EDA draft

Ricky Heinrich & Team

2023-05-10

Outline

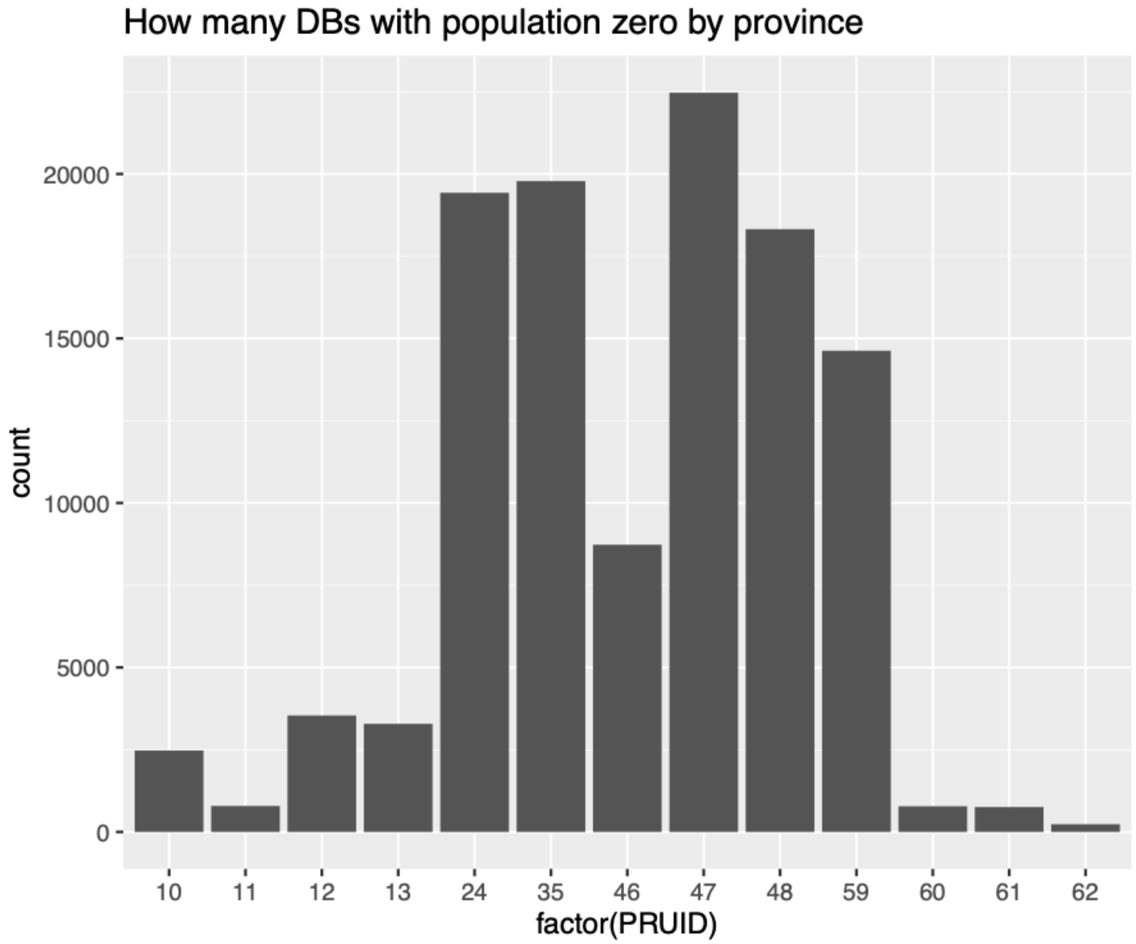
- Overall summary – how many variables, how many missing values
- Exploring missing values – logistic regression with added variables (master dataset)
- Exploring data distributions – histograms (actual numbers) – kernel densities (smooth)

Overall Summary

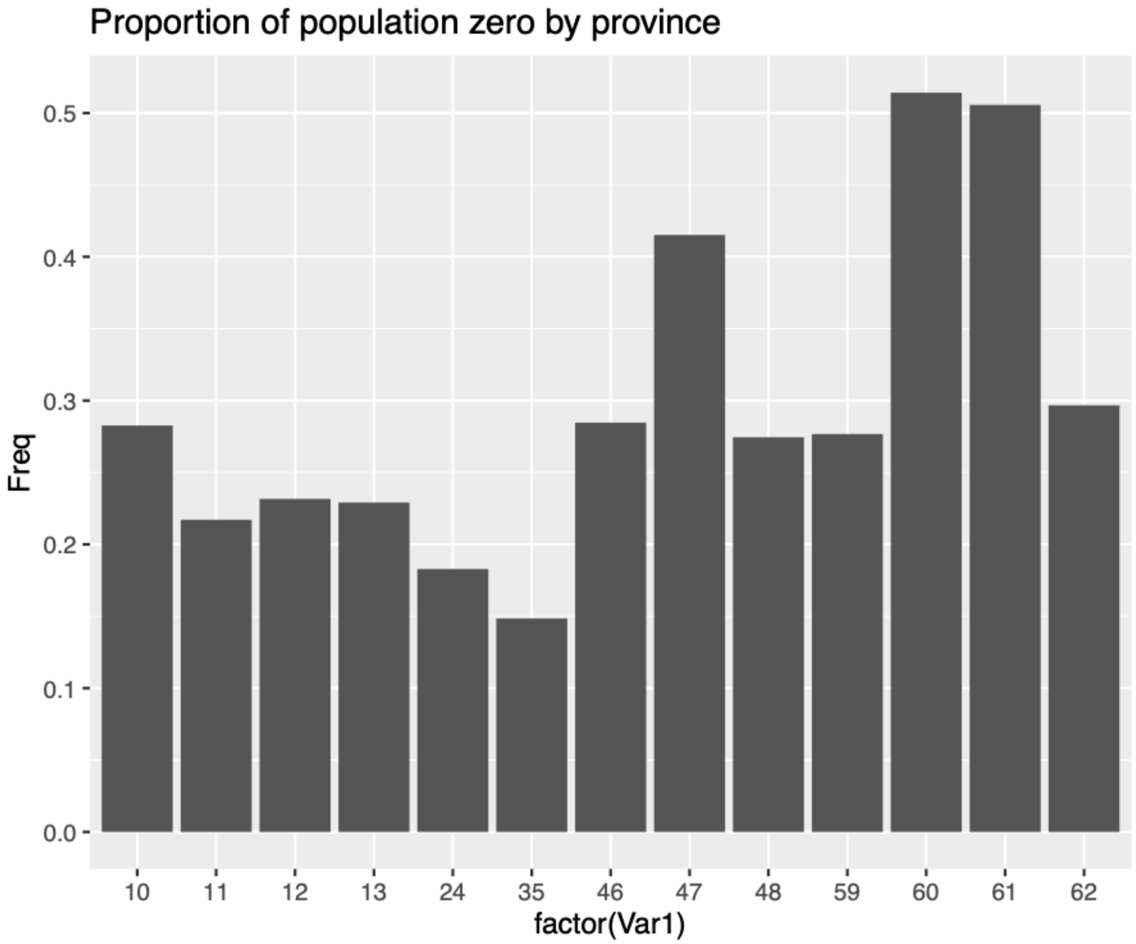
There are 489 676 rows in this data and 41 columns, meaning that 489 676 dissemination blocks are included. The 41 columns include information about the dissemination blocks themselves such as ID, population, and coordinates, as well as information about other census boundaries like dissemination areas, 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.

