source Filter
 - Add a data source filter that only keeps non-null values for life expectancy
 - Add sheet name to life expectancy heat map
- ▶ Join 1 = Left Join (Adding 3 - Country Codes UN table)
 - ▷ Hover over the table and click the dropdown arrow and open
 - ▷ Join on Country code UN and M49 Code
 - ▷ My Hypothesis
 - My goal is to have the same number of records (6,030) since I'm not changing the granularity of the data set, just adding some helpful metadata columns.
 - An inner join and left join should produce the same number of records because the metadata table should have all of the countries in the life expectancy table.
 - ▷ Actual Results
 - My hypothesis is completely off. I've got confusion with every type of join.
 - An inner join is dropping 60 records (5,970).
 - A left join keeps all records (6,030), but the inner and left join should be the same if all of the countries in the first table appear in the second. Why are they different?
 - ▷ Investigative Visualization
 - Create a new sheet and compare the two data sets Country or Area variables.
 - Taiwan and the Channel Islands aren't included in the metadata table.
 - This join decision suddenly becomes politically charged: to include or exclude Taiwan and its life expectancy data?
 - ▷ Right Join Confusion
 - A right join and full outer join aren't producing additional records. Why? If the second table has substantially more countries than the first, why aren't these additional countries showing up with a right or a full outer join?
 - I forgot that I had the data source filter that excludes all non-null life expectancy cells, which of course they don't have because they're not in the life expectancy table.
- ▶ Join 2 = Any Join (Adding 4 - World Life Expec (PIVOTED))
 - ▷ Joining 1 - Country Life Expectancy with 4 - World Life Expec (PIVOTED) on Decades Halves
 - ▷ My Hypothesis
 - Will use inner join and keep the same number of records (6,030) since both tables will find an exact match in each table.
 - ▷ Actual Results
 - Yes, that worked just as expected. This is a great example of where you can use any join and keep the same number of records because every record will find a match in the other table.
 - ▷ Adjust Default Properties
 - Create a new worksheet and pull out the decade halves to rows and put the world life expectancy on text. Change aggregation to minimum.

- Adjust the name of the variable to emphasize that it's a world aggregate, and change the default properties to aggregate at the minimum.
- ▶ Join 3 = Left Join with Calculation (Adding 5 - UN Notes table)
 - ▷ New Data Source
 - Data Menu > New Data Source > Excel > Choose same data source
 - Drag out just the 5 - UN Notes table
 - ▷ Create Calculation
 - Our goal is to extract just the notes letter or number from the rest of the string variable. To do this we'll use three different functions:
 - LEFT
 - LEFT([Notes],4)
 - Extracts the first four characters of each string
 - TRIM
 - TRIM(LEFT([Notes],4))
 - Removes any extra spaces on the shorter codes
 - REPLACE
 - REPLACE(REPLACE(TRIM(LEFT([Notes],4)), "(", ""), ")", "")
 - Removes the open and close parenthesis
 - Copy Calculation
 - Copy the calculation you've created.
 - Navigate to one of our worksheets, and you'll see that you have two data sources open at the top of your data pane. They're independent.
 - Click on your 1 - Country Life... data source, and then use the tab at the bottom of the screen to navigate back to the data source window.
 - ▷ Join on Calculation
 - Calculation
 - Pull out 5 - UN Notes Table
 - Select Join on Calculated from dropdown menu
 - Paste in your new calculation and replace the [Notes] variable with [Notes (5 0 UN Notes)], which is the additional information Tableau has wrapped around our data now that we're keeping track of multiple tables.
 - Left Join
 - We want to add notes to all of the variables that already exist in the data set and disregard any extra notes that might not be applicable.
 - Data type
 - Change the data preview from the grid to the list icon, just to the left of "Sort fields"
 - Change the data type for the first Notes variable to a string to match the new variable we just created with our calculation. Remember, in order to join two variables, they have to be the same data type.
 - Close Extra Data Source
 - Right click on the extra Notes data source that's still open at the top of the data pane and select Close.

