EDA draft

Ricky Heinrich & Team

2023-05-11

Outline

- Overall summary: how many variables, how many missing values
- Exploring missing values: are they related to population?
- Exploring data distributions
- Population zero

Overall Summary

There are 489 676 rows in this data and 41 columns, meaning that 489 676 dissemination blocks are included. The 41 columns include information about the dissemination blocks themselves such as ID, population, and coordinates, as well as information about other census boundaries like dissemination areas, census areas, and provinces. Each of the 10 amenities have two columns associated with it: one a binary indicator to track whether the amenity is present in the DB itself, and the other the calculated proximity measure. Finally there are three indicators: transit_na, amenity_dense, and suppressed.

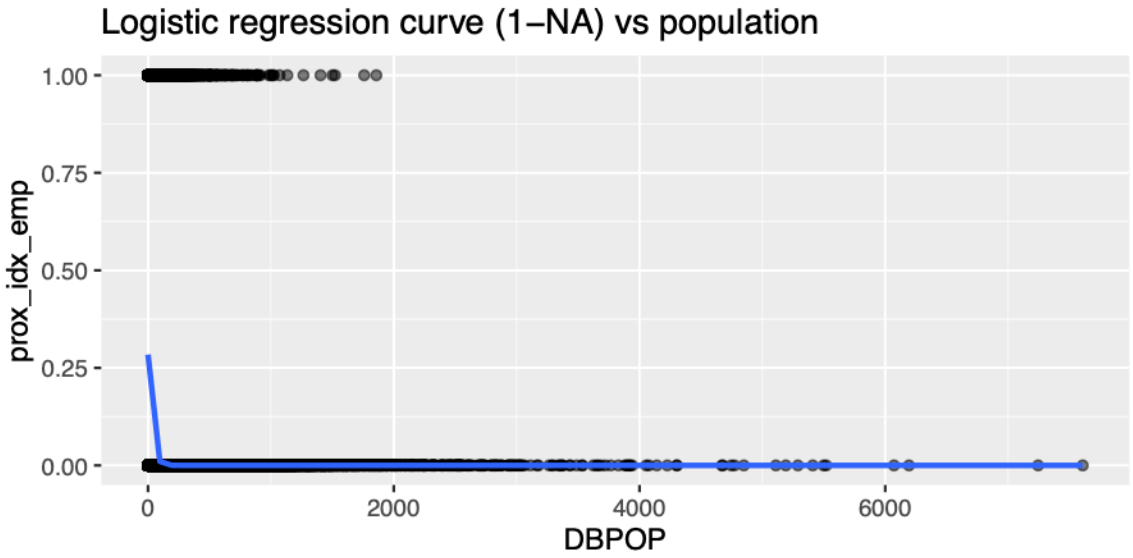
In the summary of the dataset, we see that there are many missing values. We see that the library proximity indicator contains the most missing values, at around 77%, followed by the proximity measures for grocery and secondary education. Only two out of the ten amenities have proximity measures missing proportion under 50%: health and employment.

```
##       prox_idx_lib   prox_idx_grocery   prox_idx_educsec    prox_idx_pharma
##        76.99397152        71.19258448        71.16195198        63.54303662
##   prox_idx_transit   prox_idx_educpri     prox_idx_parks prox_idx_childcare
##        62.97449742        53.97793643        52.19941349        50.17848537
##             CMAUID             CMAPUID              CMAPOP    prox_idx_health
##        43.48058716        43.48058716        43.48058716        38.64003954
##       prox_idx_emp          in_db_emp       in_db_pharma    in_db_childcare
##        13.49341197         1.09603084         1.09603084         1.09603084
##       in_db_health      in_db_grocery      in_db_educpri      in_db_educsec
##         1.09603084         1.09603084         1.09603084         1.09603084
##          in_db_lib        in_db_parks      in_db_transit      amenity_dense
##         1.09603084         1.09603084         1.09603084         1.09603084
##              DBPOP              DAPOP              CSDPOP               DBUID
##         0.06432825         0.06432825         0.06432825         0.00000000
##              DAUID              CSDUID             CSDNAME            CSDTYPE
##         0.00000000         0.00000000         0.00000000         0.00000000
##            CMANAME            CMATYPE               PRUID              PRNAME
```

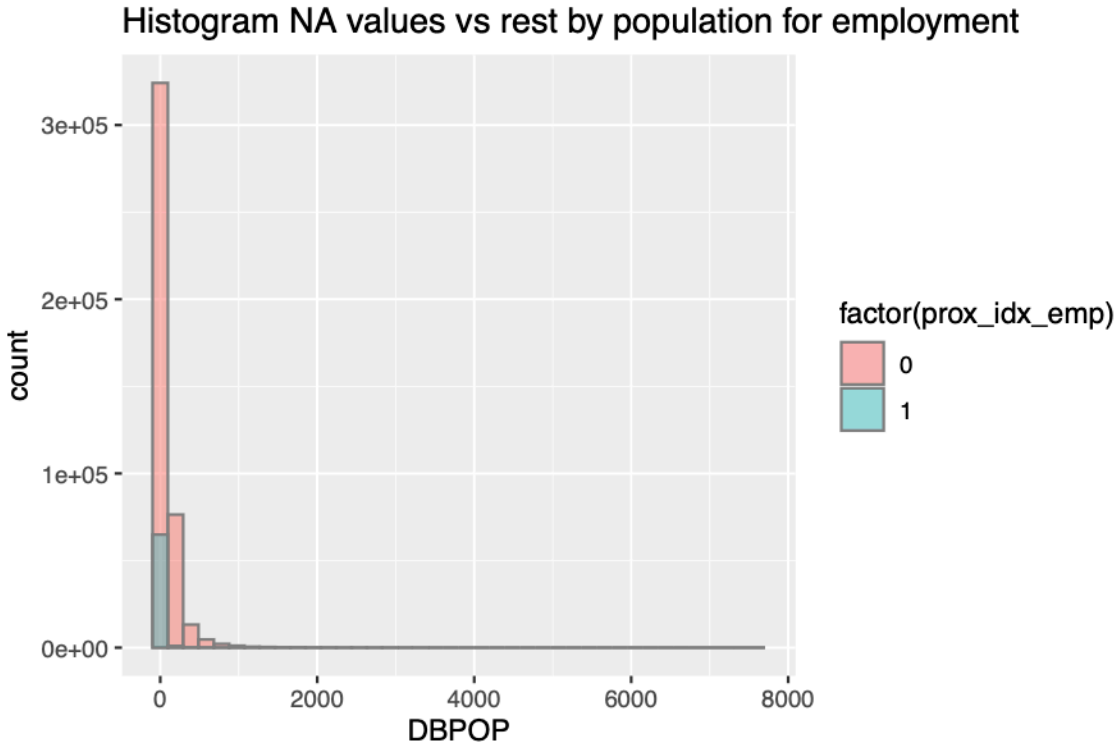
```
##          0.00000000          0.00000000          0.00000000          0.00000000
##               PRPOP                 lon                 lat          transit_na
##          0.00000000          0.00000000          0.00000000          0.00000000
##          suppressed
##          0.00000000
```

In the following chart, can see the proportion of missing values for each amenity by province. We see that overall, employment has the lowest rates of missing values, but has also more range depending on the province. Ontario and Quebec seems to have the least missing values for most amenities relative to the other regions, whereas Nunavut usually has the most. It seems like the amount of proximity measure missing for libraries are the most consistent across regions.

For each amenity, we can plot the occurrence of missing values in a DB vs its population, and plot a basic logistics curve. The employment curve is included below, and the remainder are found in the appendix. We see that for some amenities, like employment and health, the missing values are concentrated among DBs with small populations. These are the same amenities with less than 50% of values missing. Overall it seems like the population of the DB is not the only factor, if at all, affecting whether a proximity measure is missing for that DB.



We can also plot the histograms of missing values vs populations for each amenity, where '1' (blue) is a missing value and '0' (pink) is a value not missing. Again, the remainder are found in the appendix.

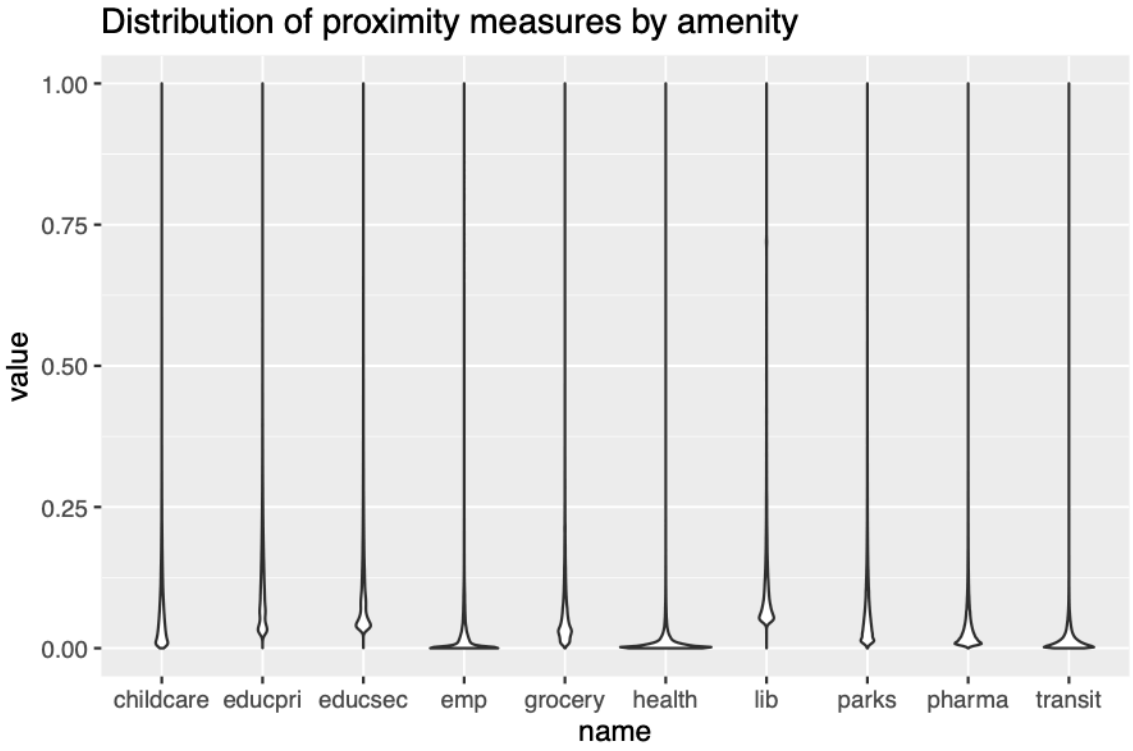


Overall, there are more DBs with lower populations than larger populations. We see that for some amenities, at smaller populations, there are a lot more missing values. Again, employment and health are the only two where there are always more actual values than missing values at every population bin.