The DBs cover all of Canada: those included in our dataset are not exhaustive, but still, many have populations of zero. The province codes correspond to each province as follows: AB: 48 BC: 59 MN: 46 NB: 13 NL: 10 NWT: 61 NC: 12 NV: 62 ON: 35 PEI: 11 QB: 24 SK: 47 YK: 60

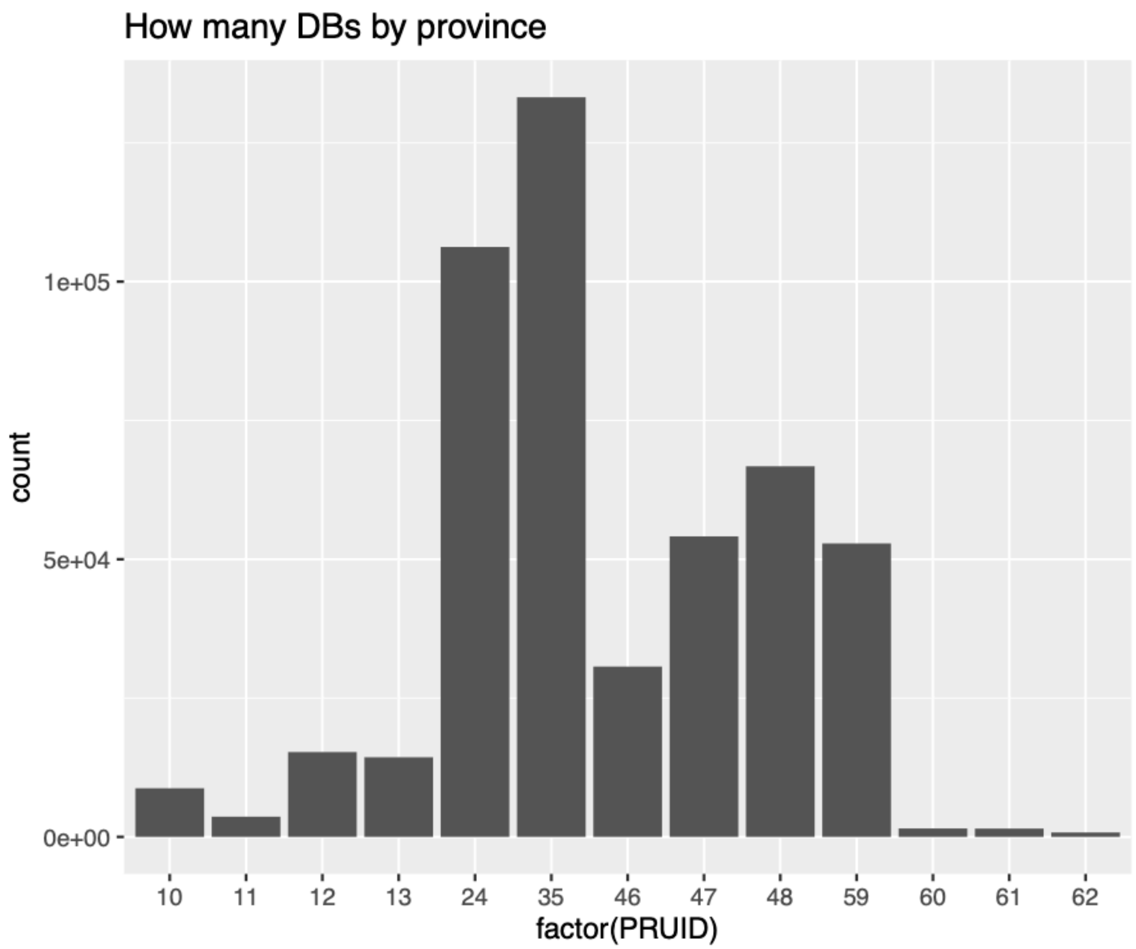
It could be reasonable to expect that if the population of a DB is 0, then the proximity measure are also near 0: it is intuitive that for the most part, amenities are further away from areas with no populations. It is thus reasonable to explore the cases where the population is zero, to see its prevalence, and deduce how it may affect the values of proximity measures. We see that Saskatchewan has the most DBs included with a population of zero, followed by Ontario, Quebec, and Alberta.



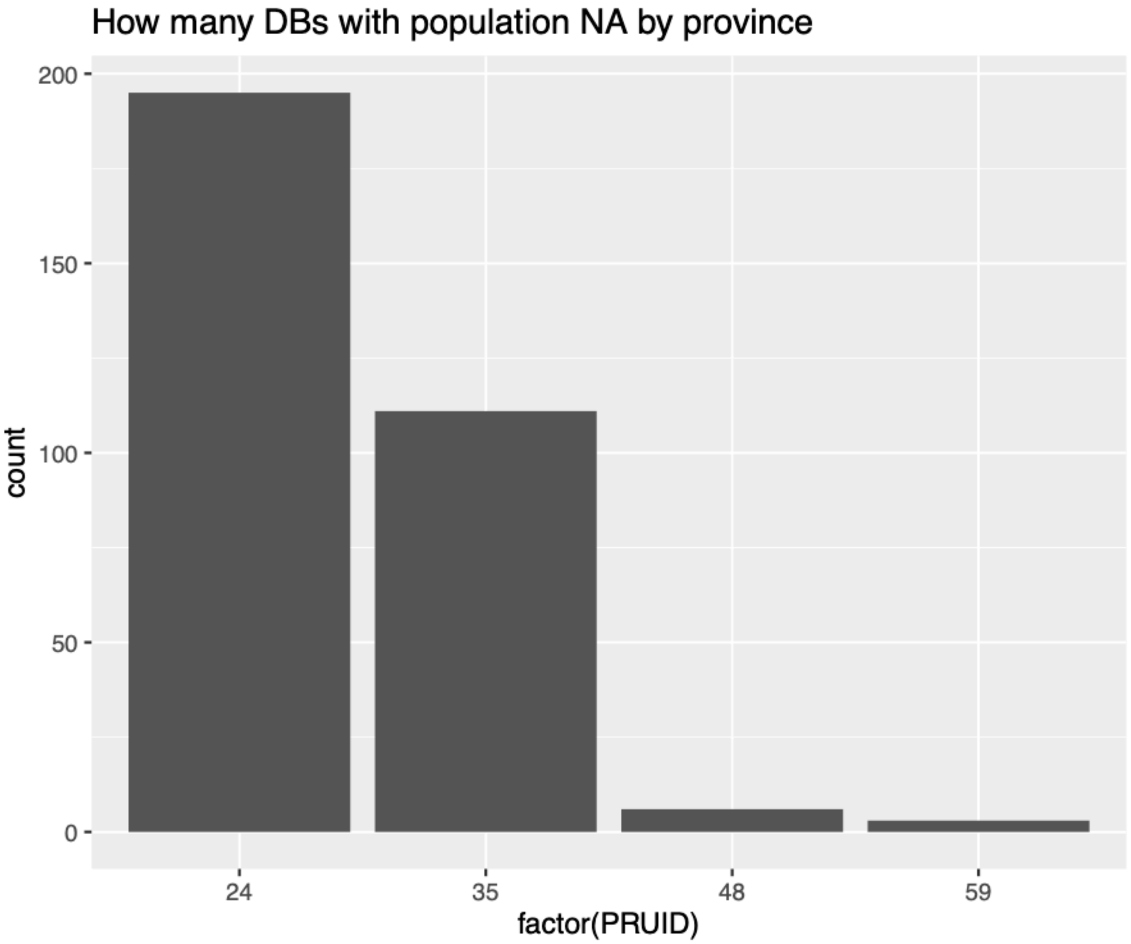
Taking the proportions however, we see that over 50% of Yukon and NWT's DBs have a population of 0, and Saskatchewan has over 40%. Ontario has the lowest at around 15%, followed by Quebec at around 18%.



We see that Ontario and Quebec have the most DBs, and the territories have the least.



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.

