**Tutorial: Data wrangling with Tableau and Excel**

Caitlin Havlak

Last edit: September 16 2016

**Introduction:** Data wrangling is the process of preparing raw data for use in a data analysis or visualization software. It can involve detecting and correcting corrupt or inaccurate data from a dataset and converting data into a more useable format by restructuring data tables. Tableau has tools for simple cleaning, but often more advanced cleaning techniques are required. There are several online wrangling tools for cleaning data, such as Trifacta Wrangler and OpenRefine, but you will find that the use of these more advanced tools is rarely necessary. This tutorial walks you through the cleanup of several spreadsheets using Tableau and Excel.

Data is rarely received in a perfectly clean format that is compatible with the software you are using. It is almost always necessary to do some data cleaning.

What are the causes of dirty data?

(1) **Data entry error**: Data entry is still commonly done by humans, so data can accidentally be corrupted at the time of data entry.
(2) **Incompatible tables**: If you want to join two datasets by a single column, the formatting of the column may not be the same for each table. Each datatable may make sense on its own, but together are incompatible.
(3) **Incompatible table format**: You'll find that some datatables will have extra headers or footers that Tableau doesn't know how to read. These headers and footers make sense for use in Excel and other software, but are incompatible for use in Tableau.

We will see examples of all three of these causes with the datasets that we will be using in today's tutorial and in the homework assignment.

What should we look out for when cleaning data?

(1) **Table formating**: Look out for extra headers and footers. Does the data need to be reshaped in some way (wide to tall)?
(2) **Variable type**: Did Tableau interpret your variables correctly (e.g., character, numeric, boolean)?
(3) **Invalid character values**: Are there extra spaces? Are there spelling mistakes?
(4) **Invalid numeric values**: What units are you working with (e.g., millions or thousands)? Do most of the data values fall within a certain range of values? Values that fall far enough outside that range may be data errors.
(5) **Grouping data**: When combining two datatables, are they compatible?
(6) **Missing values**: How does your datatable deal with missing values (e.g., NA, NULL, blanks)? Is this the same as how Tableau deals with missing values?

**Learning objectives:**

- Part I: Ideal format of data in Tableau
- Part II: Tableau's data interpreter
- Part III: Joins
- Part IV: Cleaning in Excel or Google Docs
- Part V: Pivots

## Part I: Ideal format of data in Tableau

The first step in exploring your data in Tableau is examining how the data is presented. When a dataset is formatted with sub-tables, hierarchical headers, extraneous headers and footers, or blank rows and columns, Tableau has trouble interpreting the data. To be able to use this data in Tableau, the dataset should have a specific data structure before you do any analysis in Tableau. Each variable should be its own column and each observation should be a row. Your data should follow these three rules:

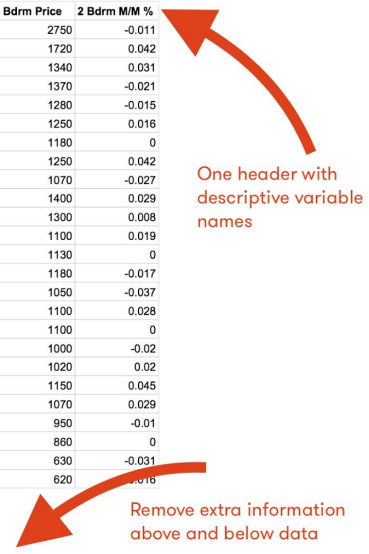
1. Start your data in cell A1. Remove all introductory information and footnotes.
2. Have the first row be the column headers/variable names
3. Have every subsequent row be one observation. No cross-tabulation!

