

Youville Centre

Exploratory Data Analysis

Data For Good Ottawa

Kenneth Chu

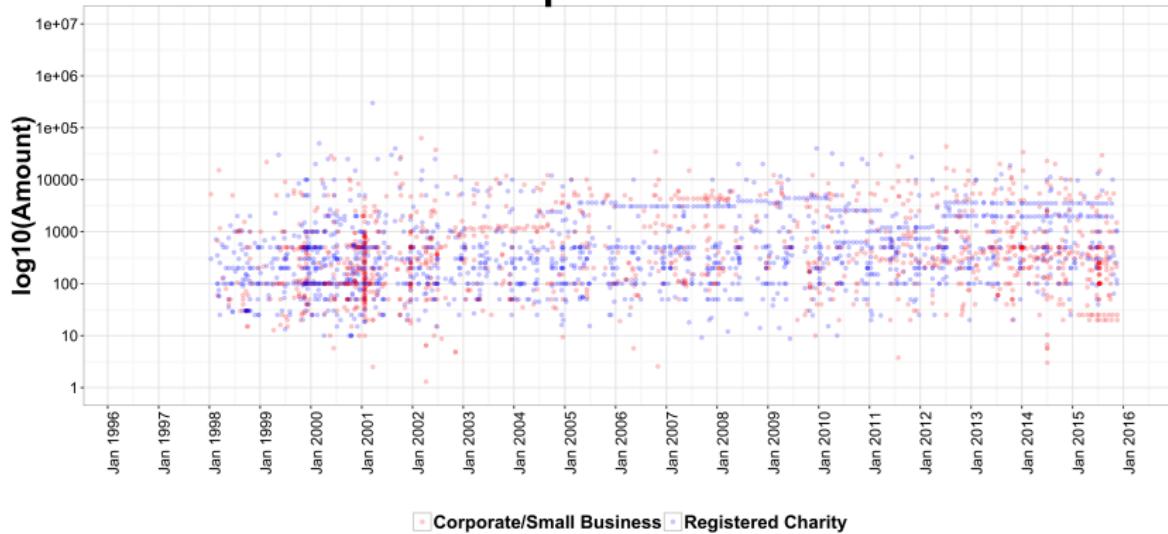
January 24, 2016

DepositItems or DonationReceipts?

- Thanks to the great discussion between Hasan and Robert, I was switching my analysis between DepositItems and DonationReceipts.
- I will refrain from making any conclusions at this point.
- Please consider the following slides as exploratory and comparative analysis of the Youville Centre data, in particular, a comparison between the data contents of DepositItems and DonationReceipts.

DepositItems or DonationReceipts?

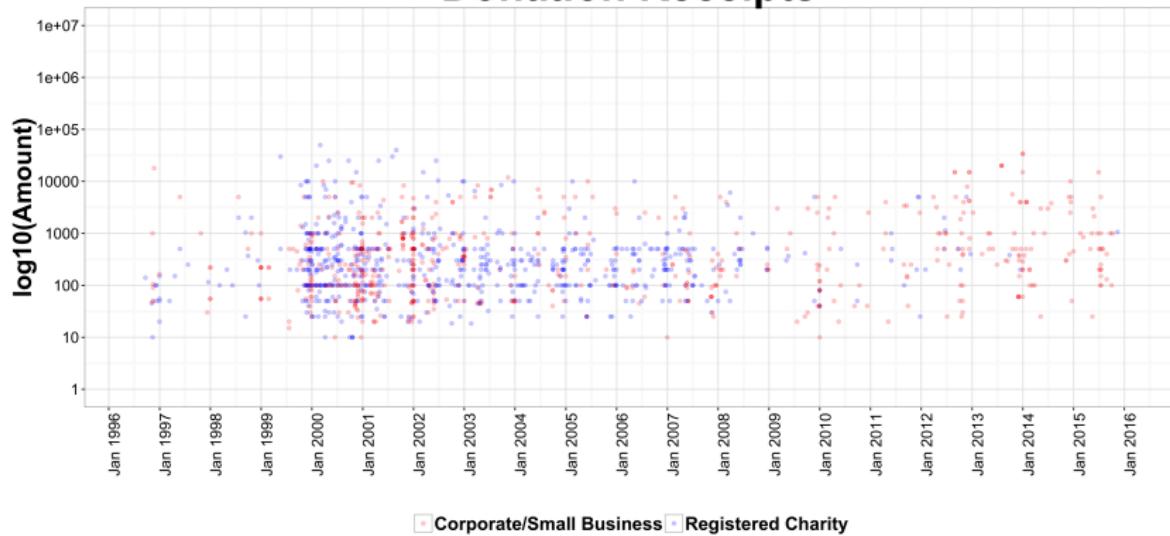
Deposit Items



Scatter plot based on rows from the table DepositItems with `Amount > 0` and `ContactTypeMain` being corporate/small business or registered charity. The dates are `DepositDate` obtained from left-joining the tables `DepositItems` and `Deposits` by `DepositNum`. `ContactTypeMain` was obtained by first joining `Contacts` and `ContactTypeMain`, and then left-joining `DepositItems` with the resulting table by `ContactID`.

DepositItems or DonationReceipts?