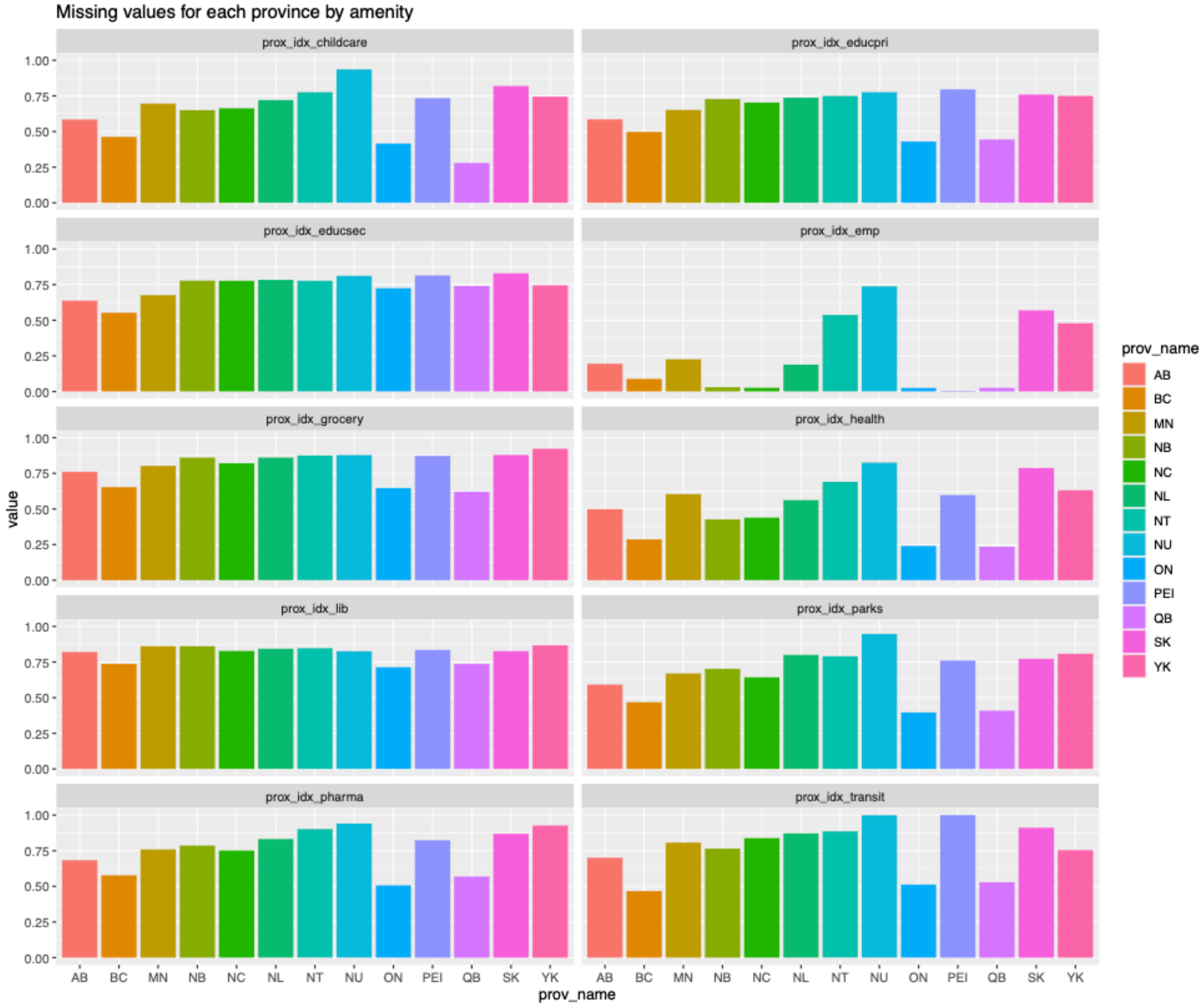


```
##
## IRI S-É
## 269  46
```

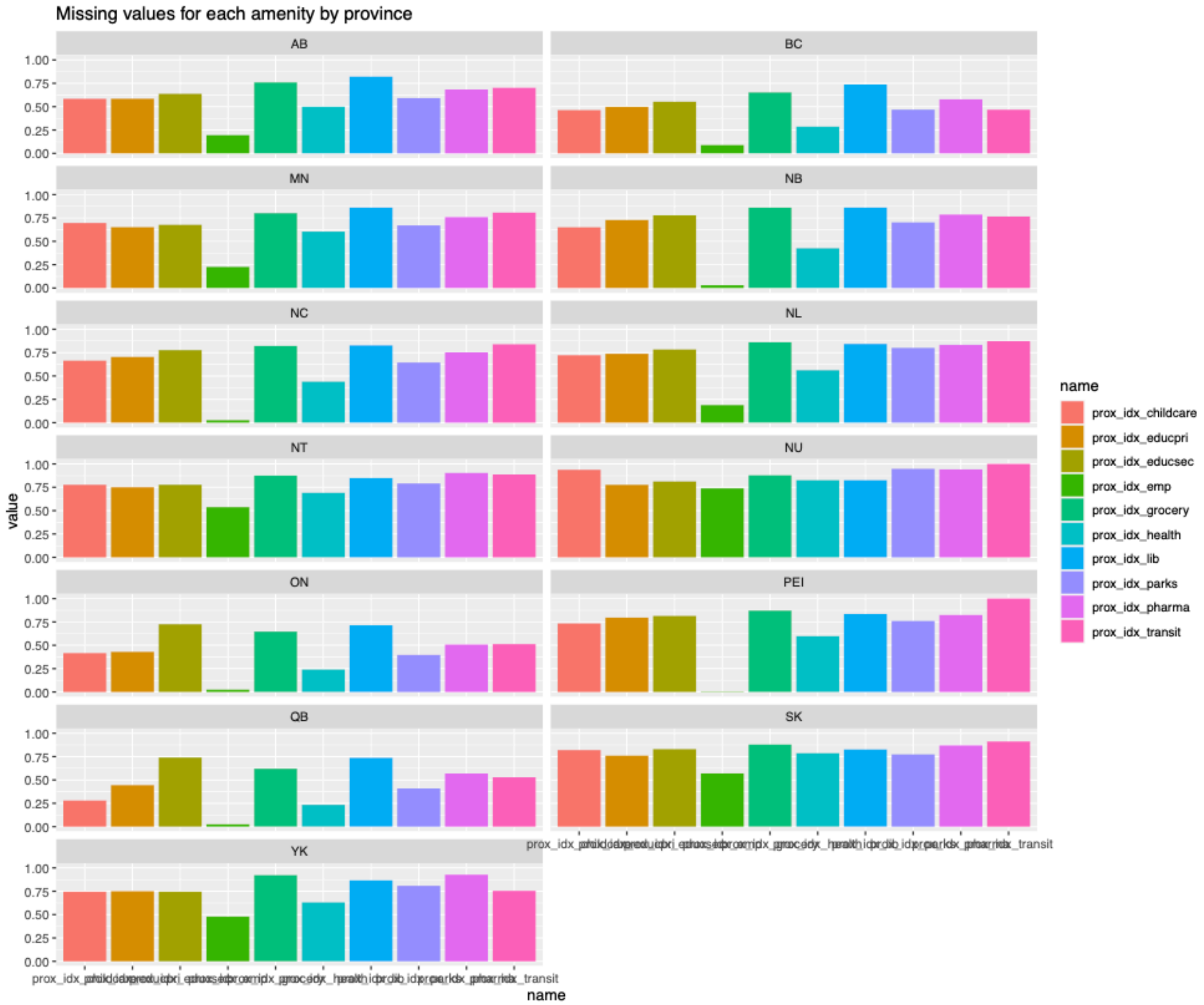
In the summary of the dataset, we see that there are many missing values. We see that the library proximity indicator contains the most missing values, at around 77%, followed by the proximity measures for grocery and secondary education. Only two out of the ten amenities have proximity measures missing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

We can see the proportion of missing values for each amenity by province:

