Data Prep

Data Visualization and Design | CUNY Graduate Center | Summer 2019

This tutorial will help you clean survey data.

Goals

By the end of this tutorial, you will be able to: - Transform data using Tableau Prep - Standardize messy entries - Remove extraneous columns - Break apart lists of values.

Data

Restaurant Recomendations DHUM 7300

Steps

1. Go to the survey we created on the Google Drive
2. Add it to your Google Drive
3. Notice a few things.
    1. Everyone entered their name a different way - we will need to standardize that
    2. Some people entered multiple food types per restaurant - we will need to break those apart.
    3. Not all of our locations have an address
    4. Not everything has a star rating.

This dataset is unnecessarily dirty because of the way that the questions were asked. The best way to get clean data is with a clean structure. Google Forms, Survey Monkey, etc. have functions such as this built in. It is easier to consider what you want and then force users to give answers that way than it is to think of all the ways that users will creatively input their data.

Getting Started

1. Open Tableau Prep
2. Under 'Connect' select To a Server >> More >> Google Sheets
3. Sign in to your Google Account
4. Select your dataset

Part I : Rename the Columns Renaming the columns just makes them easier to work with. Be sure to use names that are representative to make it easier on yourself. Many datasets use codes rather than descriptive labels.

1. Near the top of each column, click the down triangle & select 'Rename Column'
2. Give each Column 1 to 2 word names (see video for my choices)

Part II : Break apart the lists in the Food Types Field We're making an assumption here that the first food type is the most representative. That may or may not be correct.

1. We need to create a new calculated field to break up the Food Types.
2. Click the down triangle again and 'Create Calculated Field'
3. Click the arrow to the left if you can't see the options.
4. We'll make a new column of just the dominant food type. Use this formula: `SPLIT([Type], ',', 1)`
5. Now rename this column

Part III : Get first names only and standardize the orthography We've made the assumption here that there cannot be spaces in a first name. This is a bad practice and can be avoided by asking better questions i.e., for First Name and Last Name.

1. We need to create a new calculated field to break apart the names. We'll make this field all sentence case (even though we re)
2. Click the down triangle again and 'Create Calculated Field'
3. Click the arrow to the left if you can't see the options.
4. We'll make a new column of just the first name (we assume). Use this formula: `SPLIT([Name], ' ', 1)`
5. Now rename this column. **If you need to rename multiple columns, follow this tutorial

Part IV : Transform those addresses into latitude and longitude.

1. We actually can do this in Tableau, but it is typically not great. There are a variety of geocoders that work much better census gps texas A&M.

Unfortunately, the way we asked this question will not allow us to use an online geocoder. Again, it goes back to asking questions in a way that gets us the answers we want. One thing we can do is visualize how many of these restaurants are in each borough.

1. On the second page of the workbook is a key [LatLon]. Download this as a csv file
2. Select 'Add' next to 'Connections' and select 'Text File' (csv's are read as text files - NOT Excel files - because of the way the files are formatted)
3. You will be prompted to do a join. Select a Right Join. Even if a borough is not represented (i.e., sometimes Staten Island gets forgotten even though we don't want to forget it in our visualization)
4. Close the dialog box.

Part V : Typos

By default, Google Sheets connects in 'Live' mode, meaning that the data is constantly being updated. If you found a typo (maybe someone misspelled a word, then you have to fix it in the original, not here). Since we are all using the same dataset, there may not be any typos left for you to fix.

Some ways to solve this include what we did for the Borough Names. We were interested in data about NYC, so we offered our respondents a short list of acceptable responses. You can use check boxes and similar formats to constrain the answers you get.

You can also extract the data, which gives you a snapshot copy of the data in your workspace that will not be updated when new entries are found. It's really a matter of what your goals are for your project.

Congratulations! You've finished uploading the data and cleaning it. Now is time to move on to do some basic visualizations with it.

Tutorial written by Michelle McSweeney, PhD for *Introduction to Data Visualization*, a course in the M.A. in Digital Humanities at the Graduate Center at CUNY. More information about the program is available here.