Donation Receipts



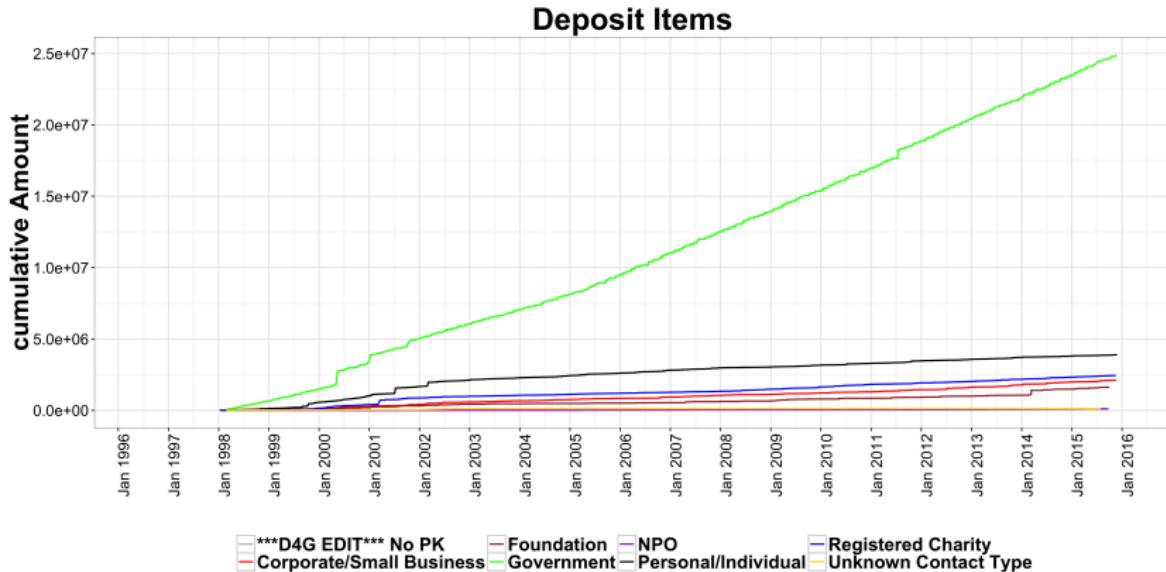
The sister plot of the preceding scatter plot based on data from the table `DonationReceipts`. The “thinning out” of receipts to registered charities, explicitly mentioned in Robert’s documentation of the data, appears visible in this plot.

Q: Does this mean we should NOT use `DonationReceipts` data to analyze trends of donations from registered charities?

DepositItems **cumulative donation** by ContactTypeMain

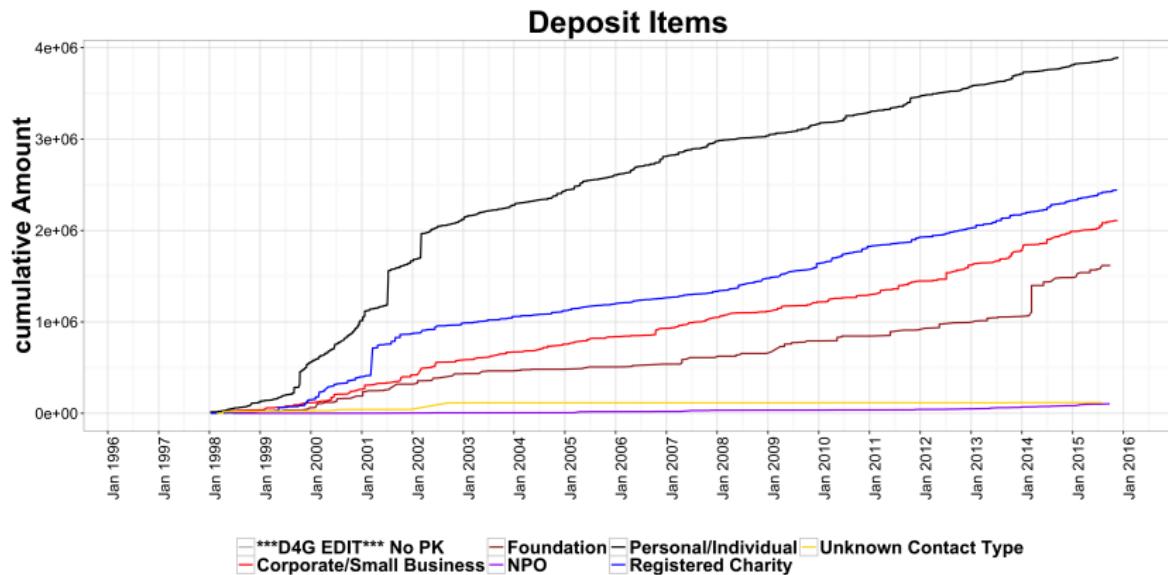
- Next, we look at plots of cumulative donation grouped by ContactTypeMain.
- These presumably should give some impression of how much donation Youville Centre has been receiving, and whether donations from the various contact types have been slowing down or speeding up.
- Comparing the plots based on DepositItems against those based on DonationReceipts also seems to indicate that the former does contain a lot more “stuff” than the latter (but, obviously, the extra stuff may or may not be relevant to the data analysis requested by Youville).