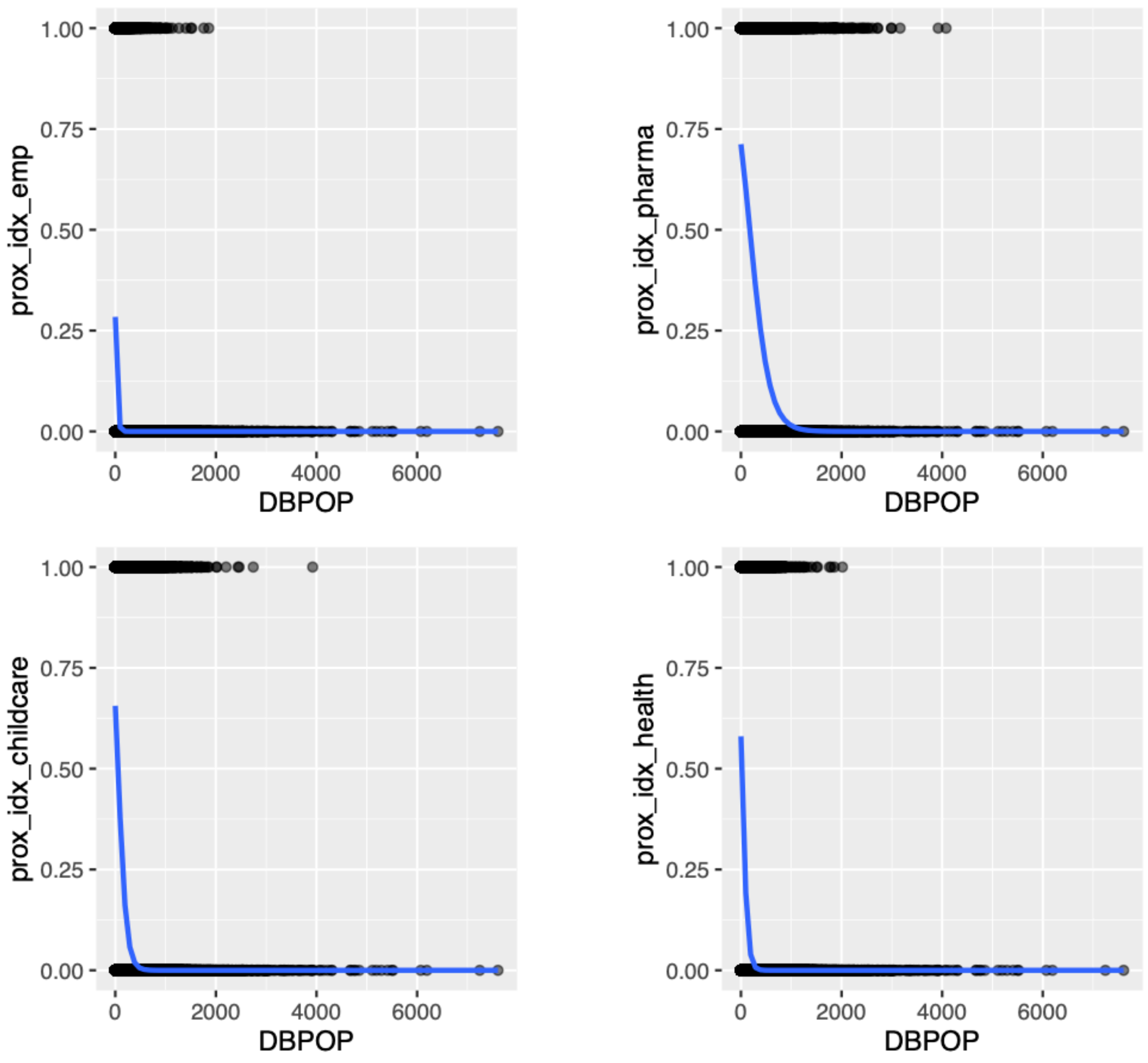


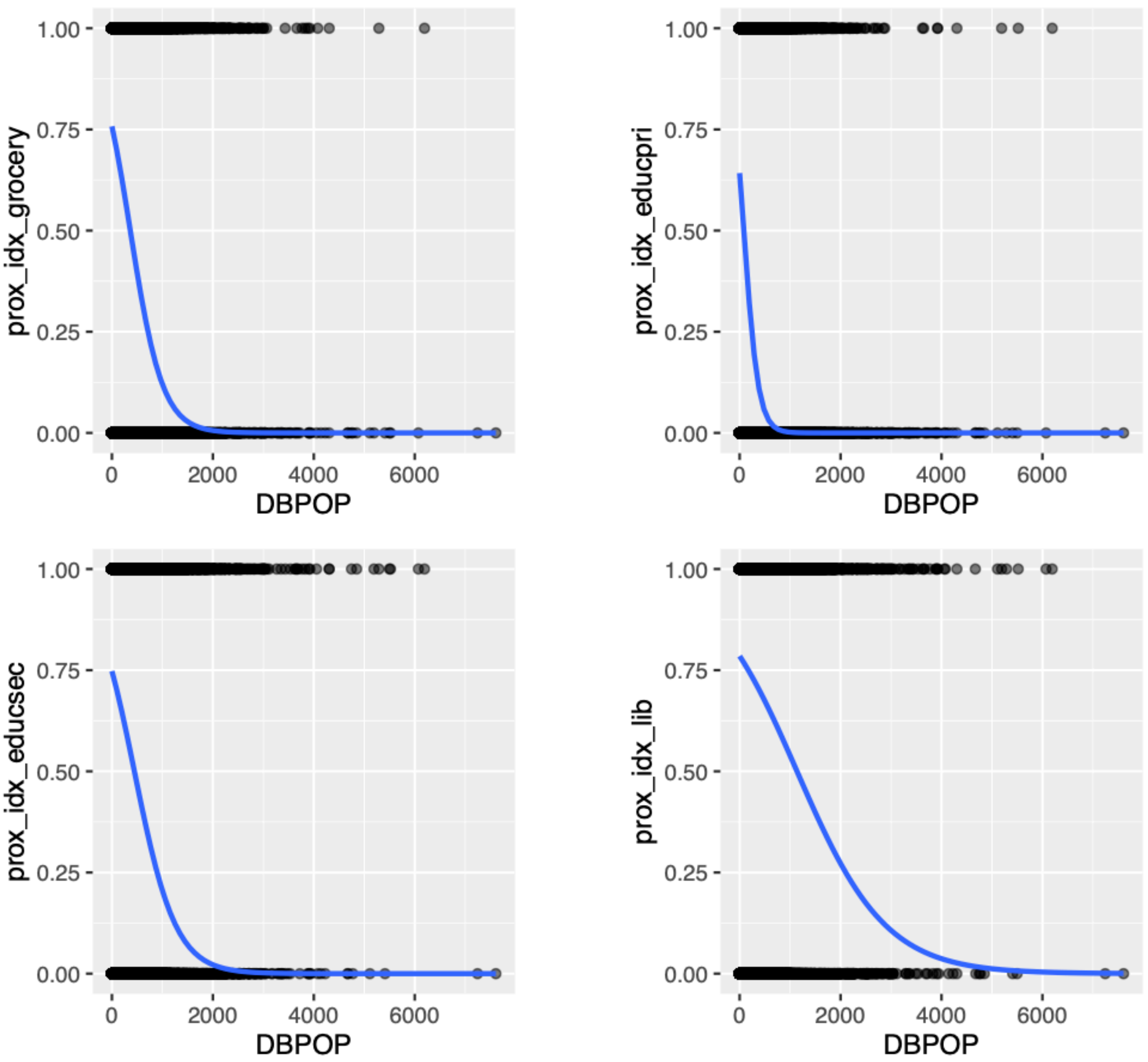
We see that overall, employment has the lowest rates of missing values, but has also more range. Ontario and Quebec seems to have the least missing values for most amenities relative to the other regions, whereas Nunavut usually has the most. It seems like the amount of proximity measure missing for libraries have the most consistency across regions.

