Merge Sum Stats

Sylvia Klosin

1/30/2018

Merge Summary Statistics

Filter

In this case we filter the patent and infutor datasets to include only

- People whose last name starts with "DI"
- Observations more recent then 2000
- People who live in the United States

Merge

We merge on last name, state, city and the first three letters of the first name. We have 3706 unique inventors that we are trying to find matches for. The corresponding infutor set has 2955052 unique person ids.

We find that doing this merge we get 1024 inventors that do not get any matches in the infutor data.

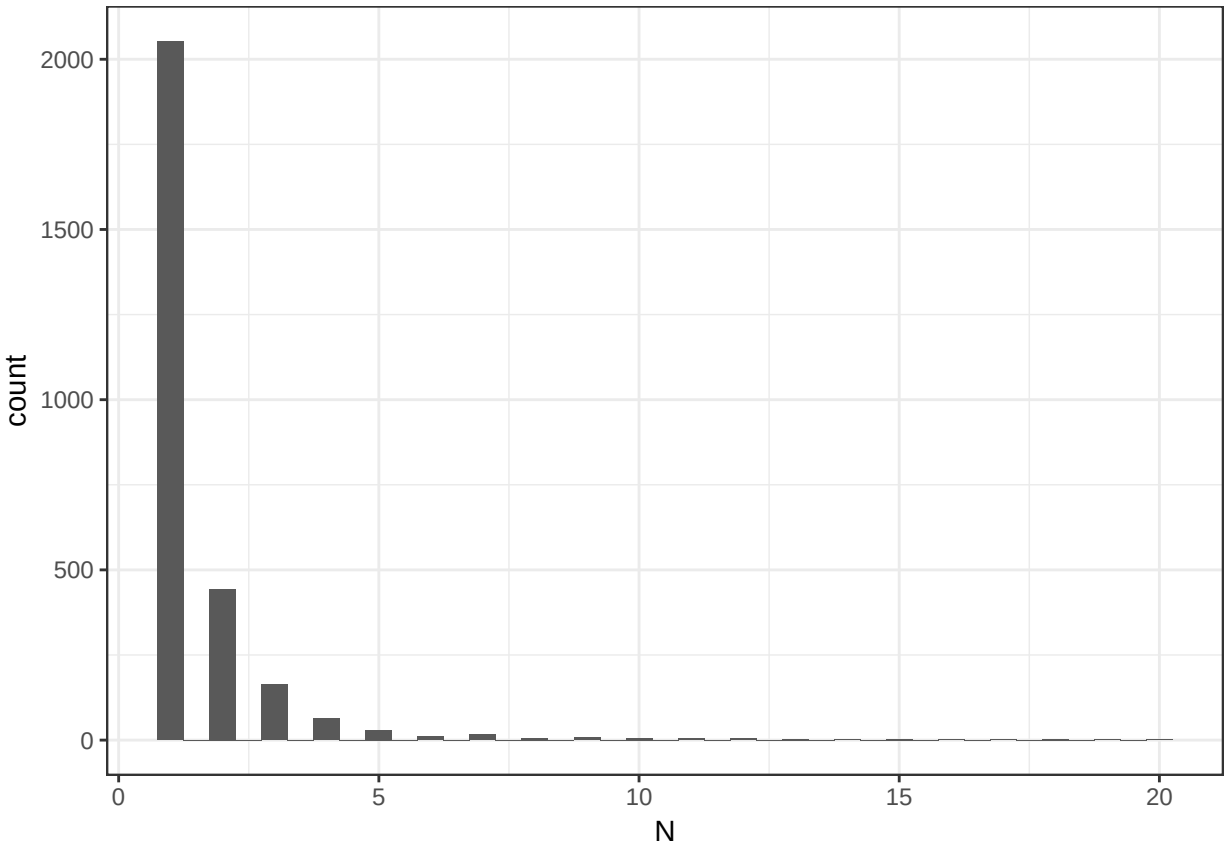
1-1 matches

There are 2053 inventors that get exactly one match.

For these 1-1 matches, we get 0.9400279 have the same clean first name. For those 7% where there is not an exact match, looking at the names show that they seem to be cases where nicknames are used e.g. Fredrick vs Fred.

$>$ 1 matches

For those with more than one match here is a graph that shows the distribution of the number of matches where we truncate the x-axis at 20. There are 26 people with more than 20 matches.



As a next step we bring in full first name information.

Adding in the first names only gives us 87 more matches.

ID problems

We have a problem with PIDS in the infutor data. There are people that are very clearly the same person but have different PID numbers. One such example is Kenneth Dicker from OH. Need to think of a way to reconcile this. For around 300 inventors, it is the case that only the PID seems wrong, and the birthdays and address are exactly the same. It also seems that for many of the rest it the case that PIDs are wrong, and there is a slight typo somewhere.

If we count the number of people that have the same date of birth, name, and exact address, but different PIDs we end up with 110 people.

Checking Matches 1-1 matches

Looking at dates

Seeing for the matches, is there at least one "true" year match. That at least one of the year/addresses in the patent data matches a year/address in the infutor data. We find that this is the case in 1593 cases. The flexibility given is 6 months. s

Looking at middle name

We still need to find the middle name variable in the infutor data.

Checking Matches > 1 matches

Looking at dates

When looking at the many matches, filtering by year only gives us 229 more 1 on 1 matches.

number_of_mathes	number_of_people
1	2053
2	444
3	164
4	64
5	30
6	11
7	16
8	6
9	8
10	6
11	4
12	5
13	3
14	1
15	3
16	2
17	1
18	3
19	1
20	2
21	2
24	1
25	1
27	1
29	1
34	1
35	1
36	1
37	1
40	1
44	2
54	1
58	1
61	1
65	1
75	1
82	2
101	1
136	1
186	1
235	1
360	1
732	1

pid	dob	add_id	add	add_city	add_state	name_first	name_last
68862832	196808	25989658	133 WILLIAMSBURG CT	CHAGRIN FALLS	OH	KENNETH	DICKER
68862832	196808	224560002	3519 STOER RD	SHAKER HEIGHTS	OH	KENNETH	DICKER
304329969	196808	25989658	133 WILLIAMSBURG CT	CHAGRIN FALLS	OH	KENNETH	DICKER

	V1
FALSE	471
TRUE	1593