DepositItems cumulative donation by ContactTypeMain



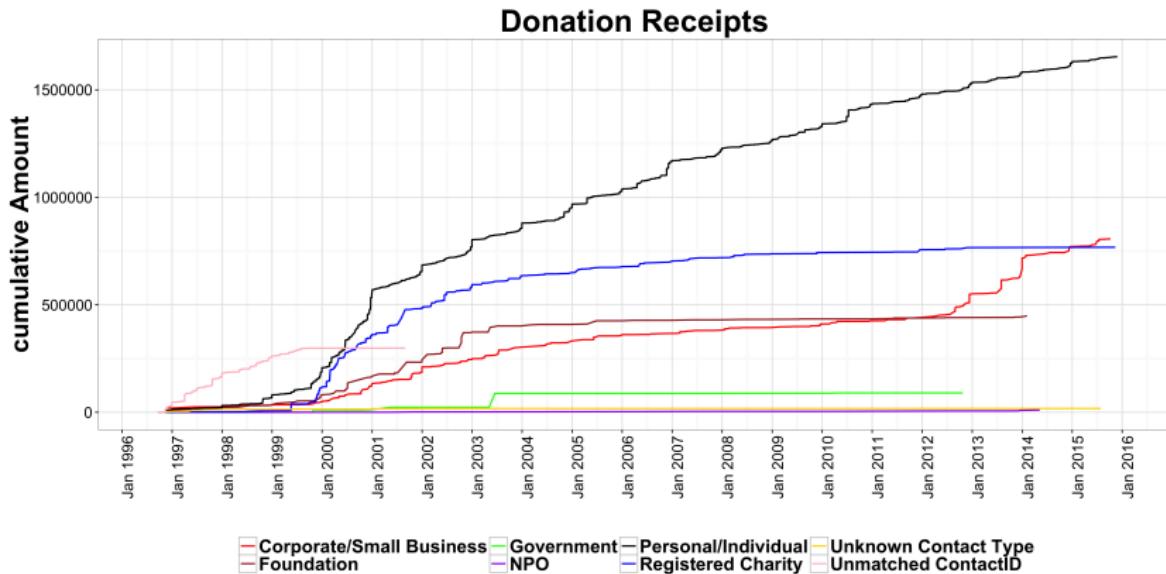
Warning: Many rows in the table DepositItems have negative values under Amount. The plot above is based on rows in the DepositItems with Amount > 0. In other words, I am tacitly assuming that all rows with Amount > 0 correspond to donations, and all rows with negative values under Amount are non-donations. Please let me know if I am mistaken, or perhaps we could verify with clients.

DepositItems cumulative donation by ContactTypeMain



Same as preceding plot but the ContactTypeMain = Government suppressed, in order to better visualize the cumulative donation profiles of the other contact types.

DonationReceipts cumulative donation by ContactTypeMain

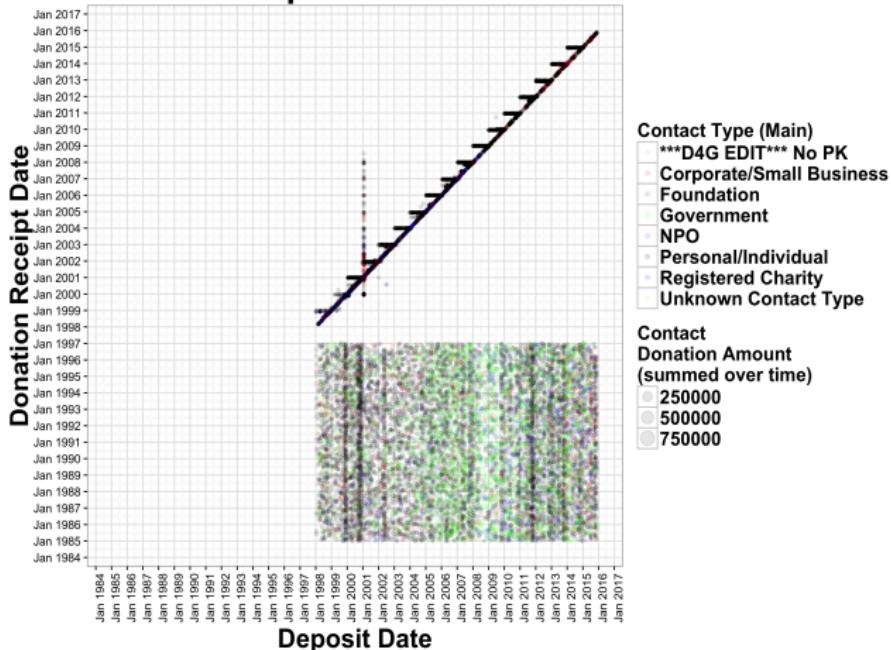


Cumulative donation plot based on DonationReceipts. Notable differences from DepositItems counterparts:

- significantly reduced vertical scale
- minute contribution of Government (among donation receipts)
- plateauing of Charity, Foundation and Government roughly after 2003
- presence of a small group of donations (pink line, labeled "Unmatched ContactID") apparently not captured in DepositItems (note the dates). The vast majority of the ContactID's of these rows on the DonationReceipts table are missing; those of the small remaining minority do NOT appear on the Contacts table.

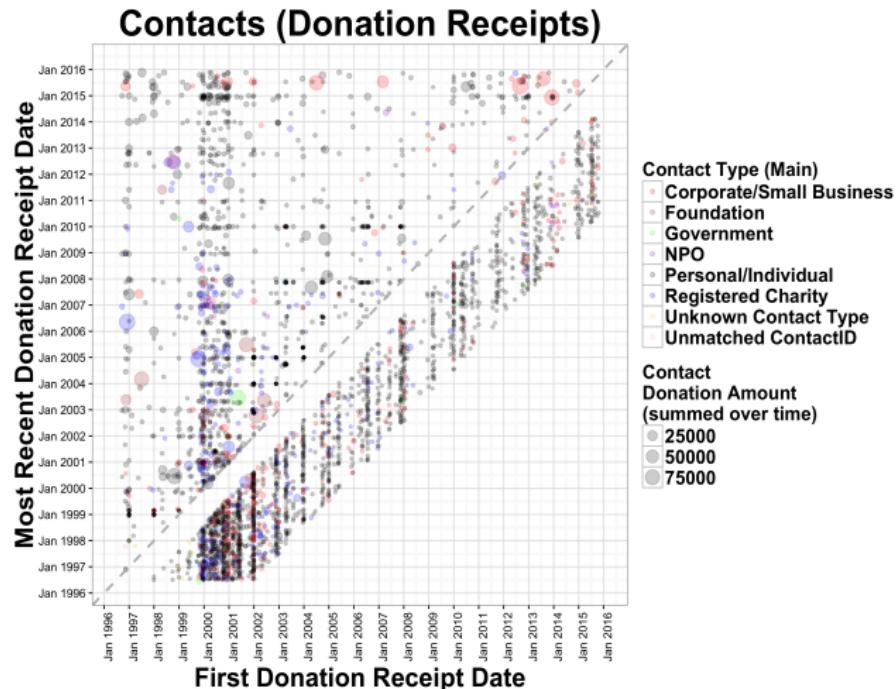
To impute or not to impute?