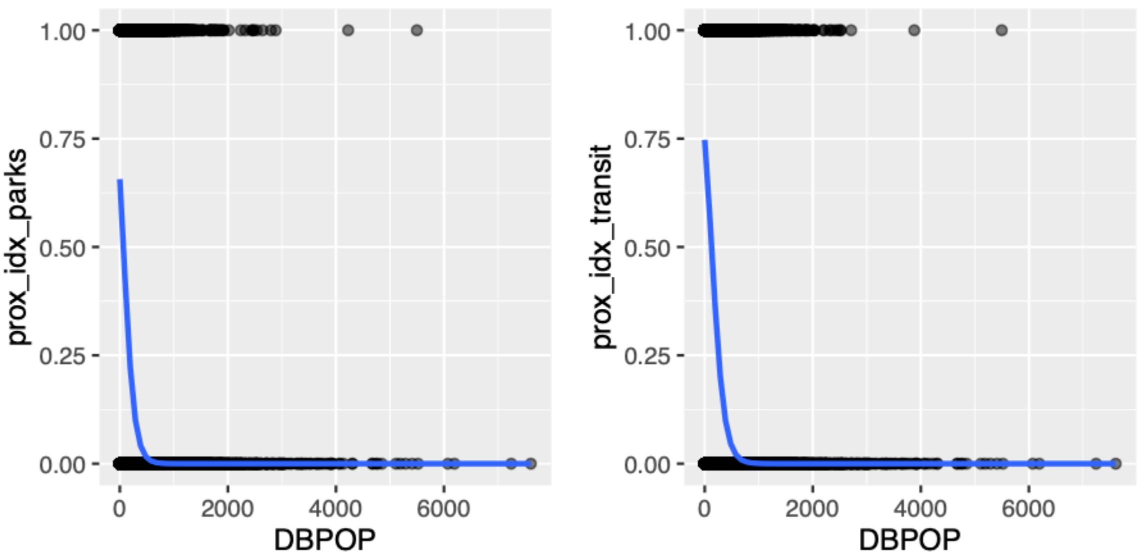


Here we see the same information, but flipped so we can compare each amenity for each province.

For each amenity, we can plot the occurrence of missing values in a DB vs its population, and plot a basic logistics curve.


prox_idx_emp
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP
prox_idx_pharma
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP
prox_idx_childcare
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP
prox_idx_health
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

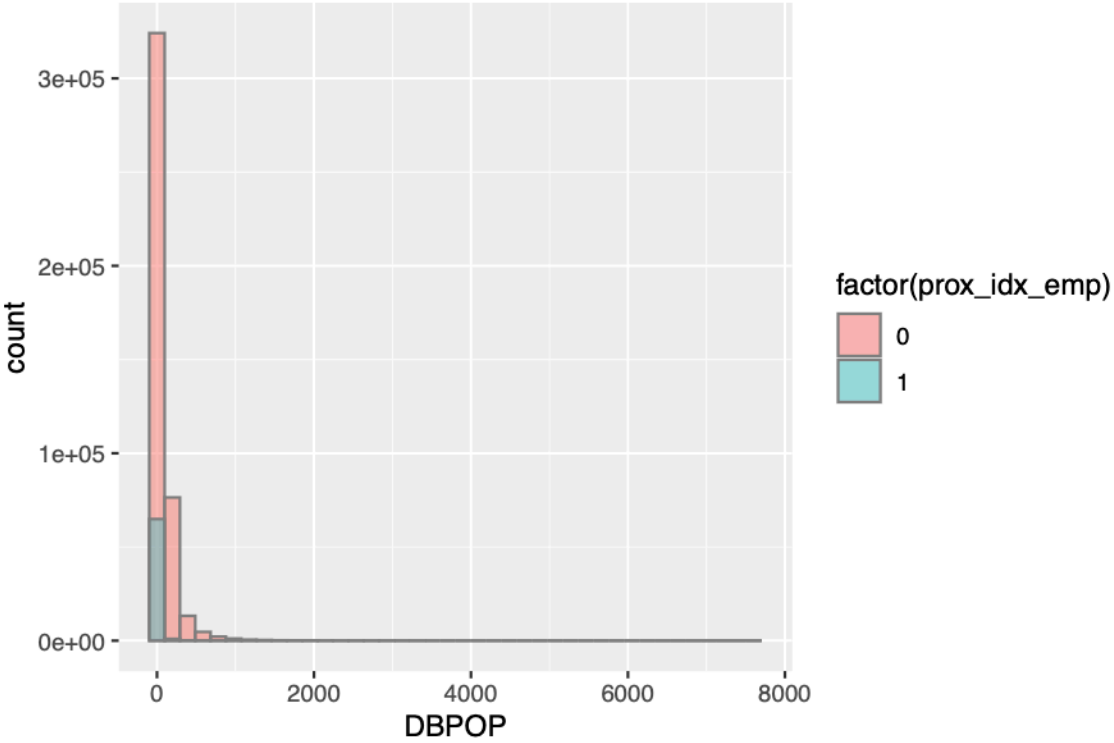

prox_idx_grocery
prox_idx_educpri
prox_idx_educsec
prox_idx_lib
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
DBPOP