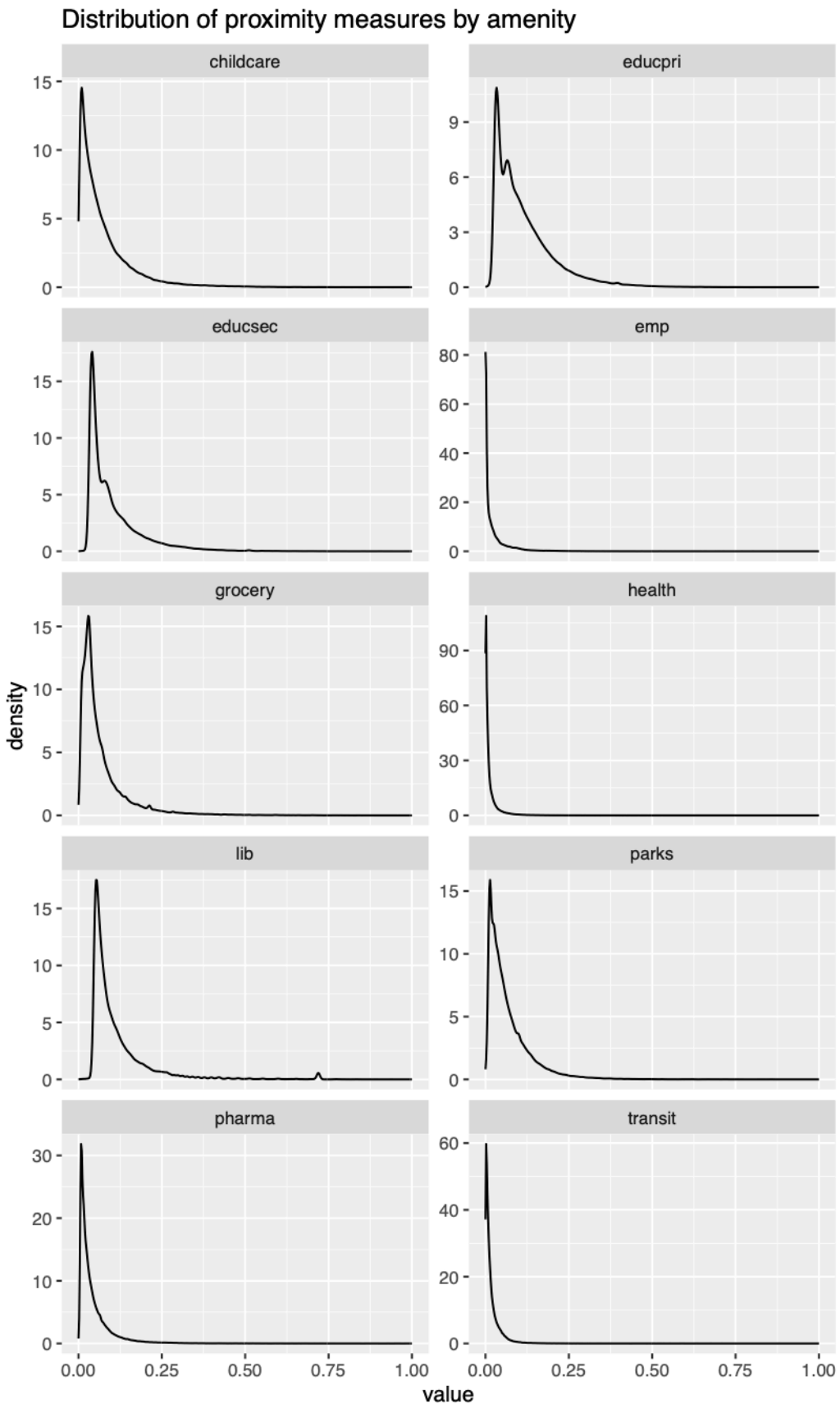
Data Distributions

We can take a preliminary look at the distribution of proximity measures for each amenity, to see if there are 'obvious' clusters.

In this violin plot, we see that the highest densities of proximity values lie below 0.12 for all amenities. We see that the amenities with the highest distribution density closer to 0 are health, then employment, then transit. Library has the lowest density right around zero, and 'starts' a bit later. Health and employment have the least amount of missing values, and library has the most; some conclusion could be made out of that.



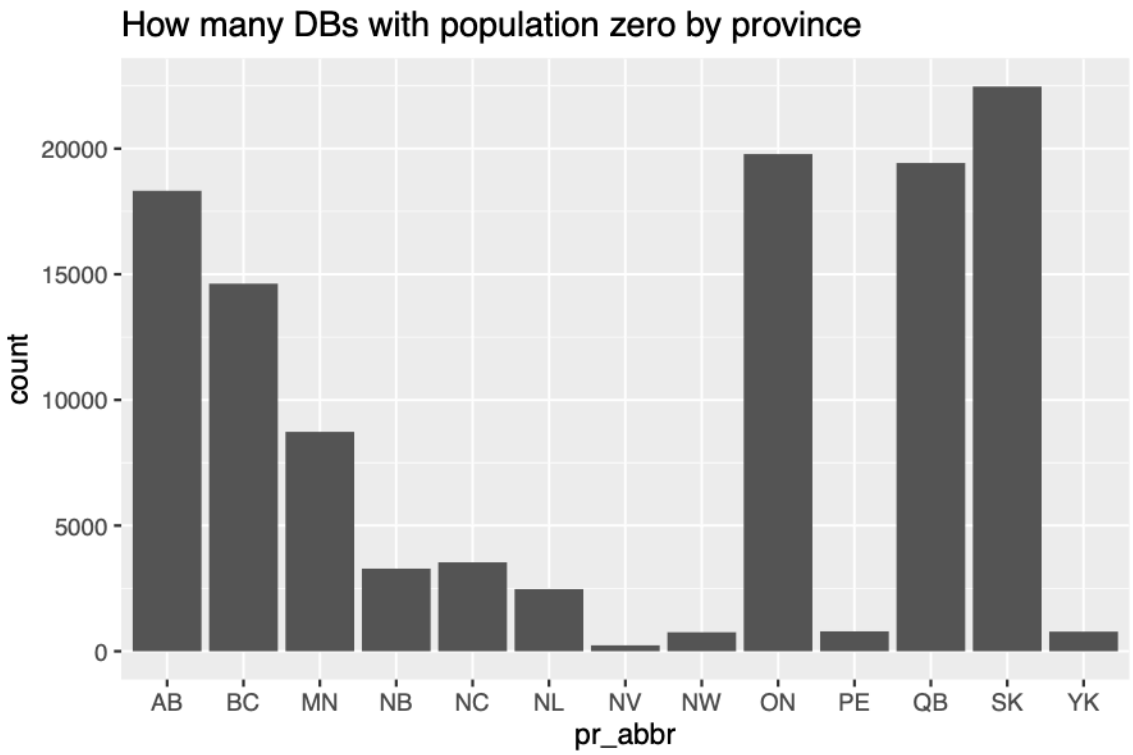
Next we see the kernel densities of proximity measures for each amenity. We see that most curves appear smooth, but some like for primary education, secondary education, and library, have 'bumps', which could indicate clusters. Overall, the naked eye is not able to perceive obvious segmentation cutoffs.


Distribution of proximity measures by amenity
childcare
educpri
educsec
emp
grocery
health
lib
parks
pharma
transit
density
value
0.00
0.25
0.50
0.75
1.00

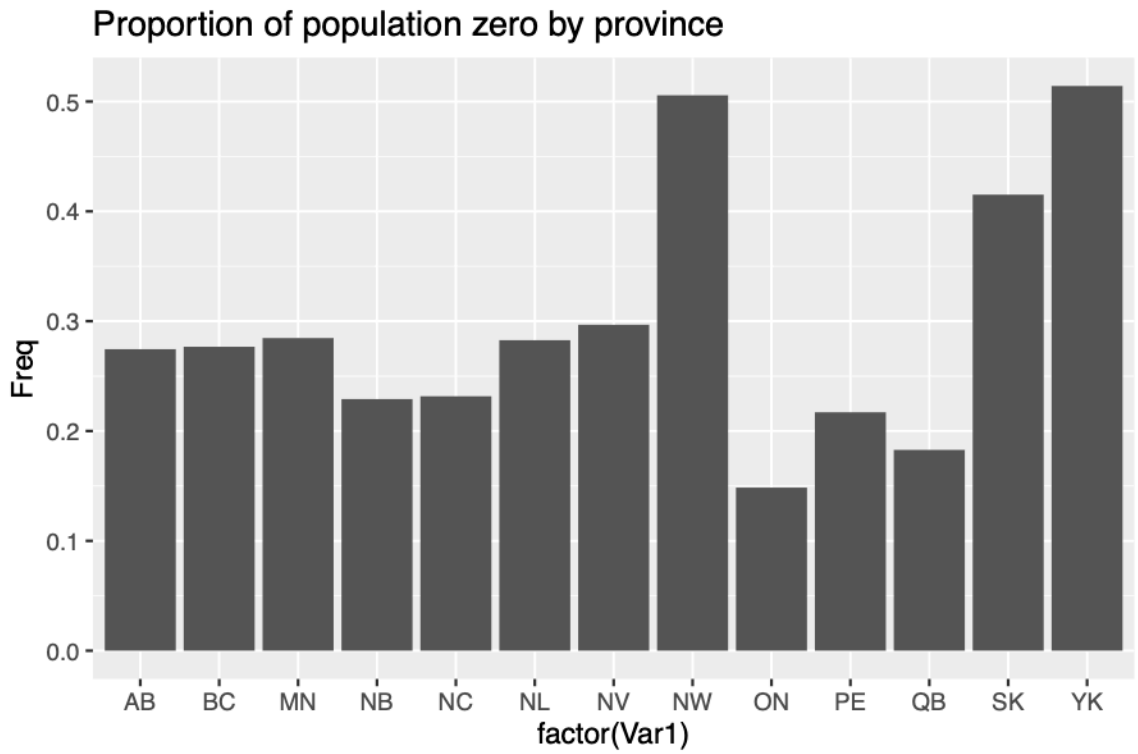
Population zero

About 24% of the DBs in Canada (in 2016) have a population of 0. It could be reasonable to expect that if the population of a DB is 0, then the proximity measure are also near 0: it is intuitive that for the most part, amenities are further away from areas with no populations. It is thus reasonable to explore the cases where the population is zero, to see its prevalence, and deduce how it may affect the values of proximity measures. In the appendix there's a barplot showing how many DBs there are per province: Ontario and Quebec have the most, whereas the Territories have the least.

Here, we that the province with the most DBs with a population of zero is Saskatchewan, followed by Ontario, Quebec, and Alberta.