Deposit Items



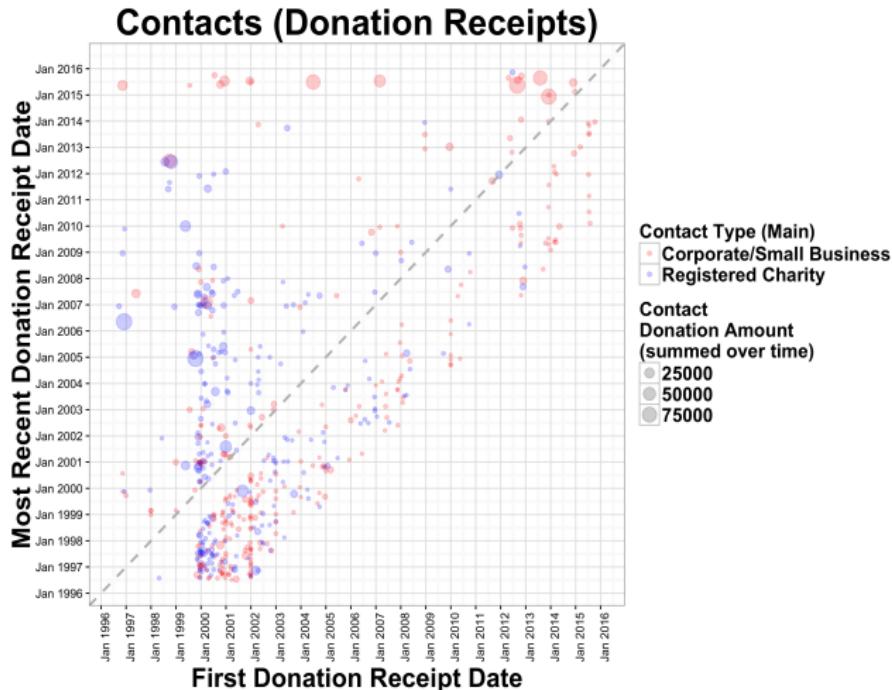
I guess this is a moot question by now, right? But since I have generated the plot already, let's have a look anyway: The above plot shows that DepositDate and DonationReceiptDate are very similar for those rows on DepositItems for which both dates are deducible (by joining tables). The large rectangular patch corresponds to rows on DepositItems whose DonationReceiptDate's are NOT deducible, and random dates were generated for them for visualization purposes.
Mini-conclusion: if we had to use DepositItems after all, using the DepositDate from Deposits is probably simpler and more reliable than imputing dates based on data from DonationReceipts.

Some retention plots à la Peter ...



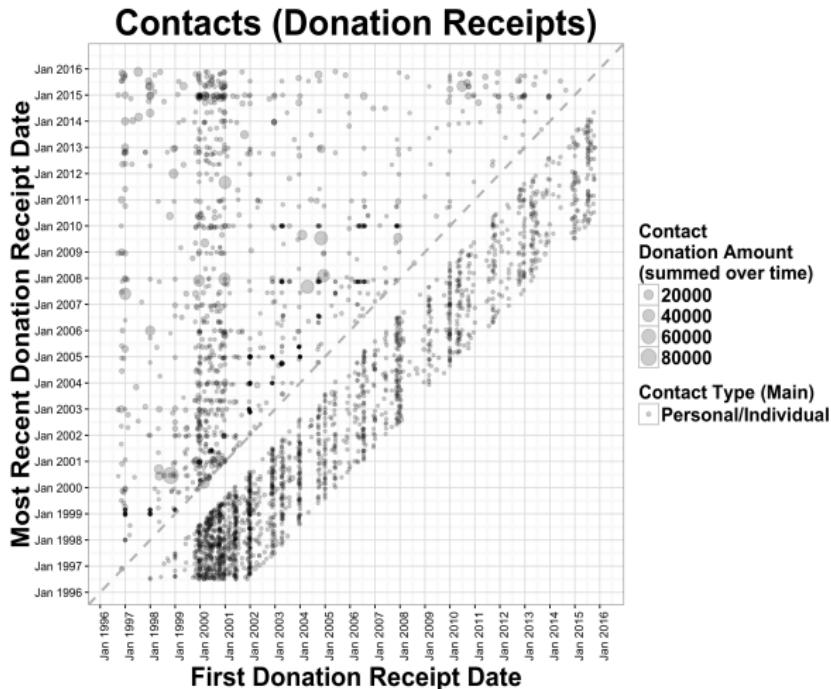
Each point corresponds to a unique ContactID on DonationReceipts and shows the first donation receipt date against the most recently receipt date of that ContactID. Colouring is based on ContactTypeMain while the bubble size indicates (not too reliably) the total amount of donation of that ContactID over time. All points should appear on or above the straight line of slope one. Those points now appearing below that line are the one-time donors with fake (random) most-recent donation dates for easy visualization (so, without tinkering, they would have lied perfectly on the slope-one line).

Some retention plots à la Peter ...