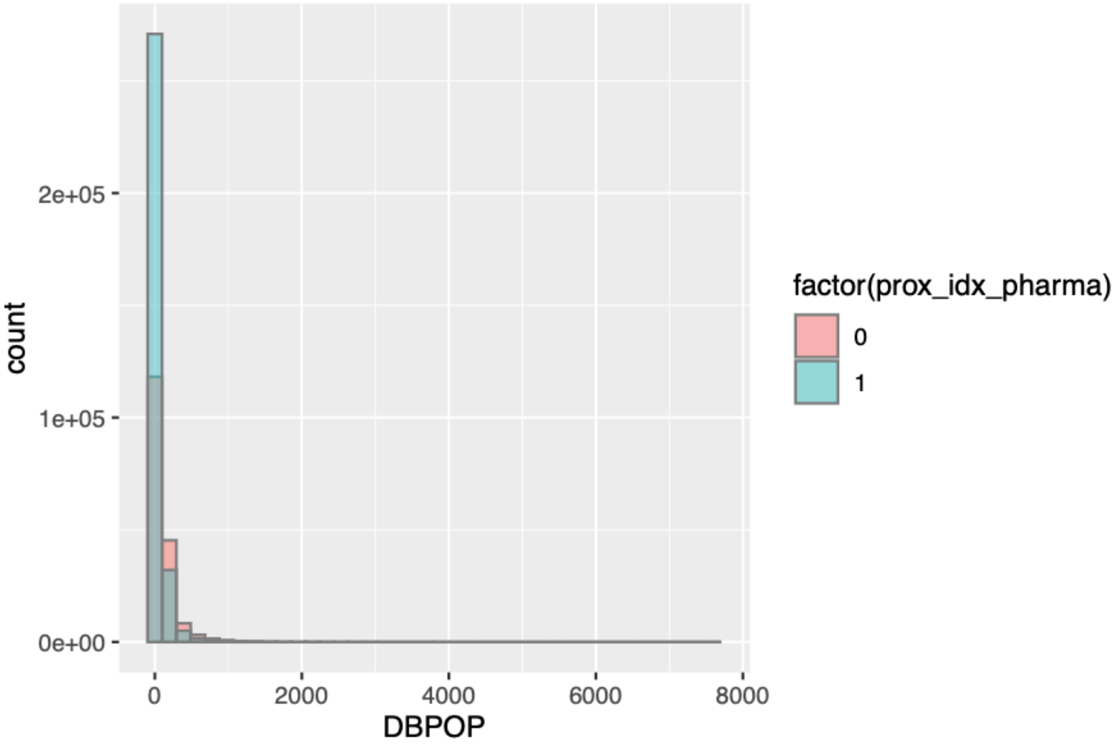


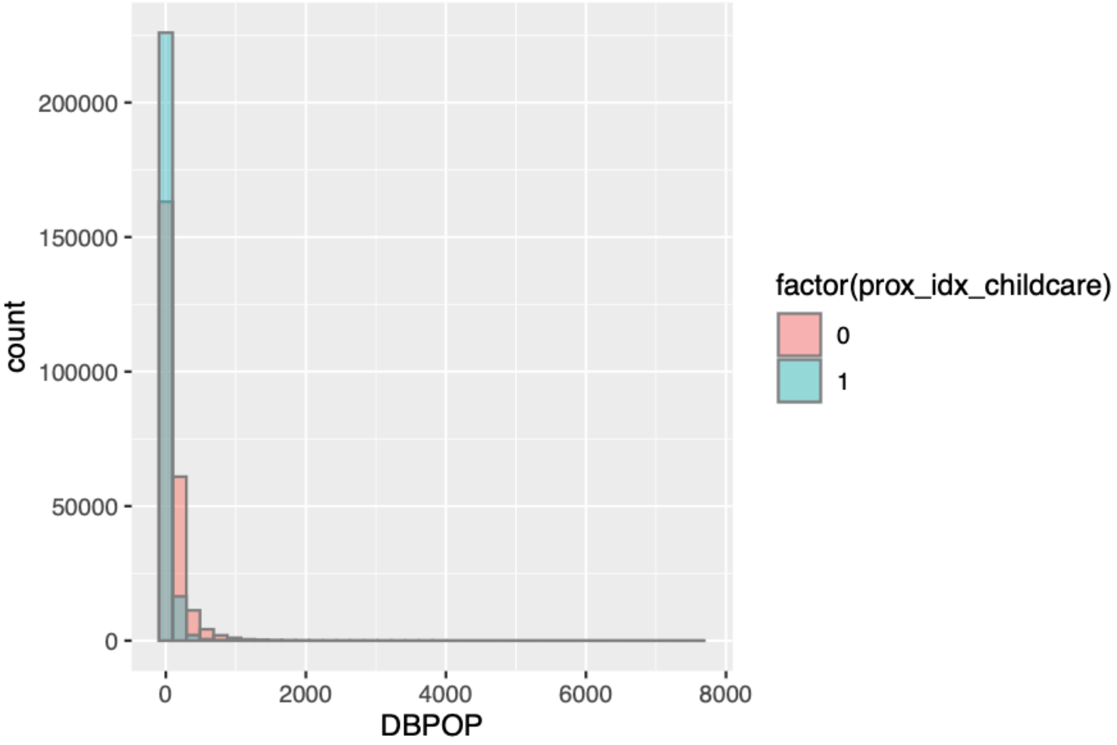
We see that for some amenities, like employment and health, the missing values are concentrated among DBs with small populations. These are the same amenities with less than 50% of values missing.

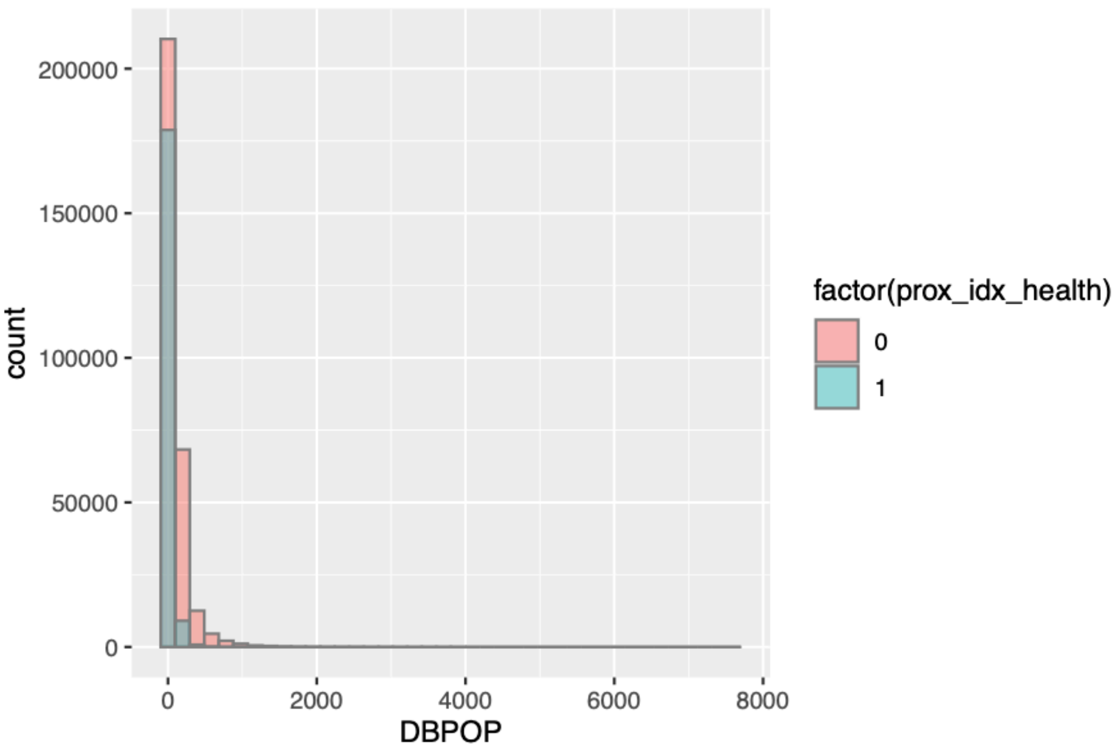
Overall it seems like the population of the DB is not the only factor, if at all, affecting whether a proximity measure is missing for that DB.

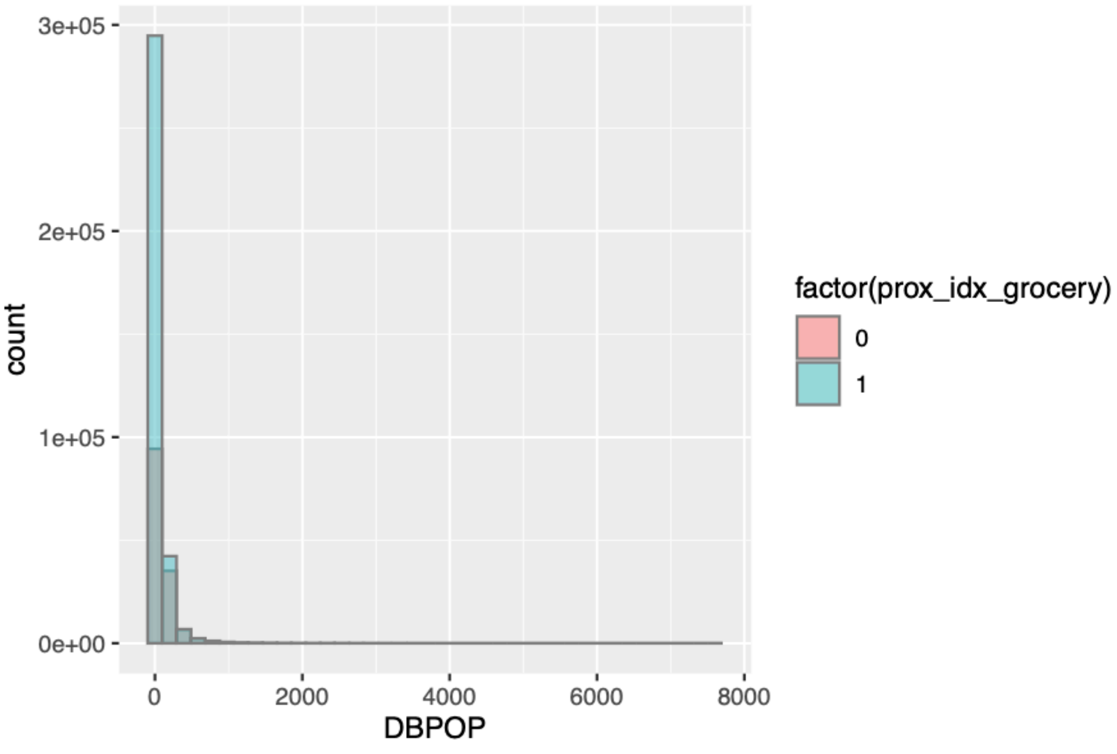
We can also plot the histograms of missing values vs populations for each amenity, where '1' (blue) is a missing value and '0' (pink) is a value not missing:

