Calculated Fields with the SSN Dataset

Data Visualization and Design | CUNY Graduate Center | Summer 2019

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

Goals

- Develop a solid command of calcuation and string manipulation
- Construct IF/THEN statements in Tableau

Data

This Tableau workbook that uses this dataset

Premise

We have a few questions:

1. What is the distribution of first letters in US Baby names?
2. Are there any outliers: Male or female names that have changed or are assigned to different genders more than others
3. What is the popularity of my name over time?
4. Any interesting trends in baby names that we can look into?

We will use a few visualizations to complete this

2. A treemap to show the distribution of names
3. A line chart and search bar to search for your own name
4. A scatterplot to show the position of a name between male and female
5. Stacked bar charts to show the % of a name for male vs female
6. An alphabet to show the distribution of first letters (shout out to Erin Waldron for this cool idea!!)

Calculated Fields

Often, the data that you have isn't in the format that you need. This tutorial will show you how to transform your data to fit the visualization that you want to make. In the next 2 labs, we are going to use the Social Security Administration database of baby names to understand trends in names over the past 150 years. In this first part, we are going to perform a series of simple operations on the variables to prepare it for our visualizations. In the second part, we are going to use those new calculated field to tell a story about names.

IF/THEN

First we need to know how many of each name are female, and how many are male. I.e., How many Females are named ‘Kelly’ and how many Males. We’ll make a new calculated field use an IF/THEN statement. This reads as “if the name is Female, then keep the occurrences”

1. Right click on ‘Occurrences’ and select ‘Create Calculated Field’
2. Call this new field ‘CALC: Female’
3. Enter this formula: `IF [Sex] = "F" THEN [Occurrences] END`

Do the same for Male.

Static Numbers

Now we’re going to use a little trick, this will help us in all of our future calculations. Currently, each record is unique: it’s the combination of name/gender/year. We’re going to want to work with the total number of occurrences for each name, irrespective of the Year. So let’s create a new variable that is just the sum of a given name. We’ll indicate this as being ‘Fixed’ since it isn’t really a variable, but a static number about our dataset.

1. Right click on ‘Occurrences’ and select ‘Calculate new field’
2. Enter this formula: { FIXED [Name]: sum([Occurrences])}
3. Call your new field CALC: Fixed Sum Occurrences

This will appear as ‘Undefined’ in your data table since it isn’t a variable.

Percent Male/Female

1. Right click on ‘CALC: Female’ and select ‘Calculate new field’
2. Enter this formula: {FIXED [Name] : sum([CALC: Female]) / sum([Occurrences])}
3. Call your new field Calc: Percent Female

This will appear as ‘Undefined’ in your data table since it isn’t a variable.

Do the same for Male

String Manipulation

Now we want to ask questions about the first letter of each name. We’ll need to make a new variable with just the first letter.

1. Right click on the Name column and select ‘Calculate New Field’
2. Enter this formula: `LEFT([Name],1)`
3. Call your new field CALC: First Letter

Boolean

Finding the interesting data points: true or false (or null)

Move to the Data Sheet, we are going to make 2 more Calculated fields here. One thing people are always interested in is name for different genders, such as a boy named Sue. Let's try to find these interesting data points. We'll find some trends in names: let's find all of the names that are owned by females less than 75% of the time, but more than 25%. We want those that are not uniformly Female, but also not just those that are given to 1 or 2 children.

1. One the drop down near 'Dimensions', select 'Create Calculated Field'
2. Enter this formula: `[CALC: Percent Female] > .25 AND [CALC: Percent Female] < .75`
3. Call your new field CALC: Female Outliers

Notice that this has a T/F data type - it's a boolean. Drag this to 'Filter' and find only the 'True' values.

In the next tutorial, we are going to work through telling a story with this data, but for now, let's play with it a bit and do some EXPLORATORY analysis.

Exploratory Analysis

Let's see what names are most interesting for males and females. We'll make a dispersion plot for male and female names that fit into our outliers criteria.

1. Drag Name to columns
2. Year, CALC: Female Outliers and CALC: Male Outliers
3. Use the drop-down on CALC: Female Outliers to create a filter and only filter for True values
4. Drag Sex to the color marks card and click on the color marks card to change the colors. I'll use one of the pre-set palettes. I don't like the colors they chose together, so I'll click on the 'F' and double click on a new color.

Great! This gives us a map of what names might be most interesting - what names have changed over time.