Setting up a Dashboard

Data Visualization and Design | CUNY Graduate Center | Summer 2019

This tutorial is adapted from one written by Erin Waldron of Data Dozen

This tutorial will help you clean messy data and prepare data to be read computationally.

Goals

1. Recognize common data structure errors
2. Identify variables vs data in spreadsheets designed for human eyes
3. Pivot data in Tableau to create long and skinny data

Data

Meal Plan Example

Population Division's World Population Prospects (Download Datasets >> World Population Both Sexes (Standard Projects // Population Indicators))

Messy Data Records vs. Unusable Data Structure

When we talk about cleaning a dataset, the changes we make usually fall into two categories: changes to the data structure or changes to the data records themselves. Think about where the 311 data got messy. Most of this came down to fuzzy categories. The taxonomy was inconsistent, and therefore the data was difficult to aggregate with complete accuracy. However, this messiness was restricted to the data itself (i.e., the content inside the cells). The data structure (i.e., the variables which dictate the column structure) was consistent and ready for a computer to process.

Today, we'll be talking about how to fix problems in this second category: problems with the structure of the data. Datasets are often designed for human eyes, which does not work well when a computer program like Tableau tries to read the content. We'll do a fair bit of restructuring by hand to ensure that you really understand what we're moving around. In the future, you will likely approach data cleaning like this with a program of some sort Let's look at a tiny text table to start, and then work through restructuring a real dataset from the United Nations.

Features of "Tidy" Spreadsheets

The idea of "tidy data" comes from a paper written by Hadley Wickham who was instrumental in the development of the R programming language. In this paper, he cites XX features of data that make it "tidy".

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table

Seems simple enough, but there is far more detail. What this means is that there should never be "data in the headers". A good way to check for this is if you can group all the headers into a category (i.e., date/time/person/etc.), then you have data in your headers. It also means that there shouldn't be 2 types of data in 2 column, i.e., if there is a code about a person that indicates 2 things (maybe pay grade and level), then that must be broken apart. We'll look at more examples shortly and transform a data table.

It also means that tables *about* different things need to be broken apart. Last week, we noted that the 311 data was *about* the complaint - so everything in the table was about the 311 data. Now, what if we wanted to combine 311 complaints with income about the neighborhoods. This data should be kept in 2 tables that are joined together because there is no inherent relationship between a 311 complaint and the income of a neighborhood. These tables should be joined on a common column (i.e., Neighborhood Tabulation Area).

Tidy data is also called "long and skinny" data

The reason we want tidy data is for our computers, and Tableau has some very helpful tools for us to work with.

The "Human Eyes" Spreadsheet

Why are spreadsheets designed for human eyes?

- Often spreadsheets are built for human eyes to make creating and visually referencing the spreadsheet easy.
- This is common with archival work, quantifiable self data, spreadsheets that are maintained by humans (which are especially common in offices with small datasets that aren't stored in databases)

Meal Plan Example

Cleaning by Hand (1)

1. Delete the 3 rows that do not contain data

2. Remove redundant information (the day of week can be directly inferred from the date)
3. Get the data out of our columns
4. Move each meal/price so that each row only is unique
5. Split the meal from the price

Our final Table will have the following headers: Meal Location Date Option Price ($)

With only three columns (one per variable), this will turn into a long and skinny dataset. It will be repetitive, which is exactly what we're looking for!

SUMMER SESSION SKIP 'Cleaning with Tableau (1)'

Cleaning with Tableau (1)

1. Load the original Meal Plan Excel file into Tableau
2. Click on 'use data interpreter'
3. Inspect the changes the Data Interpreter made.
4. Select all of the columns that have the date in the header.
5. Right click and select 'Pivot'
6. Rename the column headings.
7. Create a simple visualization
    1. Meal on Rows
    2. Date (discrete day) on Columns
    3. Option on Text and Color on the Marks Card

ALL SESSIONS RESUME HERE

Restructuring United Nations Population Estimates Spreadsheet

Now let's look at a real example with much more data: population estimates from the U.N. 1. Download the Population Division's World Population Prospects from the UN's website.

2. Go to the Data tab and then to the Download Center.
3. Download the first Excel spreadsheet called Total Population - Both Sexes.
4. Open it up in Excel and you'll see clearly this is created for human eyes.
5. If you try to point Tableau to this, it won't know which way is up. So let's clean!
6. Make a fresh copy in your Excel workbook titled Cleaned Structure and take a good hard look at the table. What do you see that needs to be cleaned up to make this Tableau ready?