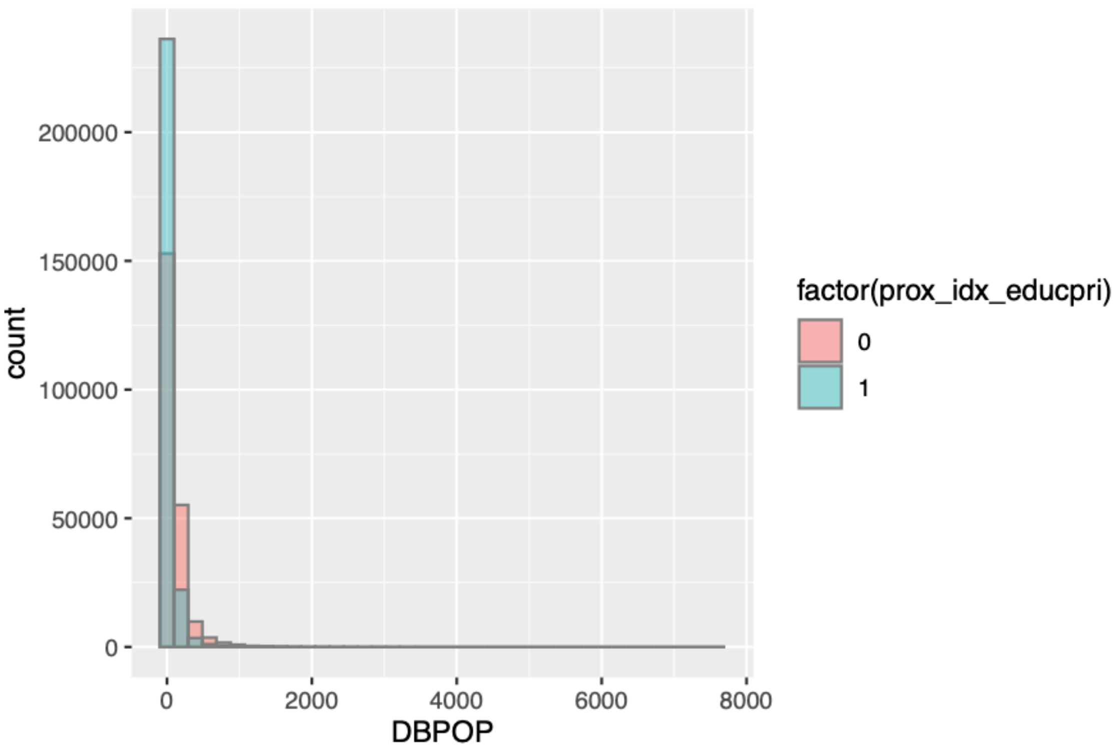


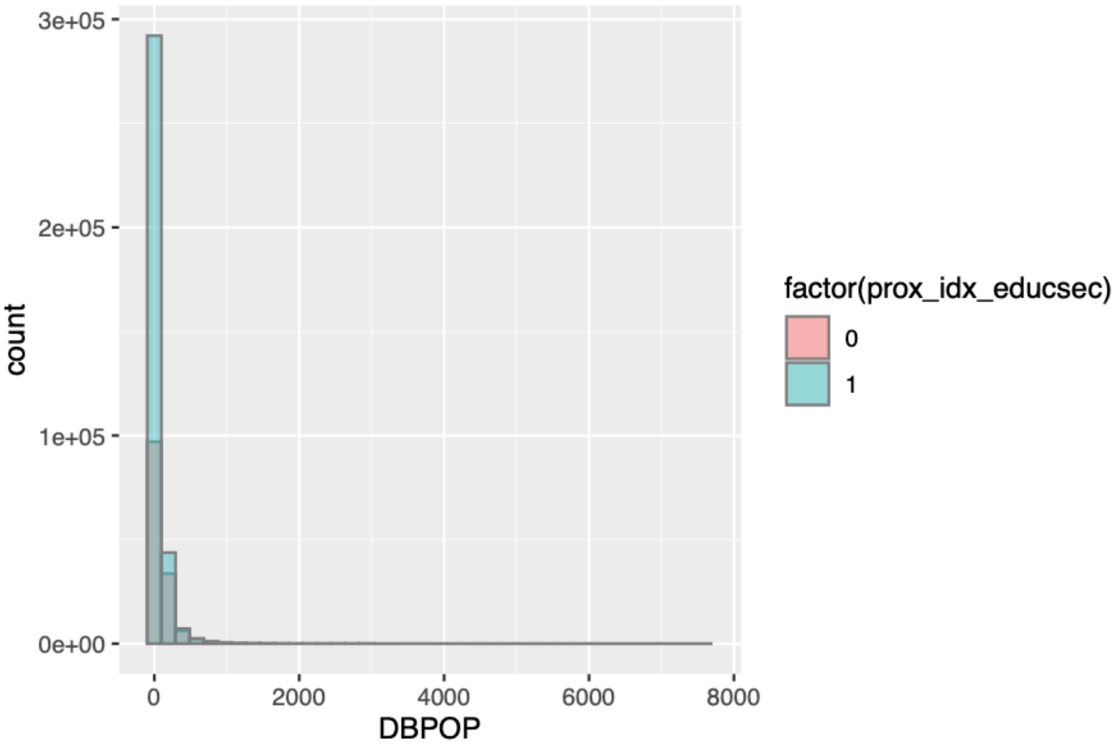

count
2e+05
1e+05
0e+00
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_pharma)
0
1

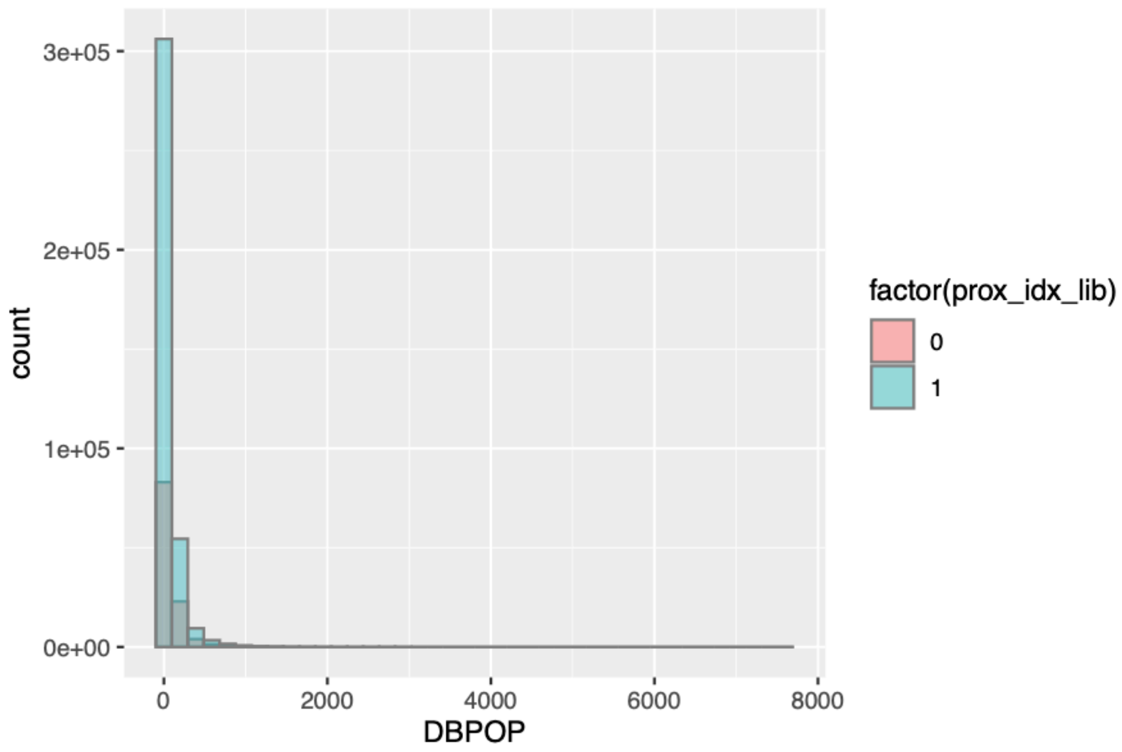

count
200000
150000
100000
50000
0
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_childcare)
0
1