- ▶ Extract Data Set
 - ▷ Once you have a complicated data set assembled, you often want to be able to extract the entire set to use independently or in other software.
 - ▷ Note: These instructions give you the same output as one of the tables in our excel workbook (B - UN Life Expectancy Data Set).
 - Navigate to your number of records worksheet.
 - Right click on the number of records (should have finalized at 6,030) and select view data.
 - In the pop up box, switch from the Summary tab to the Full Data tab.
 - Click the corner box just to the left of the Country code variable, above all the rows where the table meets. This will select all the data.
 - With all the data highlighted, select Copy to copy and paste your finalized data set or Export all to save it as a CSV.

DATA ORGANIZATION

- ▶ Data Pane Organization

After all of that structural work, there's no reason to still wade through a cluttered data pane. Here are a few tips to clean up your data pane and make sure that the variables you've worked so hard to put in the right columns and rows actually come into your visualizations in a helpful way.

 - ▷ Double Check Tableau's Assumptions
 - Dimensions and measures
 - Data types (dates and id codes are common culprits)
 - ▷ Hide Variables
 - Review and hide any unhelpful or duplicative variables. (Don't worry. You can always bring them back.)
 - ▷ Folders
 - Change the data pane sort from Group by Data Source Table to Group by Folder.
 - Then group your variables by helpful themes to streamline your work.
 - ▷ Default Properties
 - Set the default properties for your variables to useful presets to avoid repeating the same changes over and over. Adjusting the aggregation, number format, and sort up front can save you a ton of time as you build your visualizations.
 - Comment
 - Excellent place for metadata, such as a description of the variable, an original source, contact information for questions, etc.
 - ▷ Aliases
 - Useful to shorten or extend names.
 - Can be used to replace variables represented with a number, but usually it's better to join those numerical values with their descriptive counterparts.
 - ▷ Nulls
 - A quick word about null values. Tableau does a comparatively good job handling null values. There's no reason to go through and replace all your empty cells with any sort of placeholder (n/a, 0, etc.) before connecting your data to Tableau.

- Just make sure you know why the cells are empty, and document that for your future self and for your colleagues.
- ▶ Categories

Here are a few ways to extend the functionality of your data by creating quick and helpful categories.

 - ▷ Hierarchies
 - Region, Sub-Region, Intermediate Sub-Region, Country or Area
 - ▷ New Variables
 - Most granular sub-region
 - IFNULL([Intermediate Region Name],[Sub-region Name])
 - ▷ Groups
 - Decade Halves into three groups: 20th century, 21st century, future
 - ▷ Sets
 - Filter to countries by least developed and land-locked countries
 - Select those 17 countries and make a set
 - ▷ Bins
 - Create bins size of 10 from Life Expectancy
 - Histogram by counting up the countries (view data is helpful here to make sure it's working)

OTHER TABLEAU RESOURCES YOU MAY NEED DOWN THE ROAD

- ▶ [Live vs. Extracts](#)
- ▶ [Refresh Data Source](#)
- ▶ [Data Interpreter](#)
- ▶ [Tableau Data Source \(.tds\)](#)
- ▶ [Tableau Prep](#)
- ▶ Data Relationships
 - ▷ [Relate Your Data](#)
 - ▷ [How Relationships Differ from Joins](#)
 - ▷ [Questions about Relationships, the Data Model, and Data Sources](#)
 - ▷ [Don't Be Scared of Relationships](#)
 - ▷ [The Tableau Data Model \(Physical v. Logical Tables\)](#)

WRAP UP

SAVE YOUR WORKBOOK

The final step is to save your workbook! If you haven't already created a Tableau Public account, [return to Tableau Public](#), click Sign In, and create a new account. Once you've done that, here's how you save your workbook.

- ▶ Go to File, Save to Tableau Public As, and then enter your Tableau Public account credentials.
- ▶ Name your workbook.
- ▶ Edit the workbook details on your Tableau Public page. You can control whether the workbook appears on your profile page and whether someone else can download the workbook and data.
- ▶ Note that you should NEVER save anything private on Tableau Public. This is a place to only save work you've done with public data, like we've done here. To save your work locally, you will need [Tableau Desktop, which is completely free for students and instructors](#). If you're an administrator, it will run approximately \$70 a month.

TAKE THE SURVEY

Now that you're all done with the workshop, we'd so appreciate your feedback so that we can continue to improve our workshops for future attendees. [Just click here](#) to answer a few quick questions. Thank you!