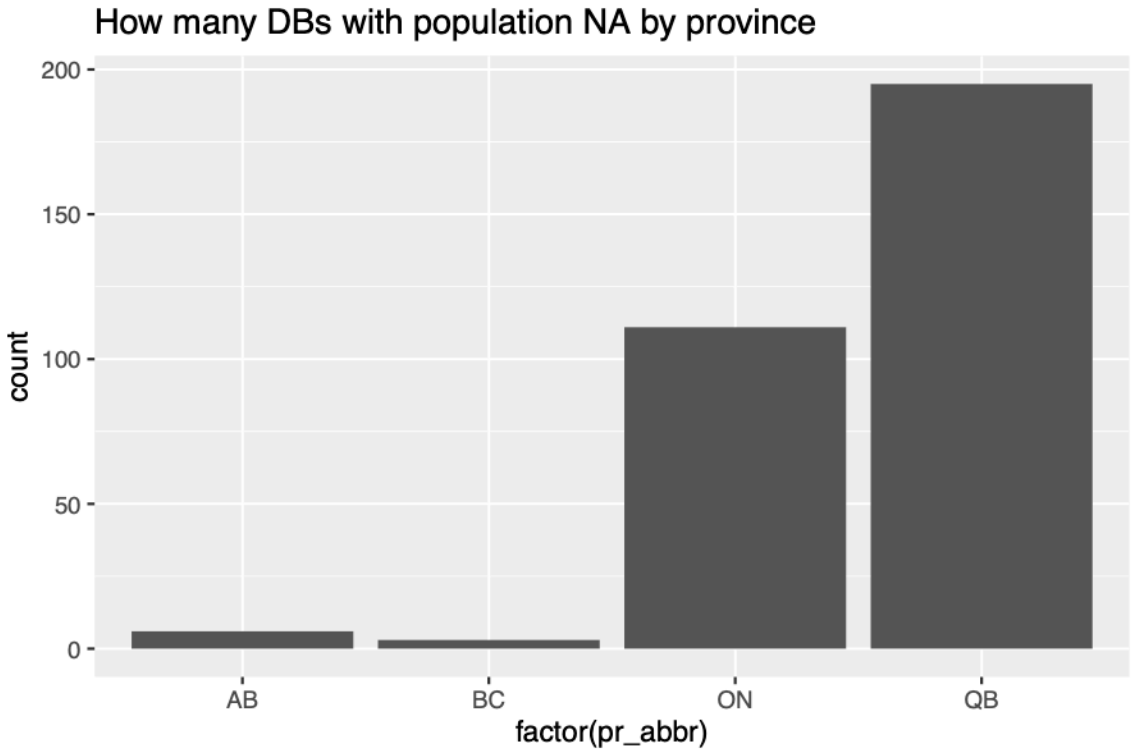
Taking the proportions however, we see that over 50% of Yukon and NWT's DBs have a population of 0, and Saskatchewan has over 40%. Ontario has the lowest at around 15%, followed by Quebec at around 18%.



In the following table, we see that the CSDTYPE with the most populations = zero are rural municipalities followed relatively closely by cities. It seems the majority of the top counts are urban areas (cities, villes, municipalities, etc), which is somewhat unexpected.

```
##
##    RM    CY     V    MD     T   RDA    MU    MÉ    NO    TP     C   IRI    VL
## 20926 18638 12454  9859  8423  7077  4723  4568  4013  3569  3023  2613  2425
##     P    DM    CV    SC   RGM   SNO    SM    PE    SA   LOT    TV   HAM    CT
##  1634  1582  1309  1277  1231  1127   701   671   594   525   429   307   304
##    RV    ID    SV     M    VN   RCR   SET    NV   COM    SÉ    NL   IGD    TC
##   147   138   100    93    80    75    72    62    51    51    44    42    37
##    CG    VC    NH   S-É    TI    CC   LGD    CU    SG    CN    IM    TK    TL
##    36    31    27    27    24    20    18    14    11     9     6     2     1
##    VK
##     1
```

Some of the population is NA. We see that Quebec has the most DBs with a population NA, followed by Ontario, Alberta, and BC. The CSDTYPE of the DB's whose population information is NA are IRI – Indian reserve and S-É – Indian settlement.



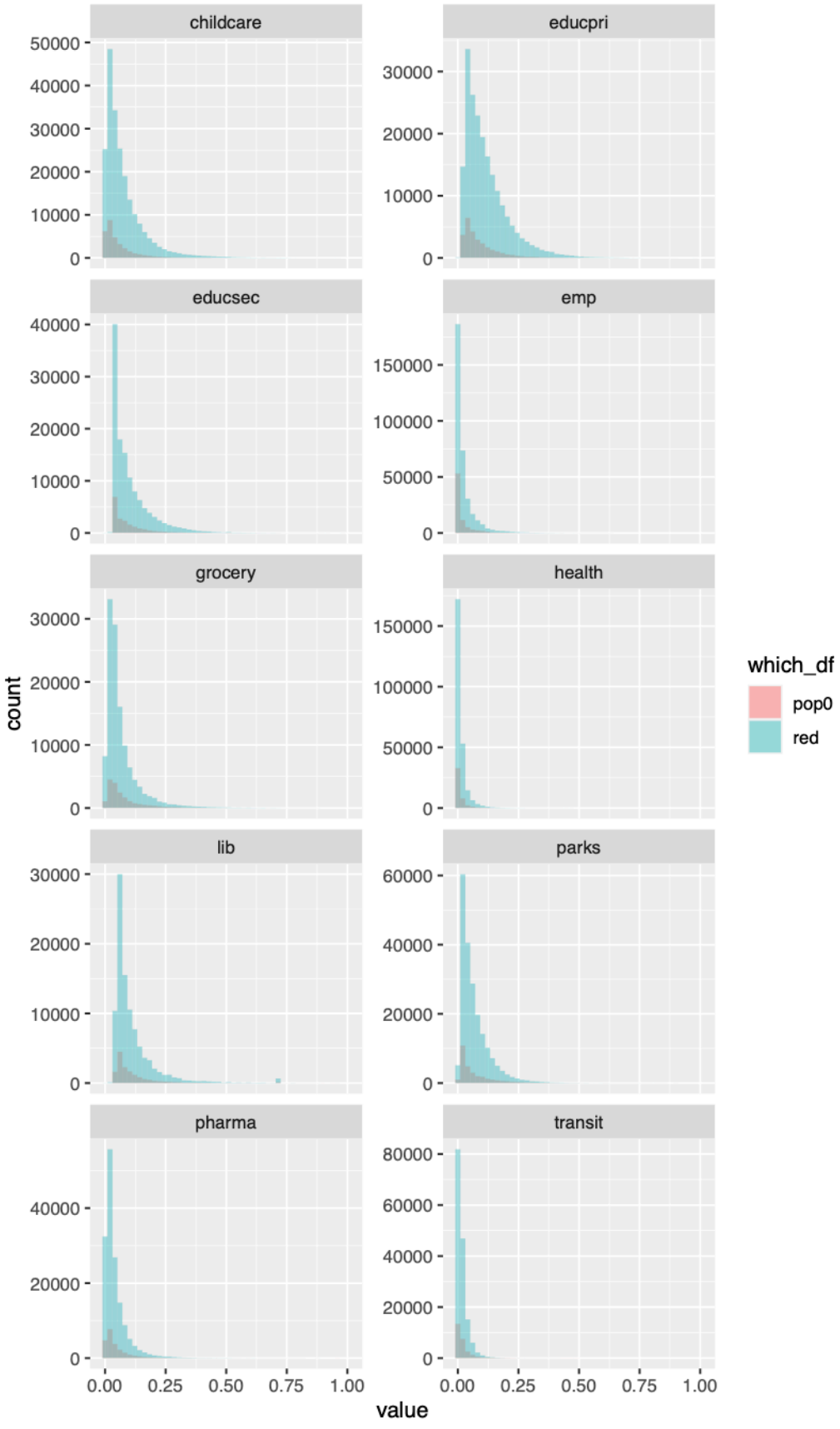
Effect of removing population = 0

The null hypothesis of the Kolmogorov-Smirnov test is that the two samples come from the same distribution. In this following table, we compare the 'sample' where the population = 0 vs the rest. We see that the p-values are very small for every amenity, thus leading us to conclude that we have sufficient evidence to say that these 'samples' don't come from the same distribution. (We can conclude that there is an effect on the proximity measures when the population is 0 ?)

```
##       amen_cols_short amen_pval
##  [1,] "emp"           "0"
##  [2,] "pharma"        "0"
##  [3,] "childcare"     "0"
##  [4,] "health"        "0"
##  [5,] "grocery"       "0"
##  [6,] "educpri"       "0"
##  [7,] "educsec"       "0"
##  [8,] "lib"           "1.44686656478044e-05"
##  [9,] "parks"         "0"
## [10,] "transit"       "0"
```

But in what ways do these subsets differ?

Here we are comparing the histogram for both, where the pink represents the count of population = 0, and the blue the rest. We see that the 'pink' appears to mirror the trends of the 'blue', but on a smaller scale. In the appendix, a 'zoomed in' plot is available. Surprisingly, see there that for some higher proximity 'bins' in transit and health, there are more cases for when population = 0. We also see in the appendix the kernel densities.


childcare
educpri
educsec
emp
grocery
health
lib
parks
pharma
transit
count
value
0.00
0.25
0.50
0.75
1.00
which_df
pop0
red

From this following table, we see that 72% of the proximity measure values where population = 0 are NA, compared to 50% of those where population !=0.

```
##
##            FALSE      TRUE
##   pop0 0.2761597 0.7238403
##   red  0.4980350 0.5019650
```

Conclusion

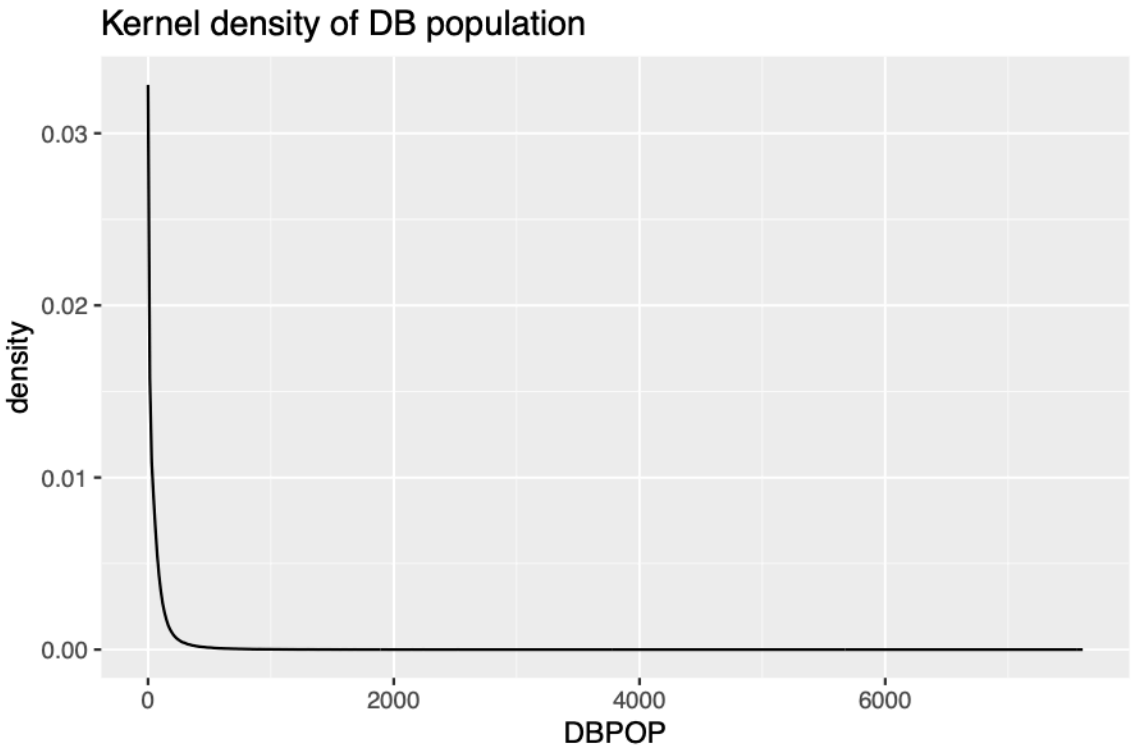
There are no obvious clusters in the proximity measures to the naked eye.