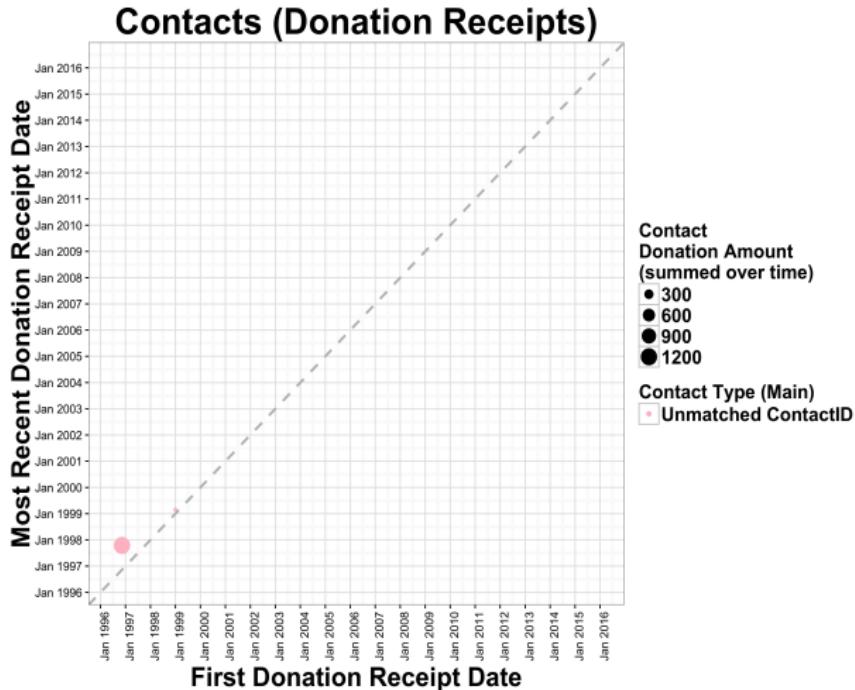
Same as preceding plot, but showing only rows with ContactTypeMain being Corporate/Small Business or Registered Charity, for easier visualization.

Some retention plots à la Peter ...



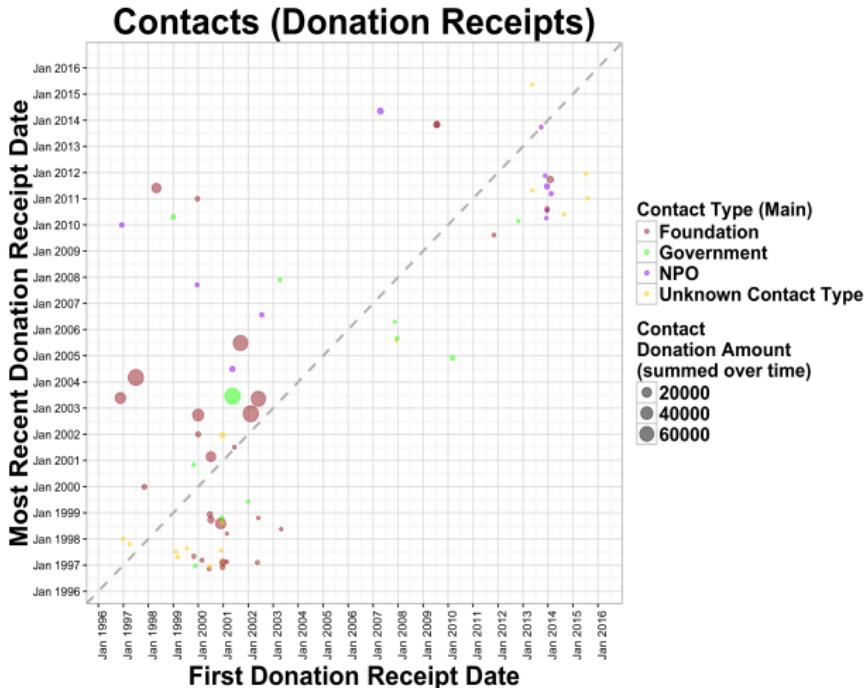
Same as preceding plots, but showing only rows with ContactTypeMain being Personal/Individual.

Some retention plots à la Peter ...



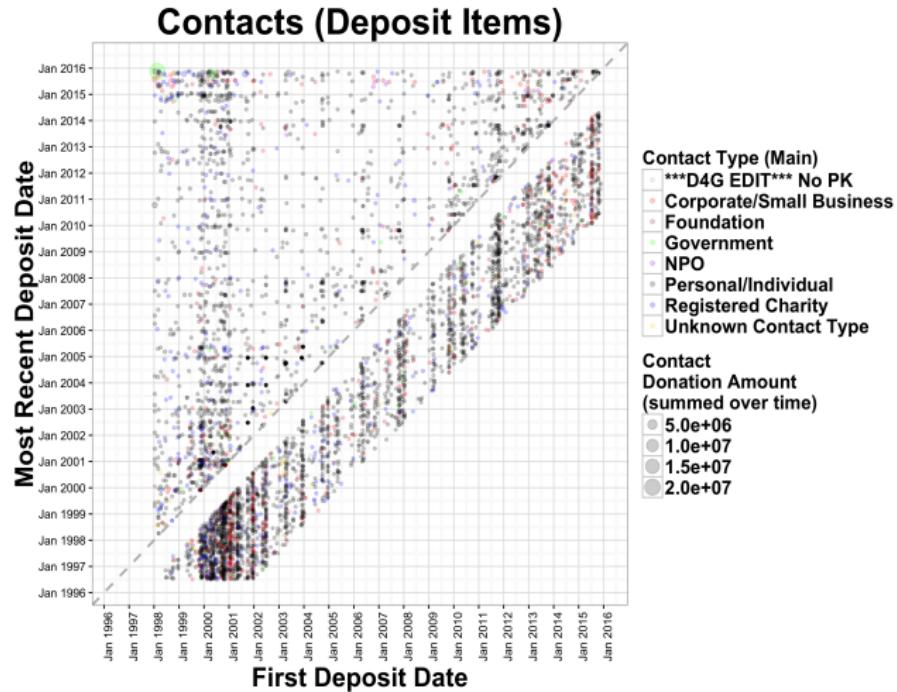
Same as preceding plots, but showing only data from rows on `DonationReceipts` whose `ContactID` are missing or do NOT appear on the `Contacts` table (hence, their `ContactTypeMain` cannot be deduced). The big bubble corresponds to the 932 rows on `DonationReceipts` with missing `ContactID`.

