Knowing Ourselves

Excuse me while I geek out on some data. I've been wondering for a while about retention rates in volunteer groups. And I just happen to have a lot of data about signups for one of them. So I thought I'd start asking some questions. The types of question that I want to start asking are:

- What are the basic demographics for this group (ages etc)?
- How many of these people were active this year?
- What's the geographical distribution... which countries are light on mappers, which timezones are light on mappers?
- What's the geographical distribution of active people?
- How long do people stay before they drop out?

First, 80% of all data science (at the moment) is data cleaning, so I had to do a few things to make this possible.

Clean all the location information - there were two sets of location fields in the original data, and issues included:

- not listing which countries they were in (generally USA people assuming we all knew that Ohio was in the US),
- listing multiple countries (which is fair - development people often move between 2 or 3 'home' sites).
- America being a continent playing at being a country - it makes more sense to break US data into states, so we can see where on the continent people are distributed instead. Looking up the abbreviation for Minnesota, so the state column was consistent.
- People living in a dependency (e.g. an island like Madeira) of a country (e.g. Portugal).
- People who only gave their timezones as an address (also fair - it's a way round declaring that you're in a country hostile to mappers; also this only happened with US addresses).
- People also got confused about US timezones (and I had to look them up too): there are 4 timezones in the contiguous United States - they're called PST, MST, CST and EST.

European timezones are also less confusing than they look (unless you're working out whether and when summertime occurs):

(blues=GMT; pinks=CET;yellow=EET;orange=FET!; green=Moscow time)

So I now have an anonymised file (a lot of the work above was to get the address fields to a state where they don't give anything away), and start feeding it into Tableau Public (which is free, so you can follow along if you want...).

- First, I drop the "country" dimension into the middle of the tableau box - this automatically gives me a map of the world with a dot on every country mentioned.
- Then I select the "pie" mark types, and drop the "last visit" dimension onto "color" in the "marks" box. This turns the dots on the map into little pie charts, coloured for each year of "last visit".
- Then I play with the "size" button a little to get good-sized marks, fiddle with the colours a bit so they don't show "never visited" as green, and produce this:

But that's just telling me the percentage dropout rates per country. What about absolute rates? So I drop "number of records" onto the "size" box, and get

Okay. That's a lot of Americans. And a lot of countries with very few mappers in them. But maybe it's a lot of Americans because it's a big place... a continent labelled as a country. So before I stop looking at this data, I have a look at the numbers by US state... I click on country, then filter, and select only the USA, then I drop the "state" dimension onto the map, and exclude Hawaii (sorry guys: I know there are two of you over there, but it was messing up the map) to give:

Yay! Go New Yorkers! Or put less emphatically - there appear to be clusters of mappers, who might be local mapping groups. Looking at those neat pie charts, I start wondering what the growth rate is like in each country - i.e. how many people joined the group when. And about how long they stayed active after they joined. But first, a question about retention: plotting year and quarter of the mappers' first visit to the site against their last visit looks like this:

A few things to say about this. First, the list of people who've never visited the community site just stops after 2011 - probably because a site visit is part of the joining process. The expected batch of people who look at the site in the quarter they join, then ignore it after that are there (the diagonal line of bigger dots). But this just tells me who dropped out when... what I really want to see is a simpler graph of how long people have stayed, and whether the date they joined is related to this in any way. I can't quite get Tableau to do that yet, but I'm working on it...