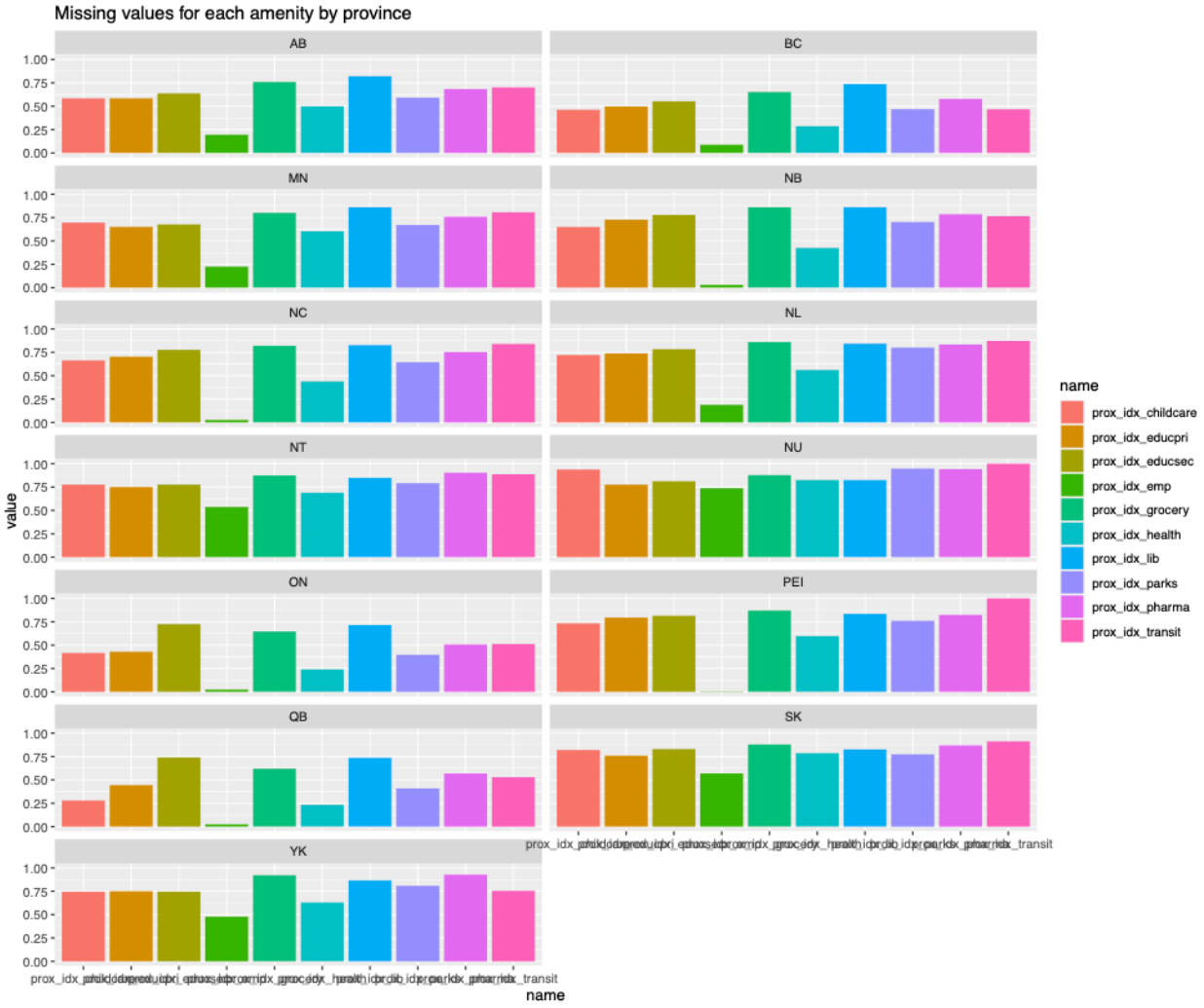
The distribution of missing values is not the same across amenities nor provinces.

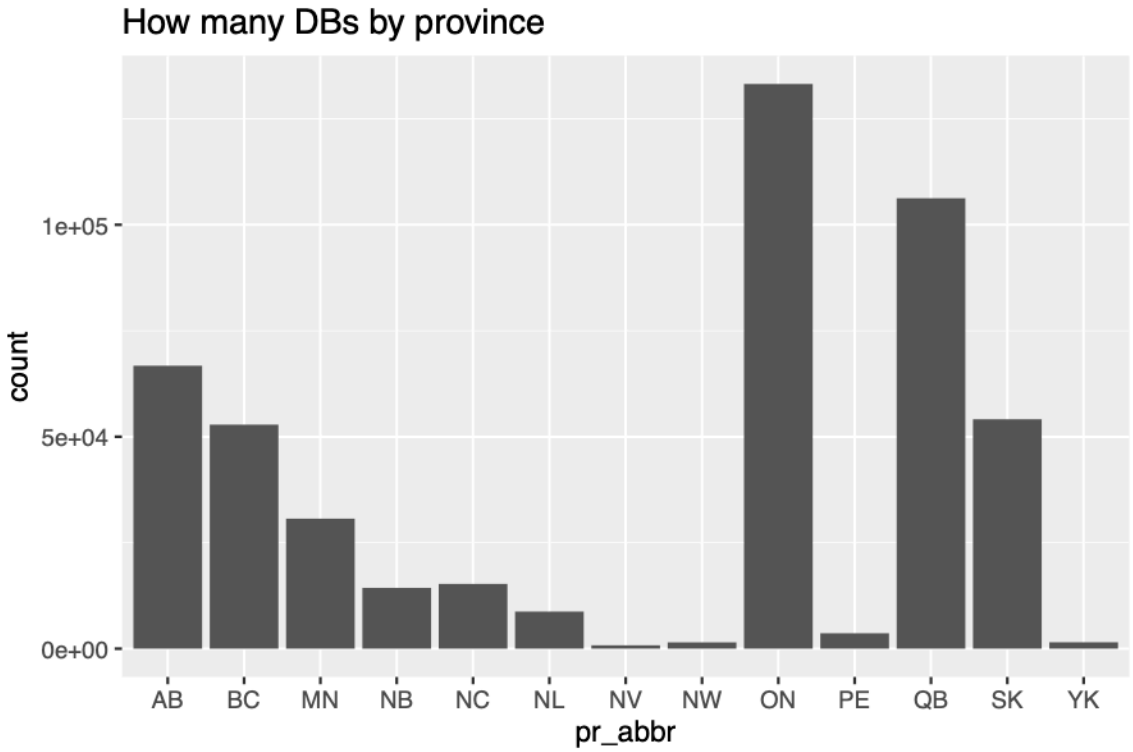
Our dataset contains a lot of rows, and there may be question about the 'usefulness' of all of them. If we were to remove all the DBs where the population is 0, we could reduce our dataset by 23%, aiding computationally. There are still proximity measures associated with these DBs with population 0: the distributions of their proximity measures are not the statistically the same as those for the rest of the DBs (those with populations), but the trends appear somewhat similar.

It doesn't appear that population of a DB is the only factor affecting whether a proximity measure is missing (NA); it was the only one tested as it was the only one included in this dataset. According to StatsCan's definition of a DB however, "only population and dwelling counts are disseminated at the dissemination block level" anyways. If we wanted to analyse other factors, we would have to look into aggregation at a higher level, which is not straighforward (need to take the mean/etc of whether something is missing or not?).

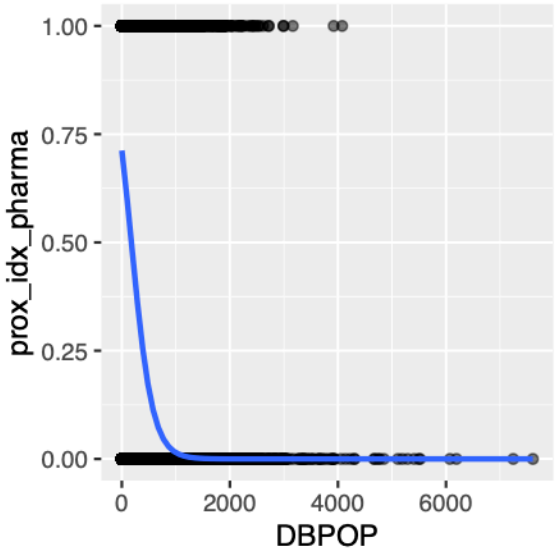
Appendix

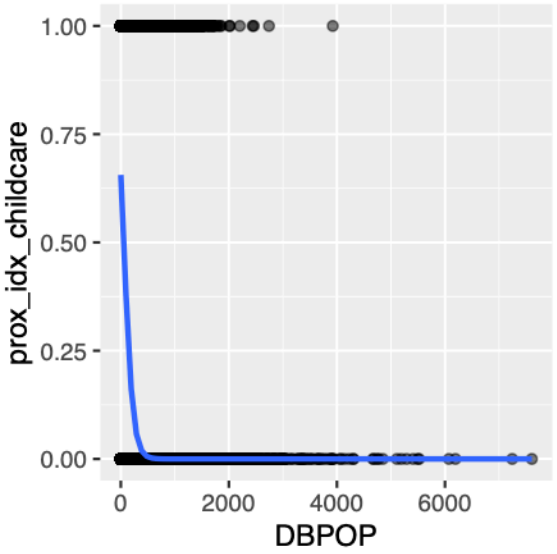

Kernel density of DB population
density
0.03
0.02
0.01
0.00
0
2000
4000
6000
DBPOP

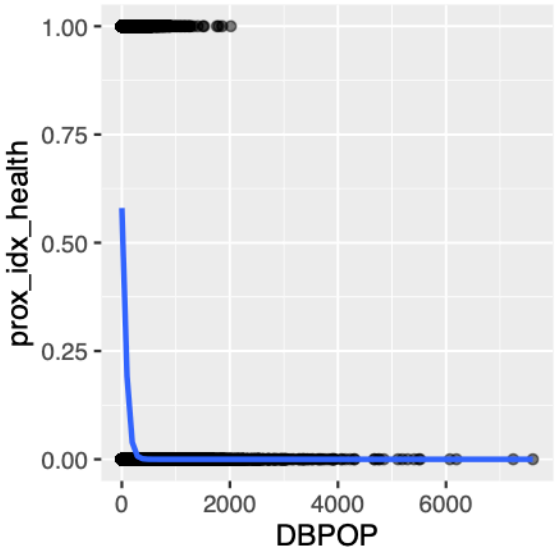

Missing values for each amenity by province
AB
BC
MN
NB
NC
NL
NT
NU
ON
PEI
QB
SK
YK
1.00
0.75
0.50
0.25
0.00
value
name
prox_idx_childcare
prox_idx_educpri
prox_idx_educsec
prox_idx_emp
prox_idx_grocery
prox_idx_health
prox_idx_lib
prox_idx_parks
prox_idx_pharma
prox_idx_transit