**Before:**

| Market rental pricing for July 2016 | | | | |
|---|---|---|---|---|
| | 1 Bedroom | | 2 Bedrooms | |
| City | Price | M/M % | Price | M/M % |
| Vancouver | 1740 | 0.024 | 2750 | -0.011 |
| Toronto | 1350 | 0.023 | 1720 | 0.042 |
| Calgary | 1110 | 0.037 | 1340 | 0.031 |
| Victoria | 1075 | -0.023 | 1370 | -0.021 |
| Ottawa | 1040 | 0.03 | 1280 | -0.015 |
| Edmonton | 1010 | 0.031 | 1250 | 0.016 |
| Regina | 980 | 0.021 | 1180 | 0 |
| Montreal | 960 | 0.011 | 1250 | 0.042 |
| Kingston | 940 | 0.011 | 1070 | -0.027 |
| Kelowna | 920 | 0.022 | 1400 | 0.029 |
| Barrie | 910 | 0.034 | 1300 | 0.008 |
| Halifax | 910 | 0.022 | 1100 | 0.019 |
| St.Catherines | 905 | -0.005 | 1130 | 0 |
| Oshawa | 890 | -0.011 | 1180 | -0.017 |
| Hamilton | 880 | -0.011 | 1050 | -0.037 |
| Winnipeg | 880 | 0.023 | 1100 | 0.028 |
| Saskatoon | 870 | 0.048 | 1100 | 0 |
| Kitchener | 860 | 0 | 1000 | -0.02 |
| Windsor | 840 | 0.024 | 1020 | 0.02 |
| Abbotsford | 800 | 0.013 | 1150 | 0.045 |
| Quebec City | 800 | 0.013 | 1070 | 0.029 |
| London | 790 | 0 | 950 | -0.01 |
| St. John's | 775 | -0.006 | 860 | 0 |
| Sherbrooke | 560 | -0.018 | 630 | -0.031 |
| Saguenay | 540 | -0.018 | 620 | -0.016 |
| | | | | |
| | | | | |
| Source: http://blog.padmapper.com/category/rental-data/ | | | | |

**After:**

| City | 1 Bdrm Price | 1 Bdrm M/M % | 2 Bdrm Price | 2 Bdrm M/M % |
|---|---|---|---|---|
| Vancouver | 1740 | 0.024 | 2750 | -0.011 |
| Toronto | 1350 | 0.023 | 1720 | 0.042 |
| Calgary | 1110 | 0.037 | 1340 | 0.031 |
| Victoria | 1075 | -0.023 | 1370 | -0.021 |
| Ottawa | 1040 | 0.03 | 1280 | -0.015 |
| Edmonton | 1010 | 0.031 | 1250 | 0.016 |
| Regina | 980 | 0.021 | 1180 | 0 |
| Montreal | 960 | 0.011 | 1250 | 0.042 |
| Kingston | 940 | 0.011 | 1070 | -0.027 |
| Kelowna | 920 | 0.022 | 1400 | 0.029 |
| Barrie | 910 | 0.034 | 1300 | 0.008 |
| Halifax | 910 | 0.022 | 1100 | 0.019 |
| St.Catherines | 905 | -0.005 | 1130 | 0 |
| Oshawa | 890 | -0.011 | 1180 | -0.017 |
| Hamilton | 880 | -0.011 | 1050 | -0.037 |
| Winnipeg | 880 | 0.023 | 1100 | 0.028 |
| Saskatoon | 870 | 0.048 | 1100 | 0 |
| Kitchener | 860 | 0 | 1000 | -0.02 |
| Windsor | 840 | 0.024 | 1020 | 0.02 |
| Abbotsford | 800 | 0.013 | 1150 | 0.045 |
| Quebec City | 800 | 0.013 | 1070 | 0.029 |
| London | 790 | 0 | 950 | -0.01 |
| St. John's | 775 | -0.006 | 860 | 0 |
| Sherbrooke | 560 | -0.018 | 630 | -0.031 |
| Saguenay | 540 | -0.018 | 620 | -0.016 |




**Source and additional information**: see this Tableau article for more information and Hadley Wickham's article, "Tidy Data"

**Part II: Data interpreter**

Data sources: Rental data (xlsx)

There is an ideal format for data in Tableau as mentioned in Part I. But we don't always have to manually clean data before we import it into Tableau. Tableau's Data Interpreter feature draws out sub-tables and removes some of that extraneous information to help prepare your data source for analysis. Note: the data interpreter only works with Microsoft Excel files, not CSV or other file types.

The data that we will be using for this exercise is from Padblogger. They collect data on monthly market rental prices. We will be looking at the data collected for June and July 2016. This dataset also includes a "M/M %" variable. This is the month to month percent change between June and July 2016.