Some retention plots à la Peter ...

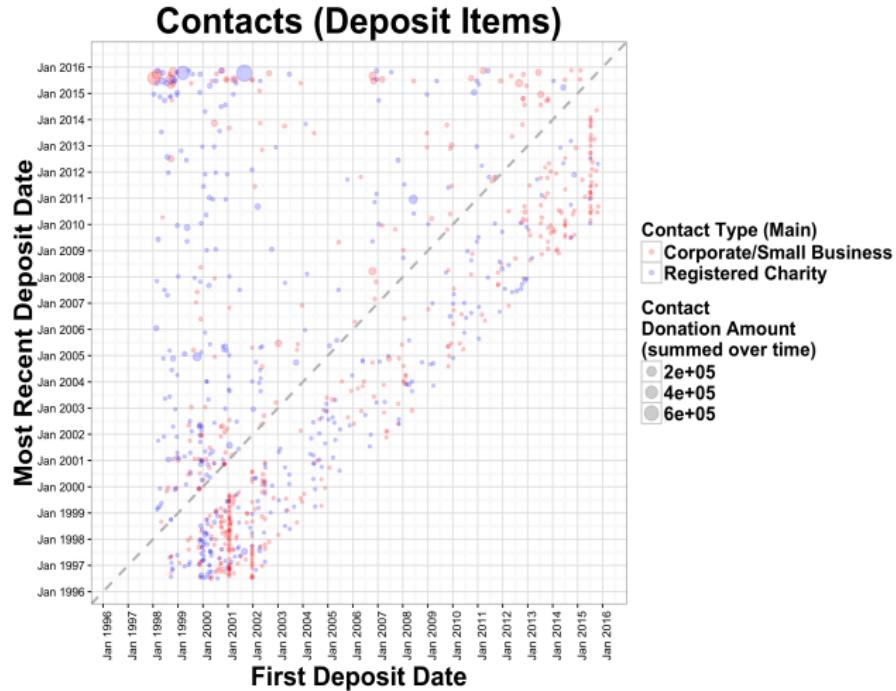


Same as preceding plots, but showing only rows with ContactTypeMain excluded in the preceding three plots.