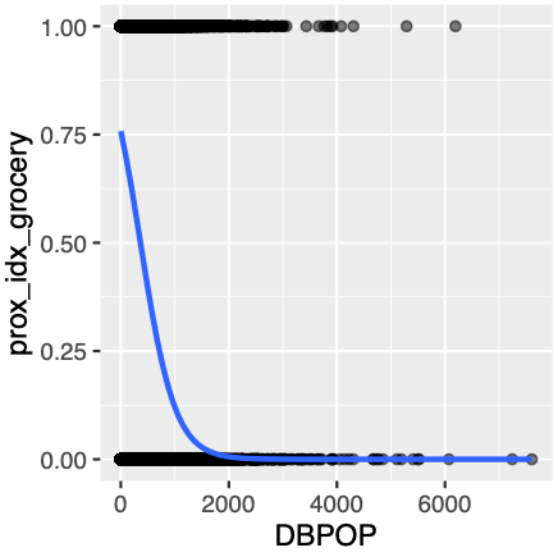


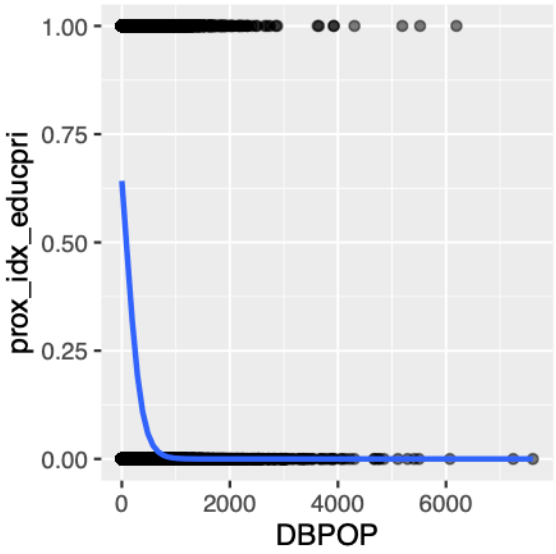
Logistics curves NAs

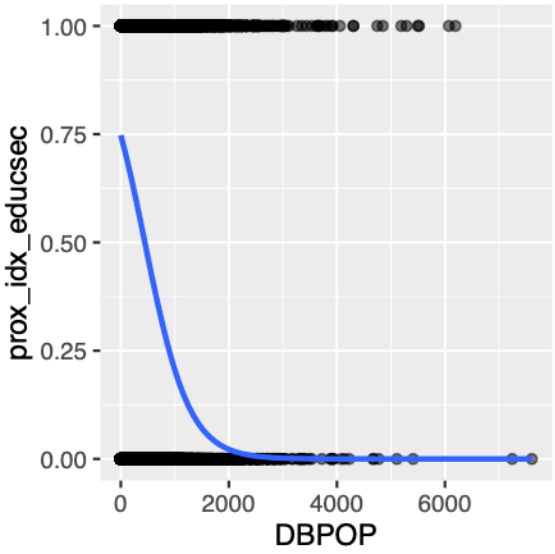


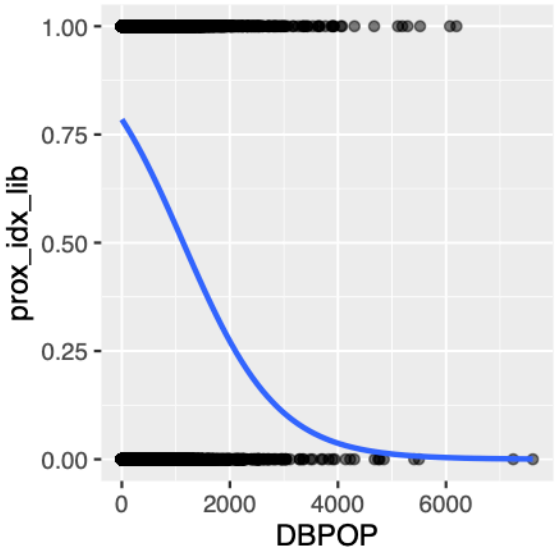


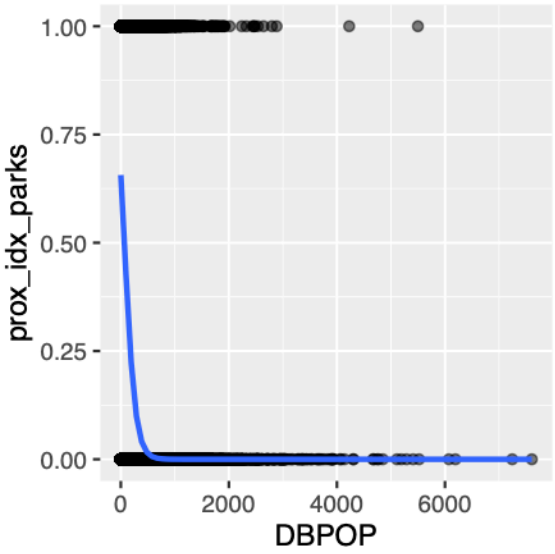

prox_idx_health
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

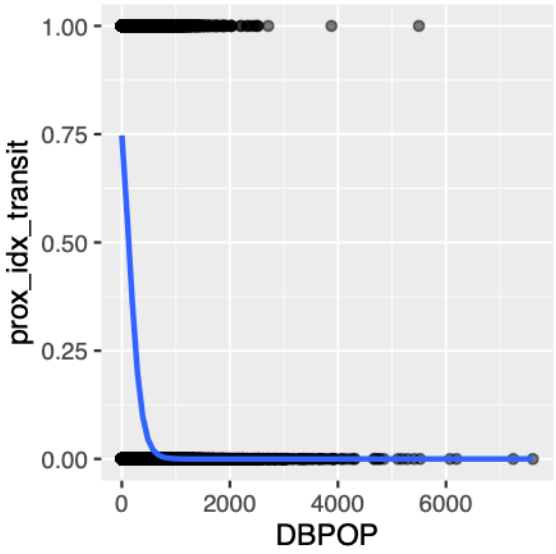

prox_idx_grocery
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP


prox_idx_educpri
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

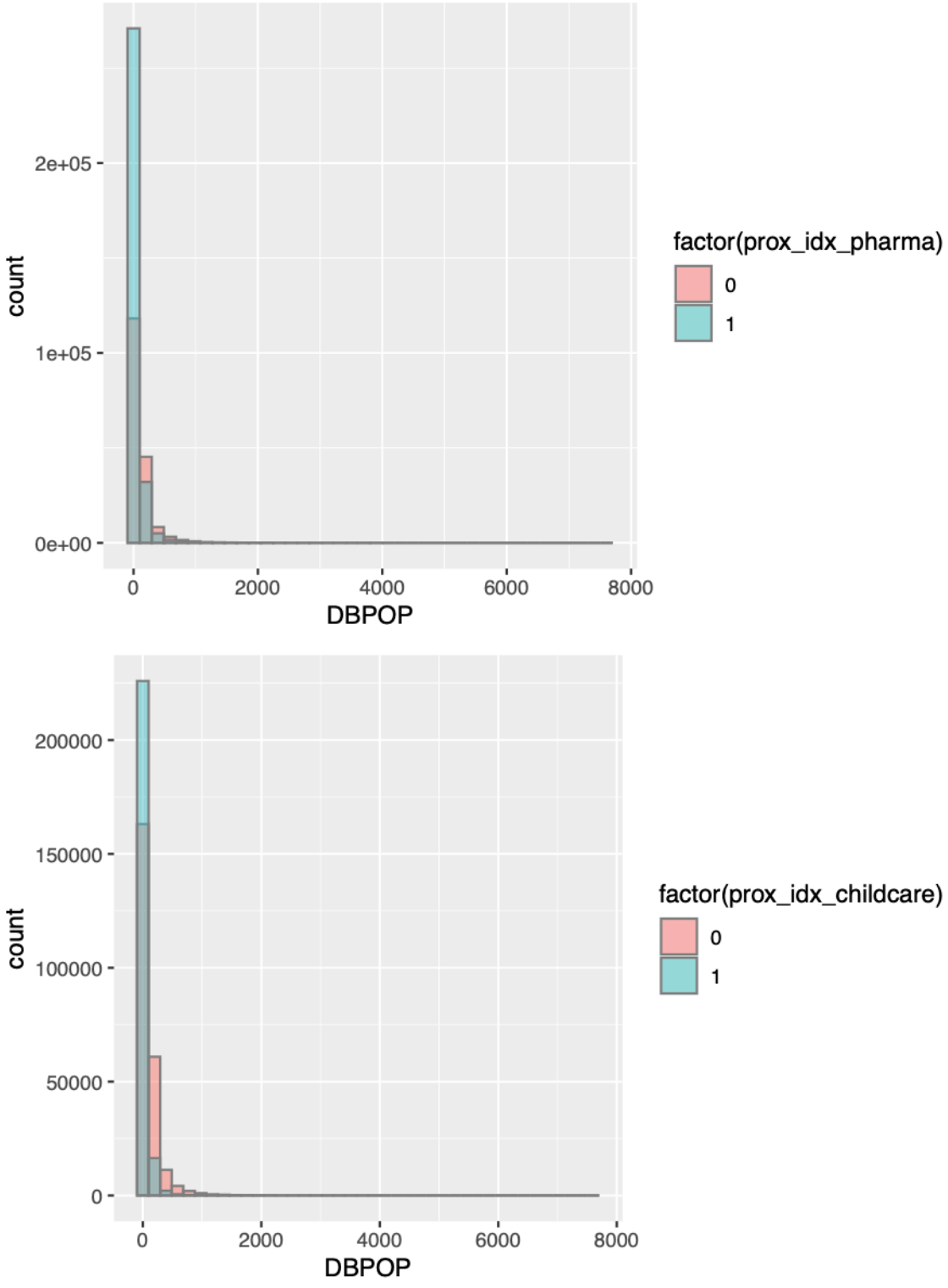

prox_idx_educsec