Cleaning by Hand (2)

1. Cut the blue header and paste into a new sheet in the workbook called source (you don't want to get rid of important information like where your data came from while you clean)
2. Delete all of the rows where the source header was
3. Delete the row that says "Total Population, both sexes combined as of July 1 (thousands)" We'll come back to the thousands conversion in a bit
4. Delete the World income developed population summaries (index rows 1-12)
5. Now we need to breakout the "Region, Subregion, Country or Area" Column into the three separate variables it contains
6. Insert a new column to the left and title it "Region"
7. Insert a new column in between "Region" and "Region, Subregion, Country or Area" and title it "Subregion"
8. Rename "Region, Subregion, Country or Area" to just "Country or Area"
9. Now it's time to copy and paste. First, we'll fill in the Region column. Copy "Africa" and paste it in Column C until you reach Asia's subtotal line. Then copy "Asia" and paste it until you've reached Europe's subtotal. Rinse and repeat until you've filled in the "Region" variable for all of the
10. Using the same technique, we'll copy and paste the "Subregion" to fill in the data for column D
11. Now delete all of the rows that are "Region" and "Subregion" subtotals. Once you've done that, you should have 233 records (or 234 rows including the column headers).
12. Finally delete the "Index" and "Variant" columns, which won't be applicable in Tableau.

Cleaning with Tableau (2)

We're going to use Tableau to do a lot of the heavy lifting here.

1. Load your spreadsheet into Tableau as an Excel File
2. Drag the Estimates Sheet over to the Data Source Pane. A lot of Null Values will appear. This is because of the header content.
3. The Data Interpreter will do a lot of the work for you. It will remove the header data and identify what should be the header. Select the checkbox on 'Use Data Interpreter'
4. This should look a whole lot better, but we want to make sure that we know exactly what the interpreter did. Select 'Review Results'.
5. Notice that we have 'Years' in the headers. Since we could refer to them as 'Years', we know that we'll need to fix that. But before we do that, let's go over to our Worksheet and see what our options are. Imagine that you want to do something completely obvious here: you want to make a visualization of the population change over years. You can't do it.

We are going to PIVOT the data.

5. Select all of the columns from 1951 to 2015 by clicking on 1951 and the Shift+click on 2015. Right click and select 'Pivot'
6. Now rename "Pivot Field Names" to "Year" and rename "Pivot Field Values" to "Population (by thousands)"
7. Click on the "Abc" data type symbol above "Year" and change the data type to "Number (Whole)"
8. Finally, let's create a population variable that's a whole number
9. Click on the down arrow in the Population (in thousands) and select "Create Calculated Field..."
10. Title your new variable: CALC: Population
11. The "Population (in thousands)" Variable should appear into the text window. Multiply it by a thousand. Your text box should look like this: [Population (thousands)] * 1000. Then say ok. You'll see your new variable show up in the Measures section of your data pane.
12. Now go back to your Worksheet, success! You can make a line chart of population over time.

Compiled values

1. Now go back to your Worksheet and trade out year for Country. What do you notice?

Check Your Work

When you finish a big restructuring job like this, you need to check that you've accomplished what you set out to do: restructure the dataset WITHOUT losing any of the data. Spot check that you've brought in all the records by comparing a few totals in Tabelau with the original UN spreadsheet.

1. Open up the original UN spreadsheet with the blue heading on the top
2. Look at the world total for 1950 = 2,536,274,721 (cell F18)
3. Now go back to Tableau. Put population on text and filter the viz by year to 1950. Do you get 2,536,274,721? If so, well done! Change the filter to another random year or two and ensure that your answer in Tableau aligns with the world population for that year in the UN table.
4. Now add the Country or Area data pill to filter and spot check a few years by country. For example, if you set your filters to 1995 in Argentina, do you get a population of 34,994,814?

Word to the Wise: Look for the CSV or Server Connection

Head back over to the Population Division's World Population Prospects website. Go to back to the Data Download Center and take a look at the top. Under "Major topic / Special Groupings" you'll CSV format. This is the format we would have wanted! When you see data formatted for statistical software, this is what you're after. I've made this mistake before, moving too quickly looking for data, and not realizing that both a "human eye" version and "computer" version are being offered. If you download one of these datasets and take a look, you'll immediately see the difference. They're Tableau ready.