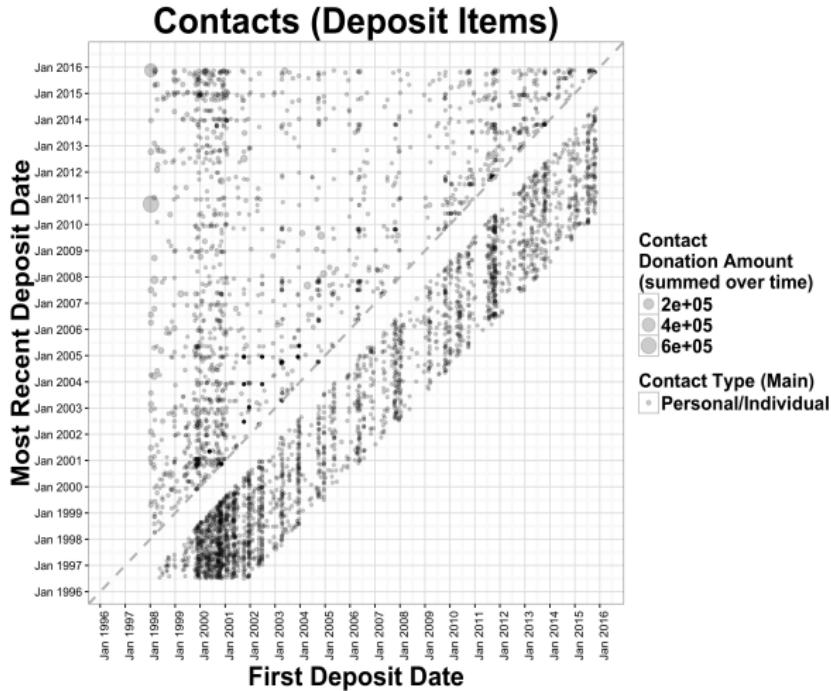
Retention plots based on DepositItems



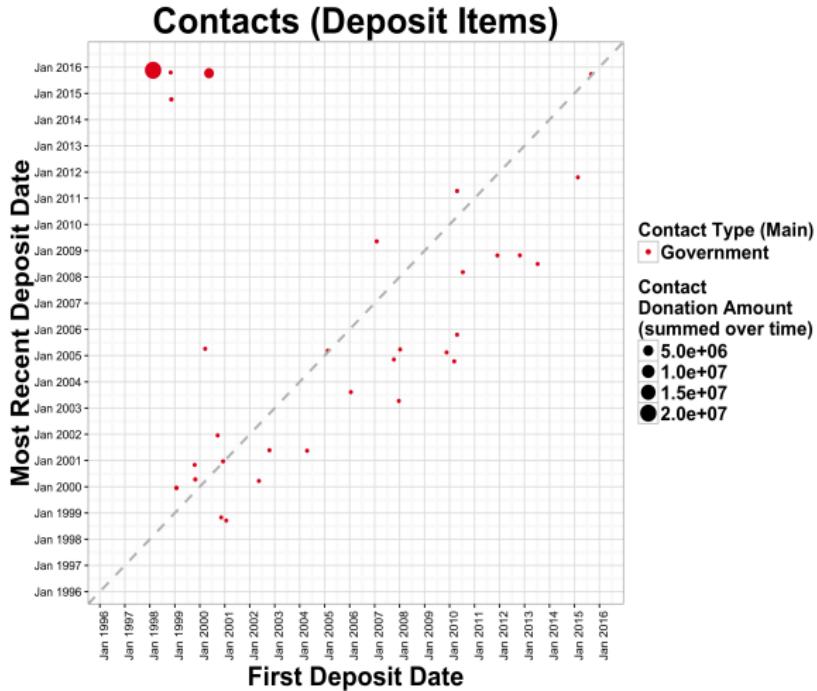
Retention plots based on DepositItems



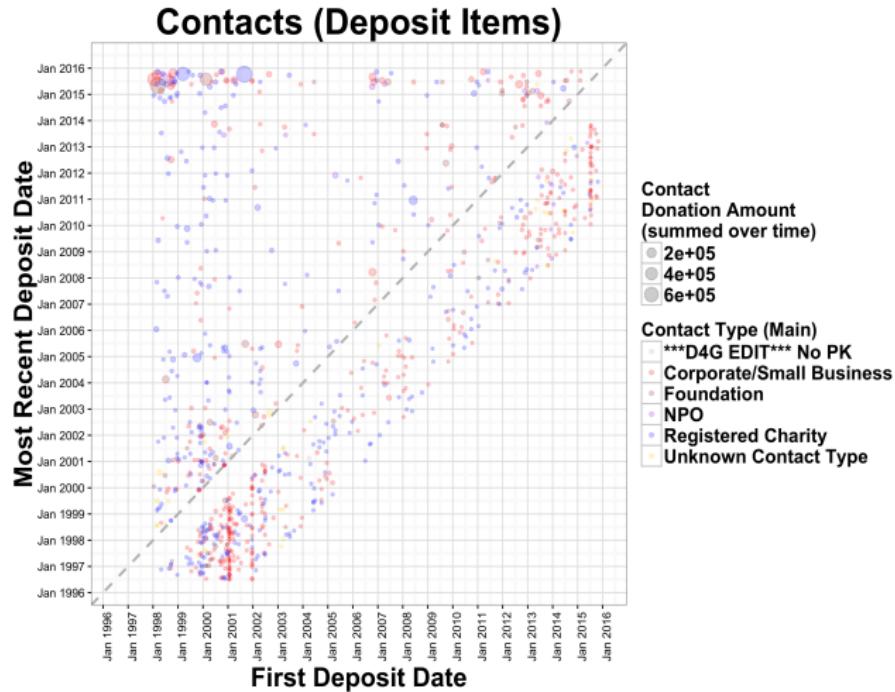
Retention plots based on DepositItems



Retention plots based on DepositItems

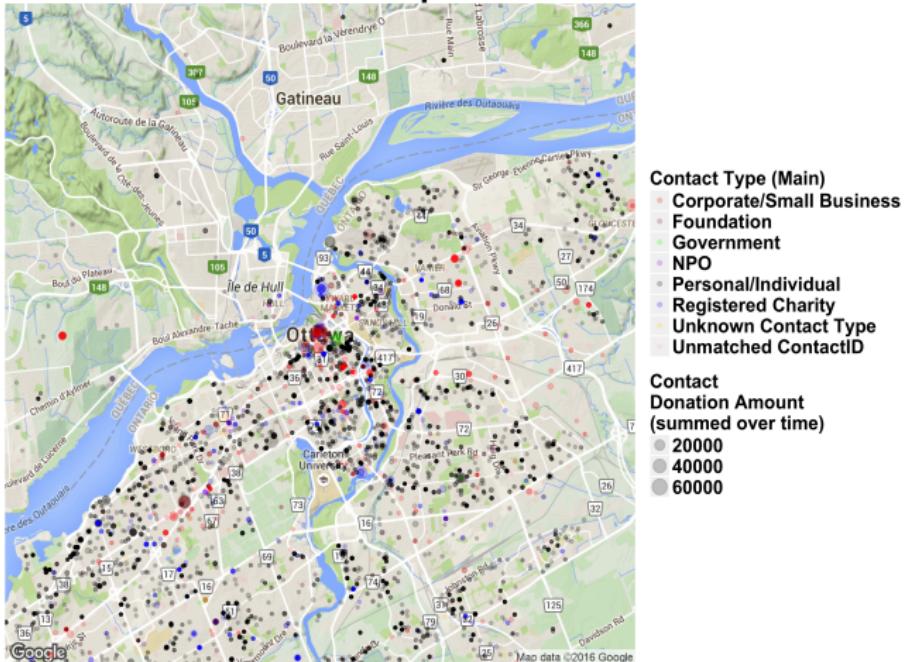


Retention plots based on DepositItems



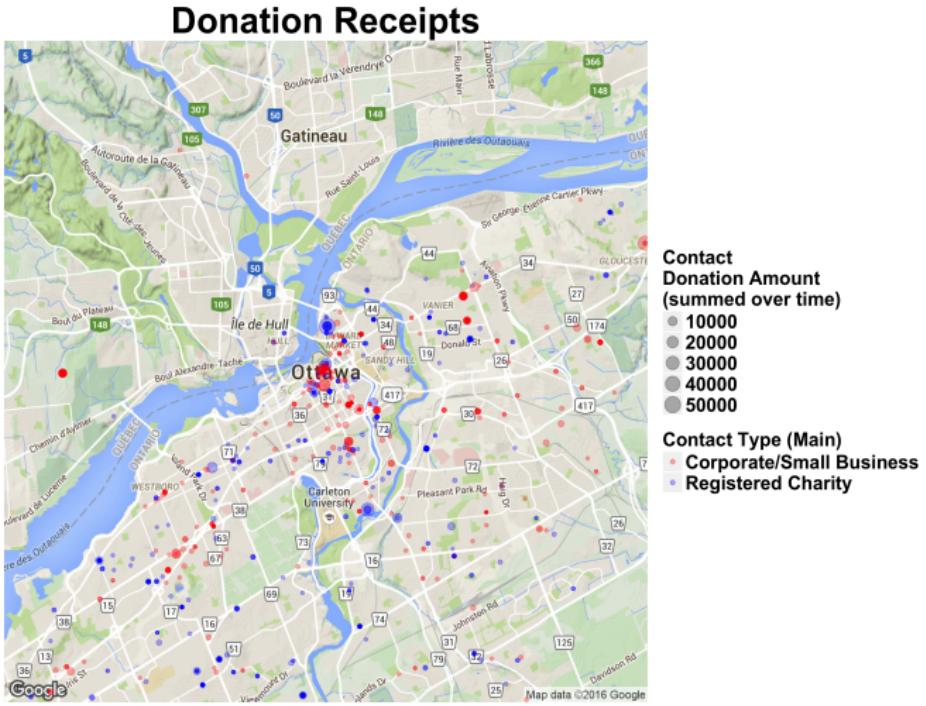
Mapping DonationReceipts data ... probably useless but definitely fun 😊

Donation Receipts



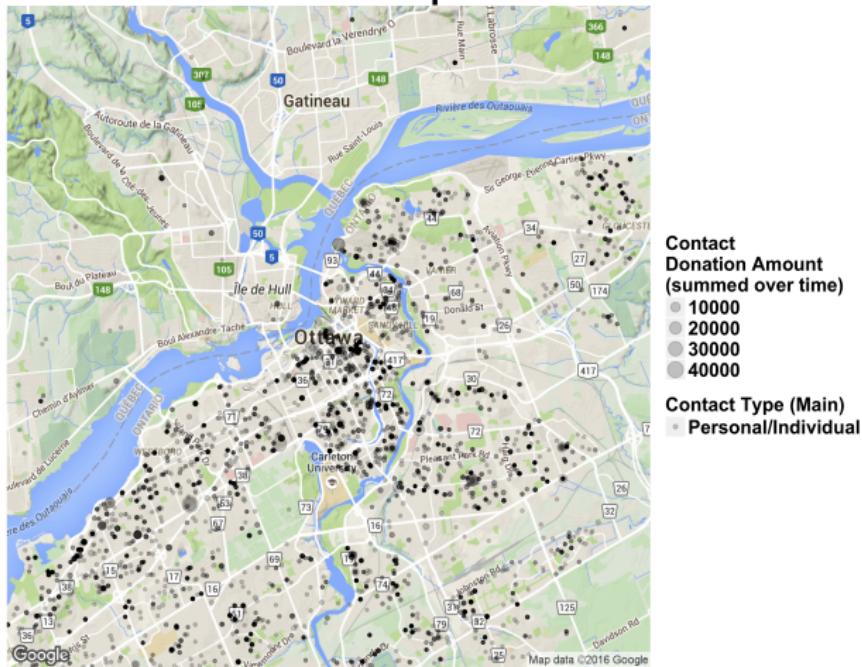
Longitude and latitude were obtained using the R command `ggmap::geocode()` based on the field `PostalCode` on the `DonationReceipts` table.