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

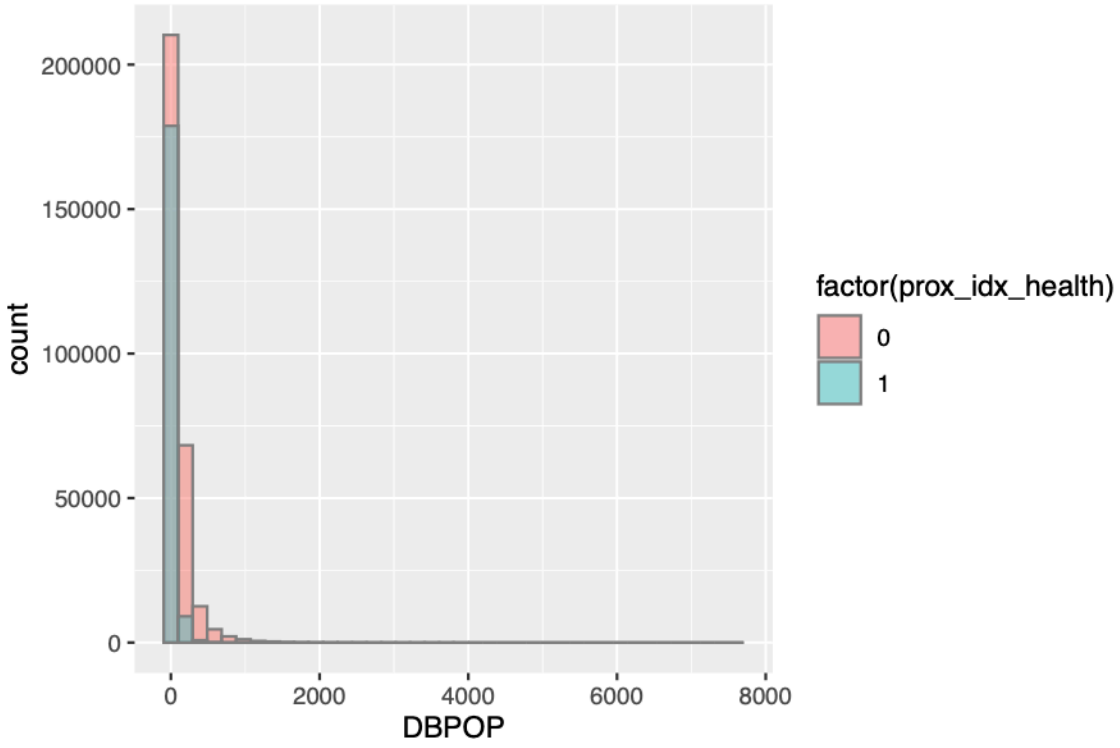

prox_idx_lib
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

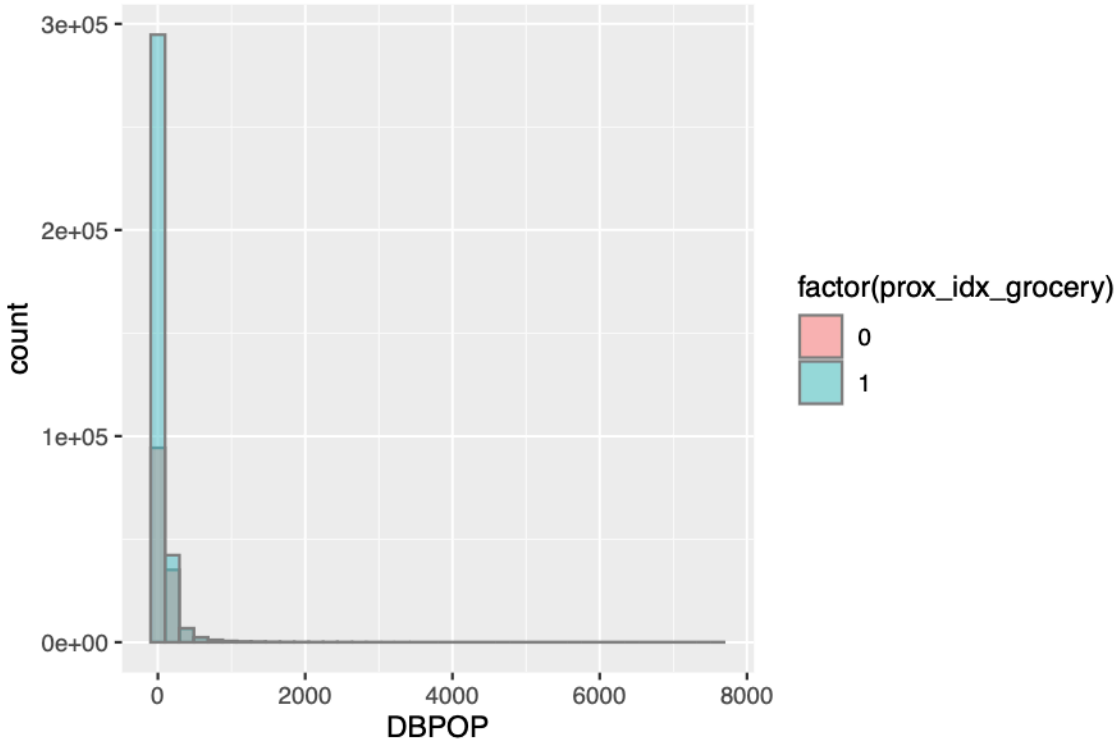

prox_idx_parks
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

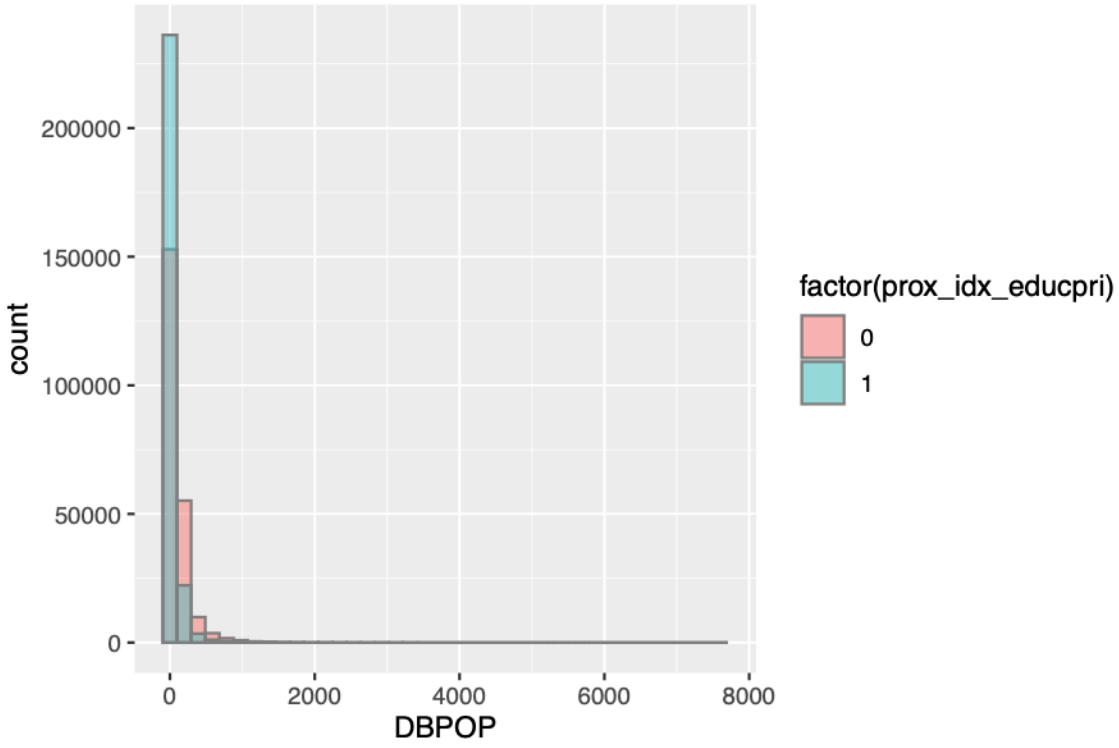

prox_idx_transit
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

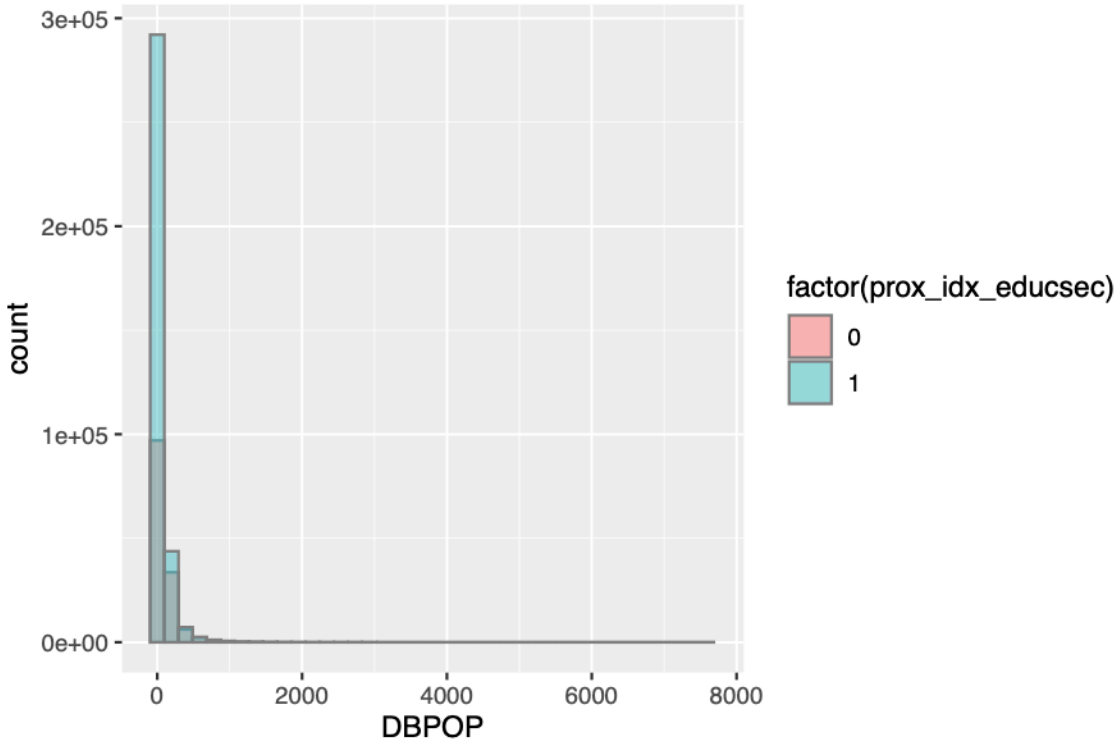
Histograms NAs

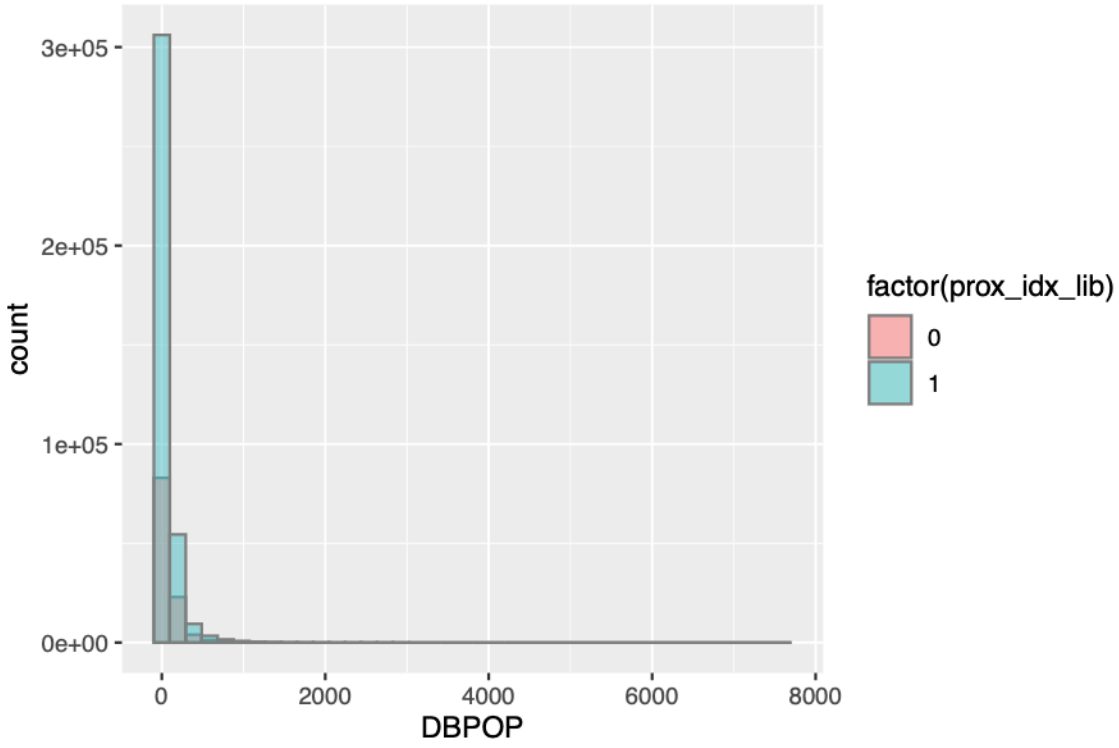


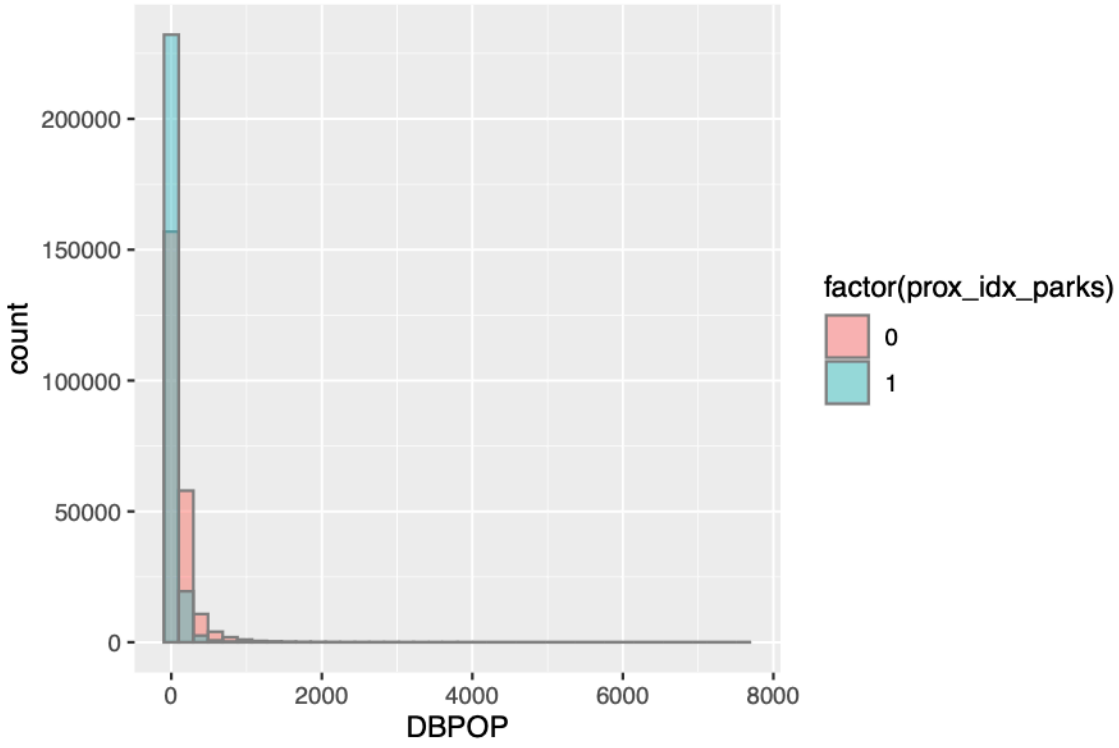

200000