**Instructions**:

(1) Download the Padblogger data (xlsx).
(2) Connect the excel file to Tableau.
(3) Drag over the "July 2016" table to the area that says "Drag sheets here".
(4) Look at your datatable. Does it look right to you? Consider the table format, header, footer and variable type.
(5) Under the sheets area on the left sidebar, look for the "Use Data Interpreter". Click the checkbox to turn the data interpreter on.
(6) Check out the datatable. How does it look now? Consider the table format, header, footer and variable type.
(7) Change the variable names to better reflect what they are. E.g., "1 bdrm price" and "2 bdrm price".
(8) Make a map displaying each city in the dataset by double clicking the City pill in the Dimensions area
(9) Visually encode each city point using the color and size properties, and the rent price and month to month growth variables.
(10) Notice that at the bottom right corner of the map, it says "1 unknown". Click the button and edit the location name of "St.Catherines" to "St. Catherines".
(11) Add a border to the circles and change the gradient to green to red.
(12) If time allows, create the same map for June 2016 rental prices.
(13) Create a dashboard that compares the one and two bedroom rentals.

**More information here and here.**

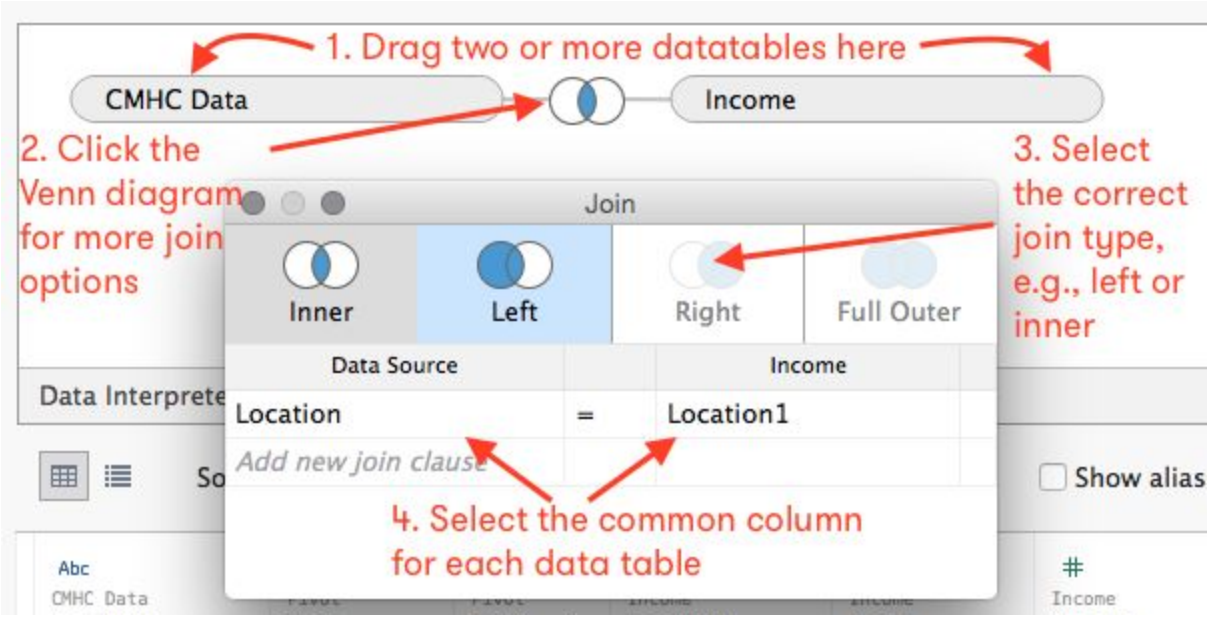
**Part III: Joins**

Data sources: Rental data (xlsx)

A JOIN is a means for combining columns from one or more tables by using values common to each. For example, we have rental data for June 2016 and July 2016. It would be nice to have

both of these datatables combined so that we can compare the two months. We can do this by joining the two datatables by a common column, “City”.

In Tableau, you can join two datatables by dragging each table into the area that says “Drag sheets here”. If you click the Venn diagram, you can select the correct join type and common column to perform the join on.




There are four main join types: inner, left, right and full outer.