Mapping DonationReceipts data ... probably useless but definitely fun 😊



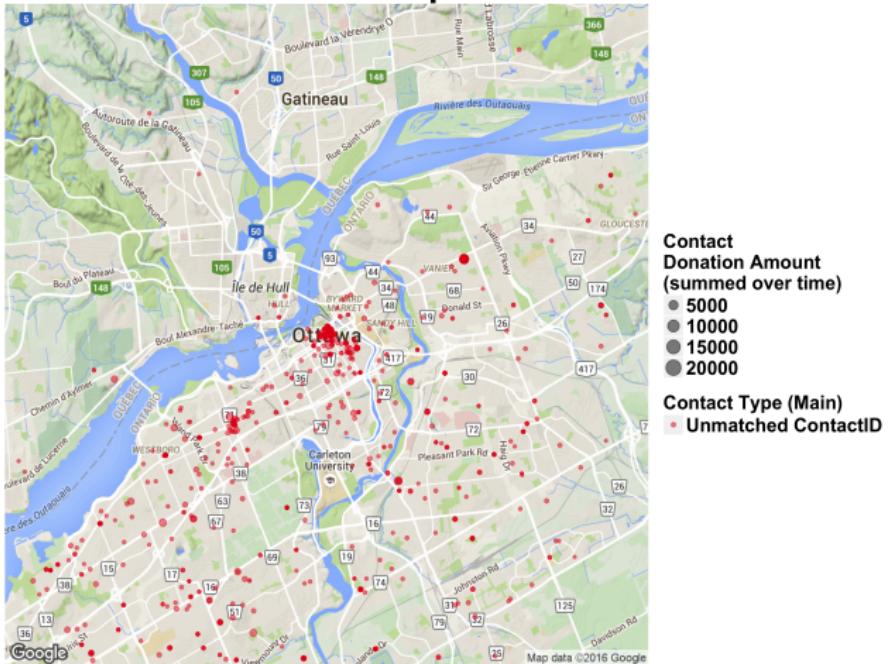
Mapping DonationReceipts data ... probably useless but definitely fun 😊

Donation Receipts



Mapping DonationReceipts data ... probably useless but definitely fun 😊

Donation Receipts



Well, those rows on `DonationReceipts` with non-matchable `ContactID`'s (either missing, or non-missing but not present on the `Contacts` table) probably didn't originate from the same physical real-world individual, in case there were ever any doubts 😊.