# Merge Sum Stats

*Sylvia Klosin*

## 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 3103215 unique person ids.
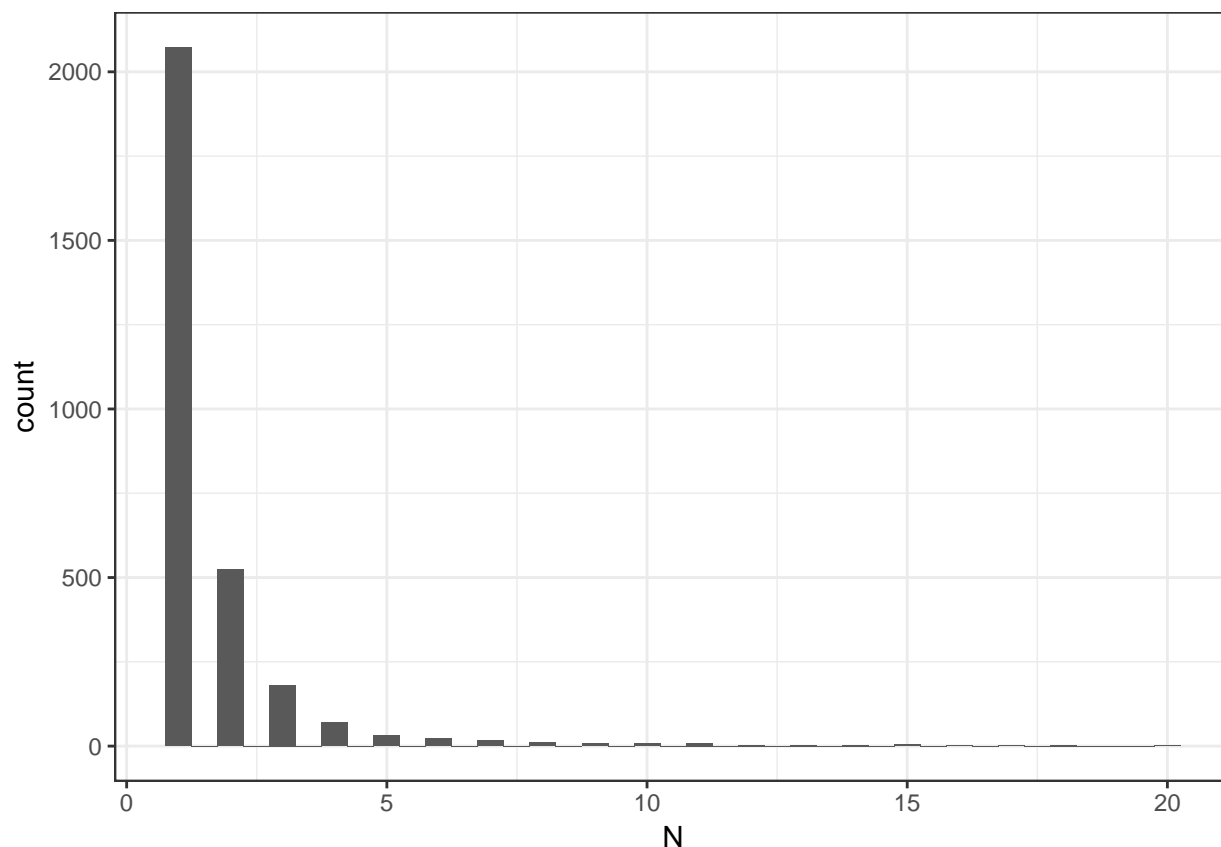
#### 1-1 matches

We find that doing this merge we get 700 inventors that do not get any matches in the infutor data. We were able to lower this number a few different ways. 1) we expanded the infutor time period 2) we cleaned last names by removing things like "JR" and "III".

There are 2073 inventors that get exactly one match.

For these 1-1 matches, we get 0.9395848 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 30 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 80 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 163 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 294 more 1 on 1 matches.

| number_of_mathes | number_of_people |
|---|---|
| 1 | 2073 |
| 2 | 524 |
| 3 | 181 |
| 4 | 70 |
| 5 | 32 |
| 6 | 23 |
| 7 | 17 |
| 8 | 11 |
| 9 | 7 |
| 10 | 8 |
| 11 | 9 |
| 12 | 3 |
| 13 | 3 |
| 14 | 3 |
| 15 | 4 |
| 16 | 1 |
| 17 | 2 |
| 18 | 3 |
| 20 | 2 |
| 21 | 2 |
| 22 | 1 |
| 23 | 1 |
| 24 | 1 |
| 25 | 1 |
| 28 | 1 |
| 30 | 1 |
| 33 | 1 |
| 35 | 1 |
| 37 | 1 |
| 39 | 1 |
| 40 | 3 |
| 45 | 1 |
| 51 | 1 |
| 62 | 1 |
| 67 | 1 |
| 68 | 1 |
| 69 | 1 |
| 84 | 1 |
| 92 | 2 |
| 114 | 1 |
| 166 | 1 |
| 208 | 1 |
| 271 | 1 |
| 397 | 1 |
| 822 | 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 |
| 68862832 | 196808 | 198208621 | 2765 HAMPSHIRE RD APT 204 | CLEVELAND | OH | KENNETH | DICKER |
| 68862832 | 196808 | 222995957 | 8362 RIVERSIDE DR | NOVELTY | OH | KENNETH | DICKER |
| 304329969 | 196808 | 25989658 | 133 WILLIAMSBURG CT | CHAGRIN FALLS | OH | KENNETH | DICKER |