150000
100000
50000
0
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_health)
0
1

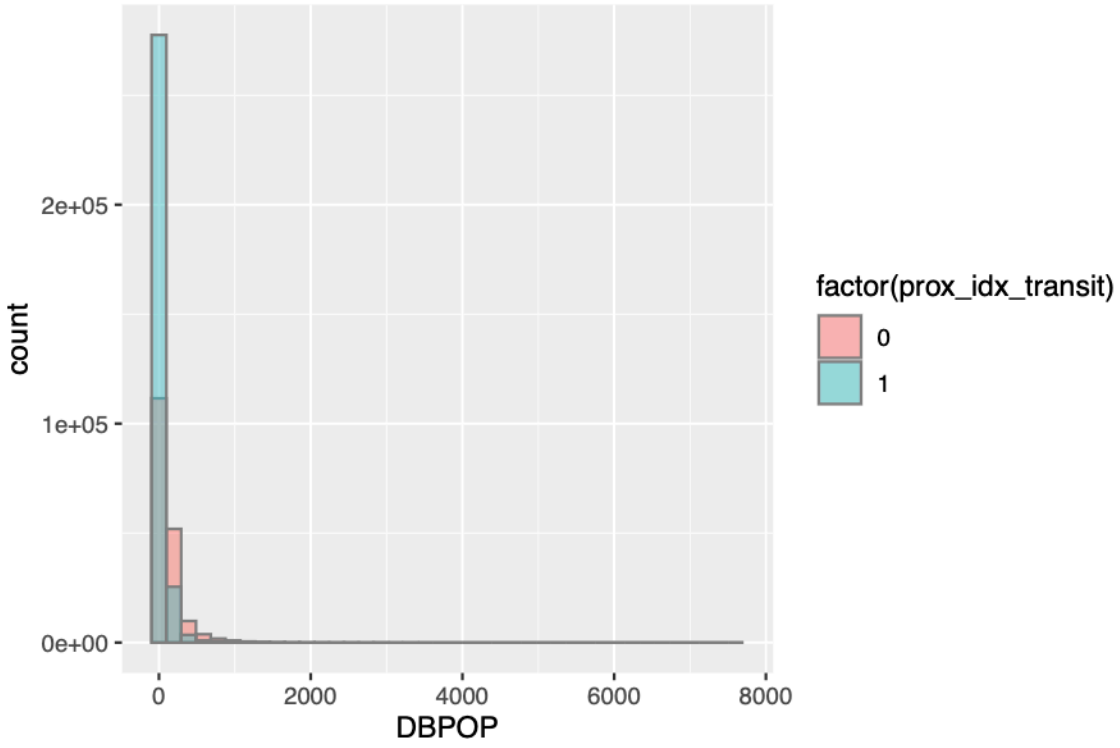

3e+05
2e+05
1e+05
0e+00
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_grocery)
0
1


200000
150000
100000
50000
0
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_educpri)
0
1

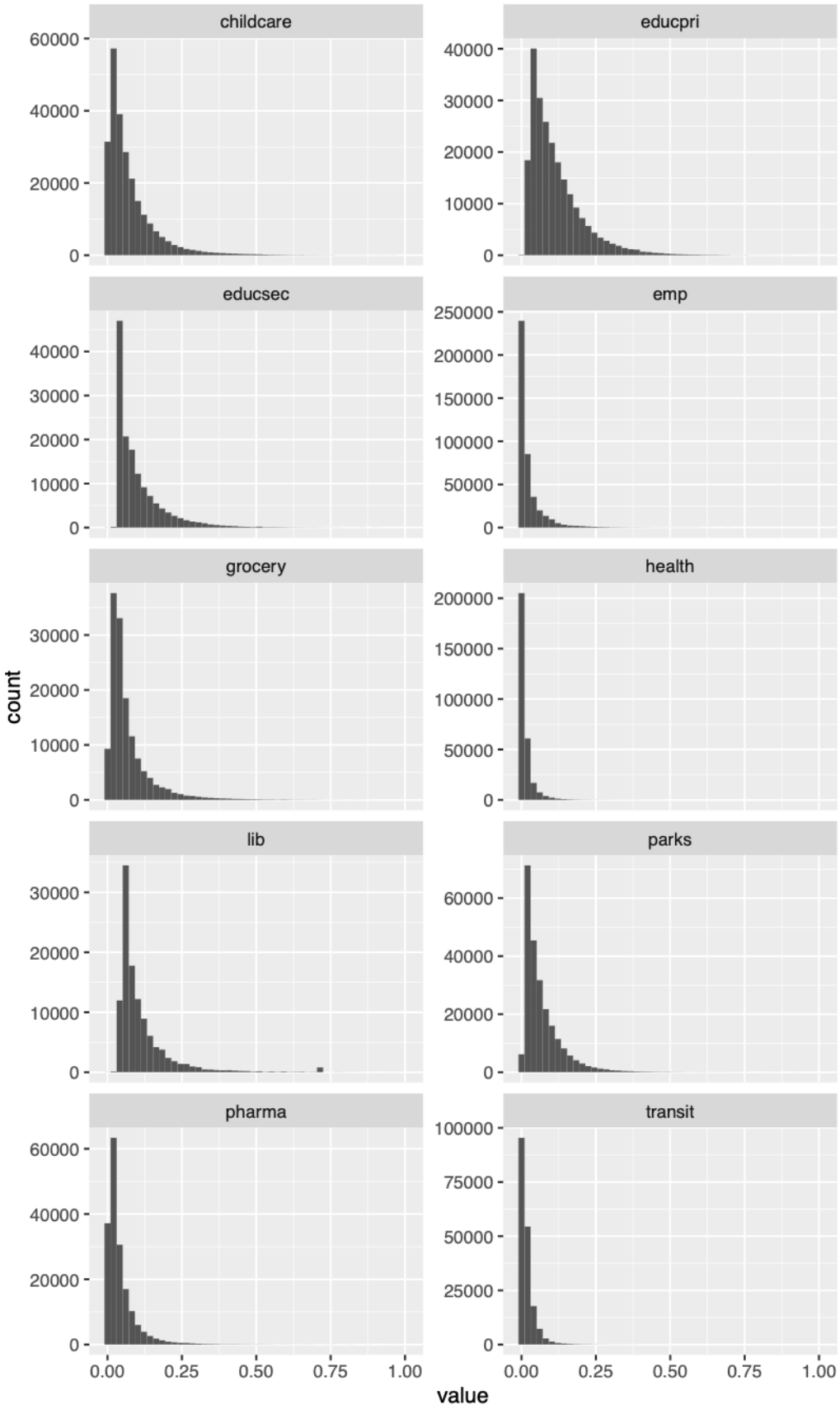

3e+05
2e+05
1e+05
0e+00
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_educsec)
0
1


3e+05
2e+05
1e+05
0e+00
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_lib)
0
1