- An **inner join** preserves only the rows that have the same key field, in our case name, between both tables. We only get information for names that are listed in both tables—and there are no nulls.
- A **left join** brings in all the information for the rows from the table on the left (siblings) and any information from table on the right (eye color) for rows with the same name as the sibling table.
- In a **right join**, it’s the reverse. We have all the names from the right (eye color) table, and for those rows we get the sibling (left table) information if it exists.
- An **outer join** brings in all names listed in all tables, and fills in nulls wherever there isn’t information for a given column for that row.

**Source and more information here.**

**Instructions**:

(1) Return to the “Data Source” tab.
(2) Drag over the “June 2016” sheet, ensuring that the data interpreter is still turned on.
(3) Tableau has automatically assumed that you want to join these two tables using an inner join. It also automatically updates the June 2016 variable names to match the July 2016 variable names. Confirm that Tableau interpreted these variables correctly.

(4) Hover over the Venn diagram. It should say “Inner Join of ‘July 2016 and June 2016.’ City=City (June 2016)”. This is what we want. Since the City column in each of these datatables is exactly the same, it doesn’t matter if we choose an inner, left or right join.
(5) Make a new dashboard that compares June and July data.

**Part IV: Cleaning in Excel**

Data sources: Income data (xlsx) or (Google spreadsheet)

Sometimes the data interpreter in Tableau isn’t able to detect all of the errors in the dataset. In cases like this, you will need to manually clean the data in Excel.

For this exercise, we will add a third table to our analysis: mean total income by city. Review this dataset and consider table formatting, variable types, invalid character values, invalid numeric values, grouping data and missing values as described above.

**Instructions**:

(1) Open xlsx file or google spreadsheet file.
(2) Manually remove the extra information in the header and footer.
(3) Change the variable names in the header so that Tableau will recognize the header as a numeric value. I.e., change “Projected 2015” to “2015”.
(4) We want to eventually join this table with the June and July rental data using the “City” column. Remove the provinces and make sure that the city names match.
(5) Download the new Excel File.
(6) Return to the Data Source tab in Tableau.
(7) Click on “Add” next to the connections area on the left sidebar. Select “Excel” under “Add Connection”
(8) Drag the income sheet and join the income sheet with the June and July sheets. Join by the city column.
(9) Make a new map that shows major Canadian cities, sized by price of a 2 bedroom rental.
(10) A common affordability index is to consider the proportion of one’s income that goes towards housing. Create a new calculated field (under analysis on the main menu). Name the calculation whatever you like. In the calculation field, write, “[2 bdrm price]*12/[2016]*100”. This finds the annual rent cost and divides it by the projected 2016 income level. Times by 100 to make it a percent. Color the dots by this new calculation.
(11) Format the dots in a way that makes sense and update the tooltip.

**Part V: Pivots**

Data source: Average cost of rental housing data

For this part of the tutorial, we will look at a new dataset from the Canadian Mortgage and Housing Corporation (CMHC) that contains data on average rental prices. The Padblogger dataset from before is for the current market price for newly listed rentals, whereas, the new

CMHC dataset is for the average rental price paid, regardless of how long renters have been living in their unit. For example, a renter could pay $800/month on rent for a two bedroom apartment but that's only likely if they have been in the same apartment for the past five or ten years. Whereas, market rate for a two bedroom apartment is more like $2000/month if it were to come onto the market today.

This data is stored in a cross tabular format. Tableau has trouble interpreting this format of data and will need us to convert it to a columnar format.

**Instructions**:

(1) Make sure that the dataset is in an appropriate table format before connecting to Tableau.
(2) Open a new Tableau workbook.
(3) Connect the new dataset and drag it over.
(4) You may find that you need to use the data interpreter to get Tableau to read the data's column names correctly.
(5) Take a look at the data table. Tableau has trouble visualizing data in cross tabular format. Instead of having years in the column headers, we need one column header called "Year".
(6) To do this, highlight all of the columns with a year column header. On a Mac you can do this by holding down the command and shift keys.
(7) Scroll all the way over to the 2015 column and click the triangle in the top right corner. Select "Pivot".
(8) Our data is now in columnar format. Rename the variables to "Year" and "Rent price"
(9) Make a new map that colors the city point by rent price and disaggregates the map by year. What does this show? What do this not show?

**More information here.**