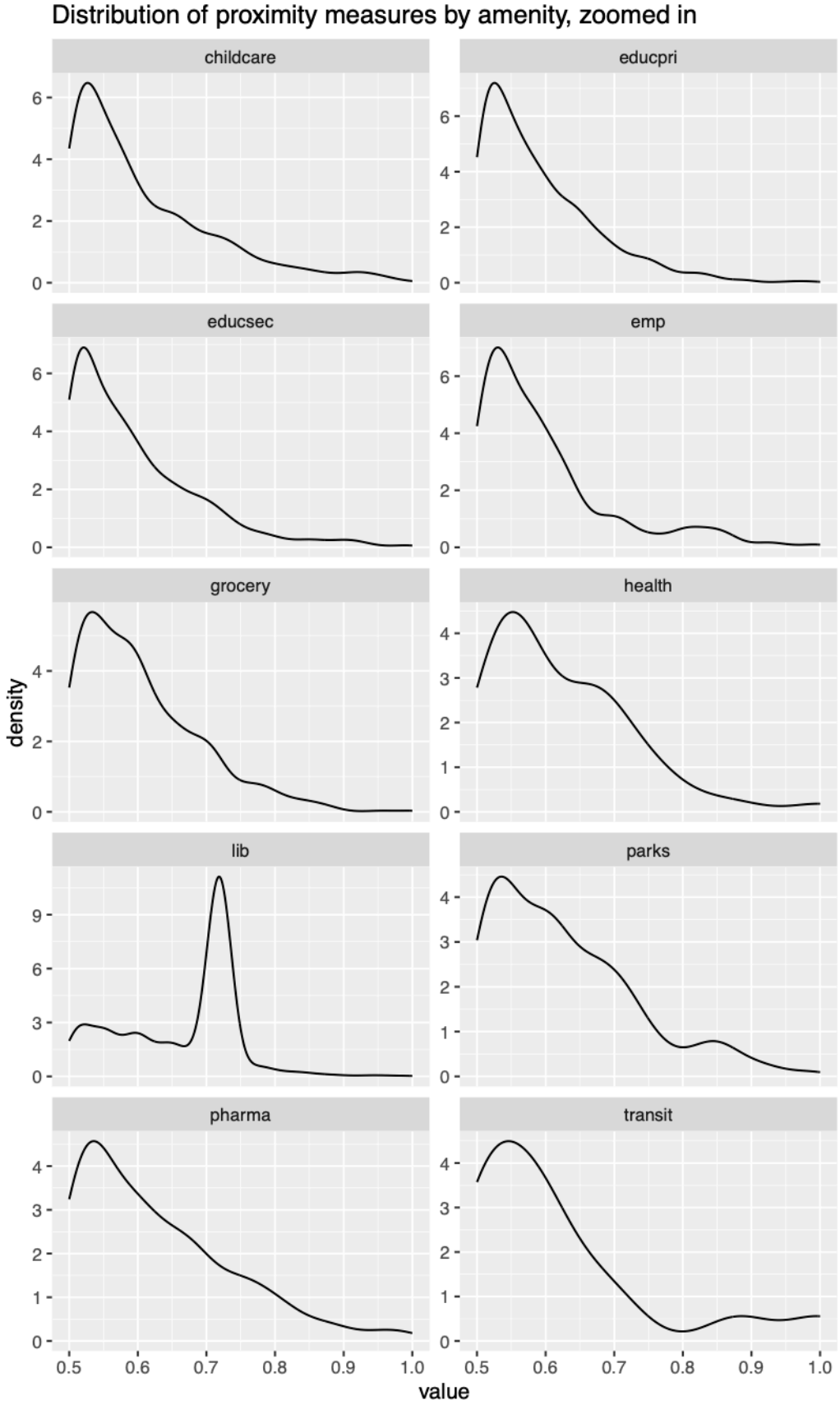

200000
150000
100000
50000
0
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_parks)
0
1


count
2e+05
1e+05
0e+00
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_transit)
0
1

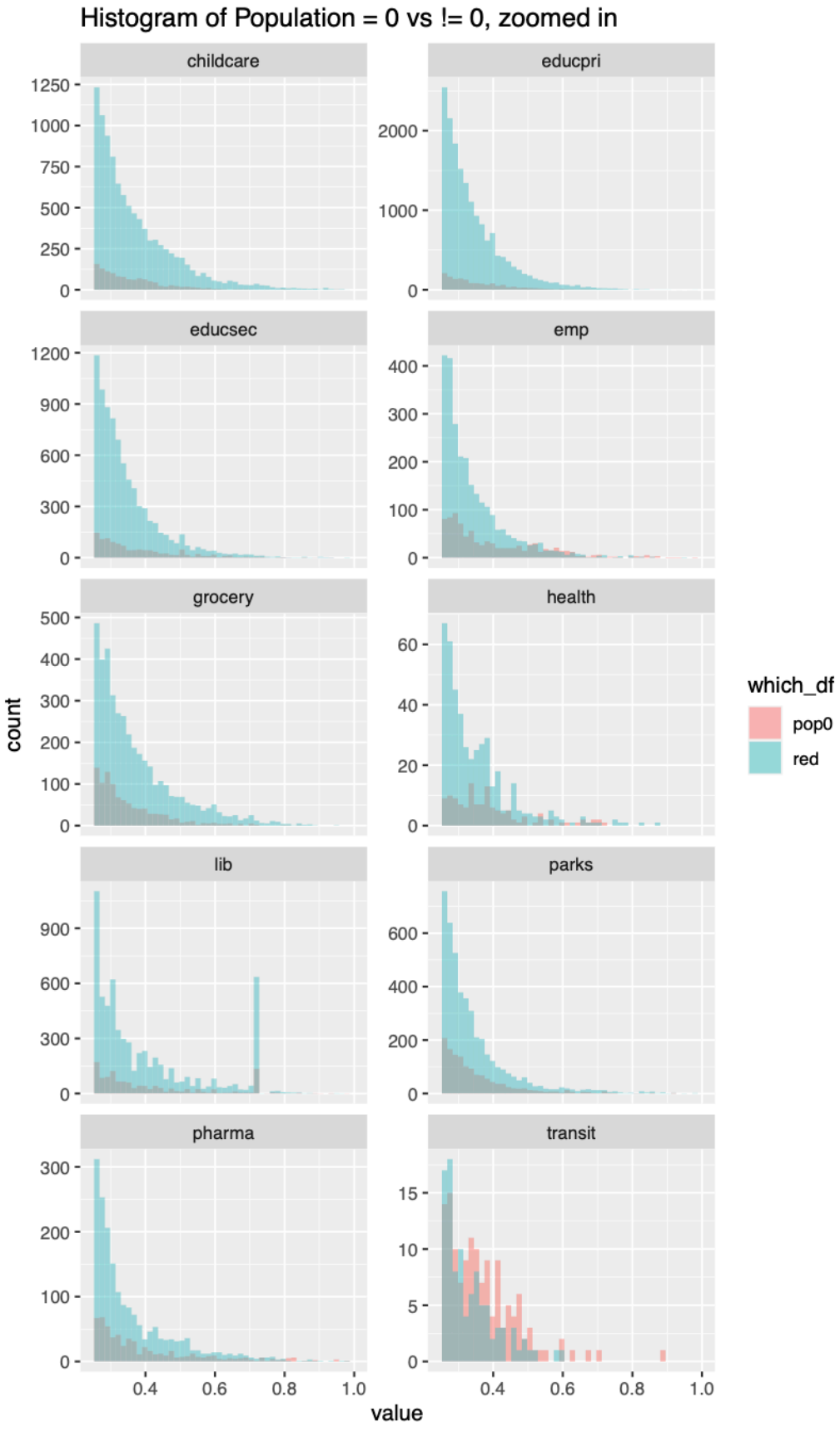
Proximity measures histograms



Proximity measures density zoomed in ? I think something is off



Population 0


Histogram of Population = 0 vs != 0, zoomed in
childcare
educpri
educsec
emp
grocery
health
lib
parks
pharma
transit
count
value
which_df
pop0
red

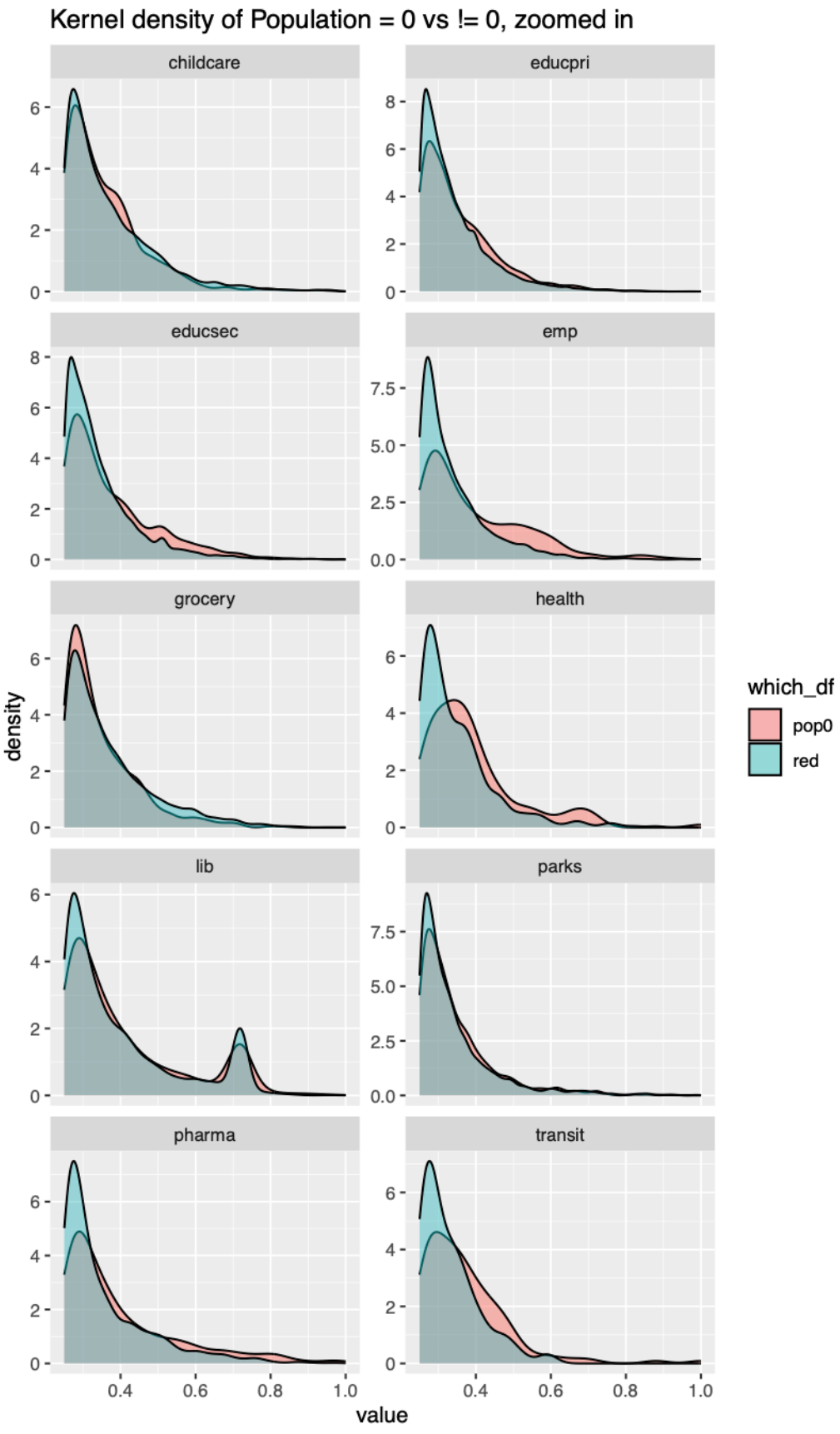

Kernel density of Population = 0 vs != 0, zoomed in
childcare
educpri
educsec
emp
grocery
health
lib
parks
pharma
transit
density
value
which_df
pop0
red

Table of counts for populations and NAs

There are still many more DBs that have populations > 0:

```
##
##          FALSE    TRUE
##   pop0  318194  834016
##   red  1863348 1878052
```