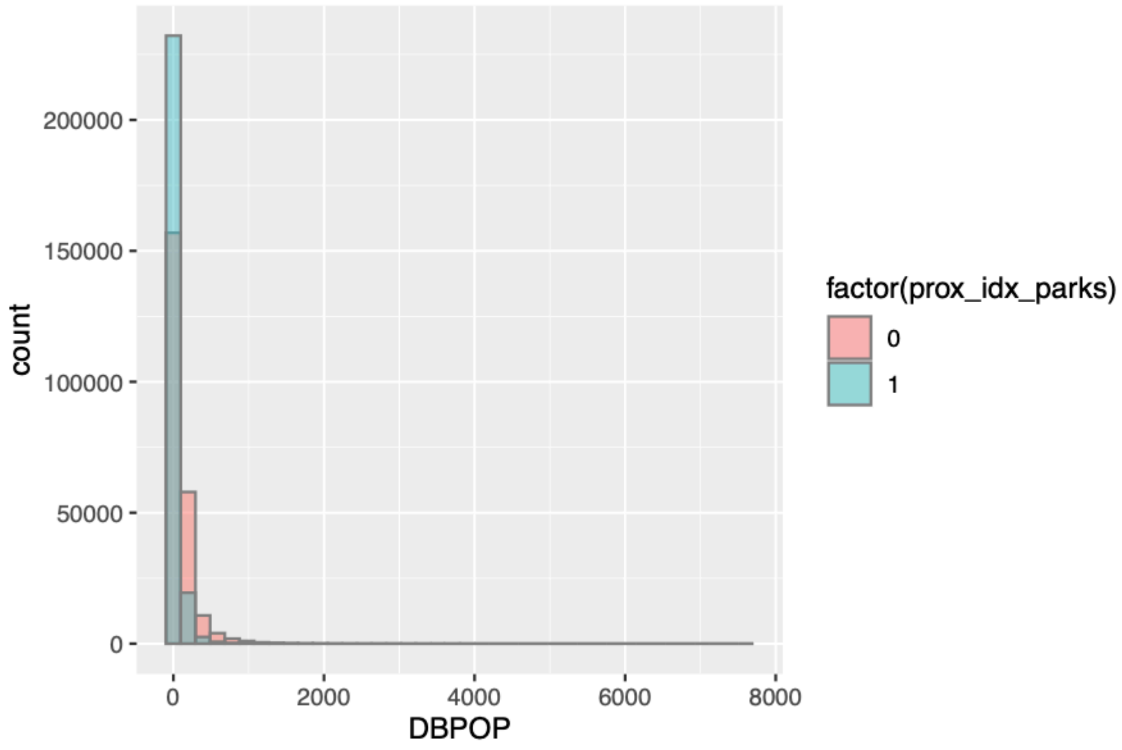

count
200000
150000
100000
50000
0
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_health)
0
1

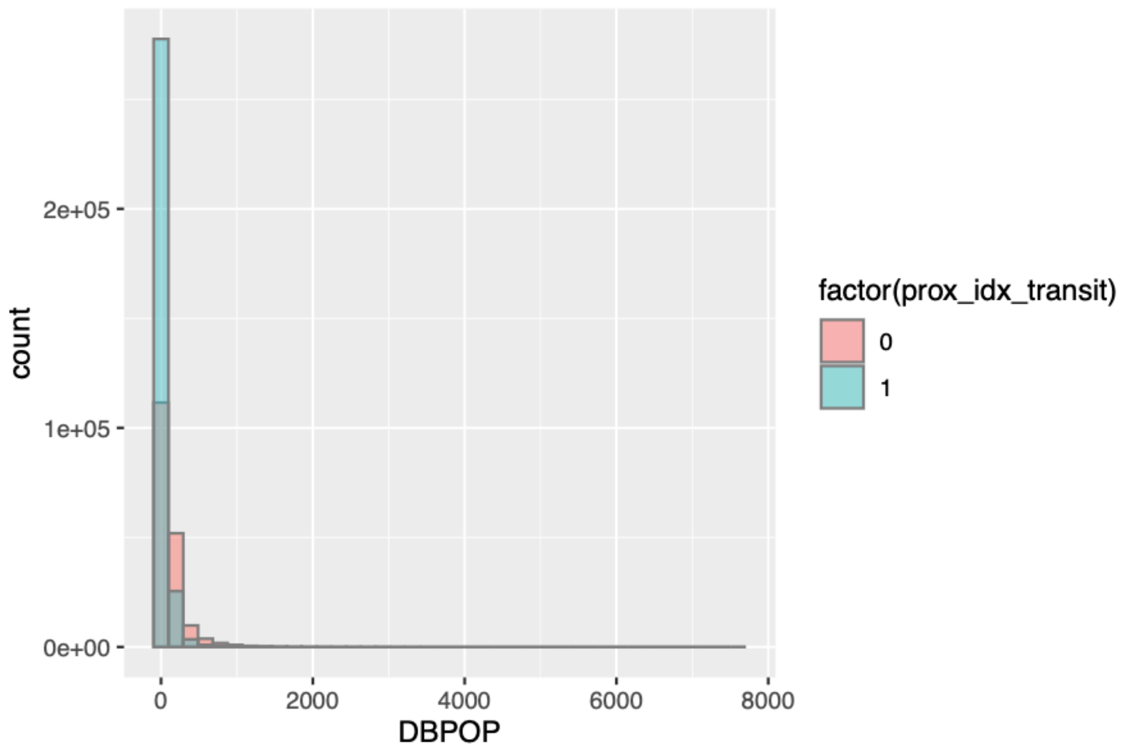

count
3e+05
2e+05
1e+05
0e+00
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_grocery)
0
1


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


count
3e+05
2e+05
1e+05
0e+00
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



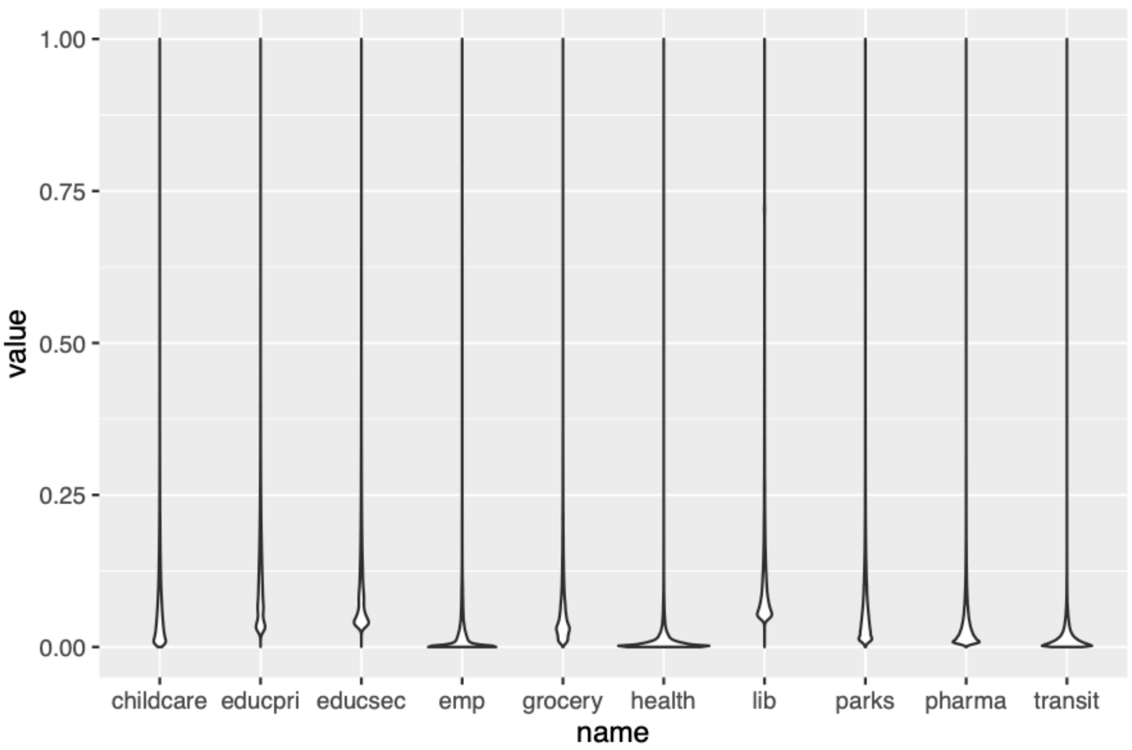
Overall, there are more DBs with lower populations than larger populations. We see that for some amenities, at smaller populations, there are a lot more missing values. Again, employment and health are the only two where there are always more actual values than missing values at every population bin.

Model with other variables to see relationship: need add 'master' df

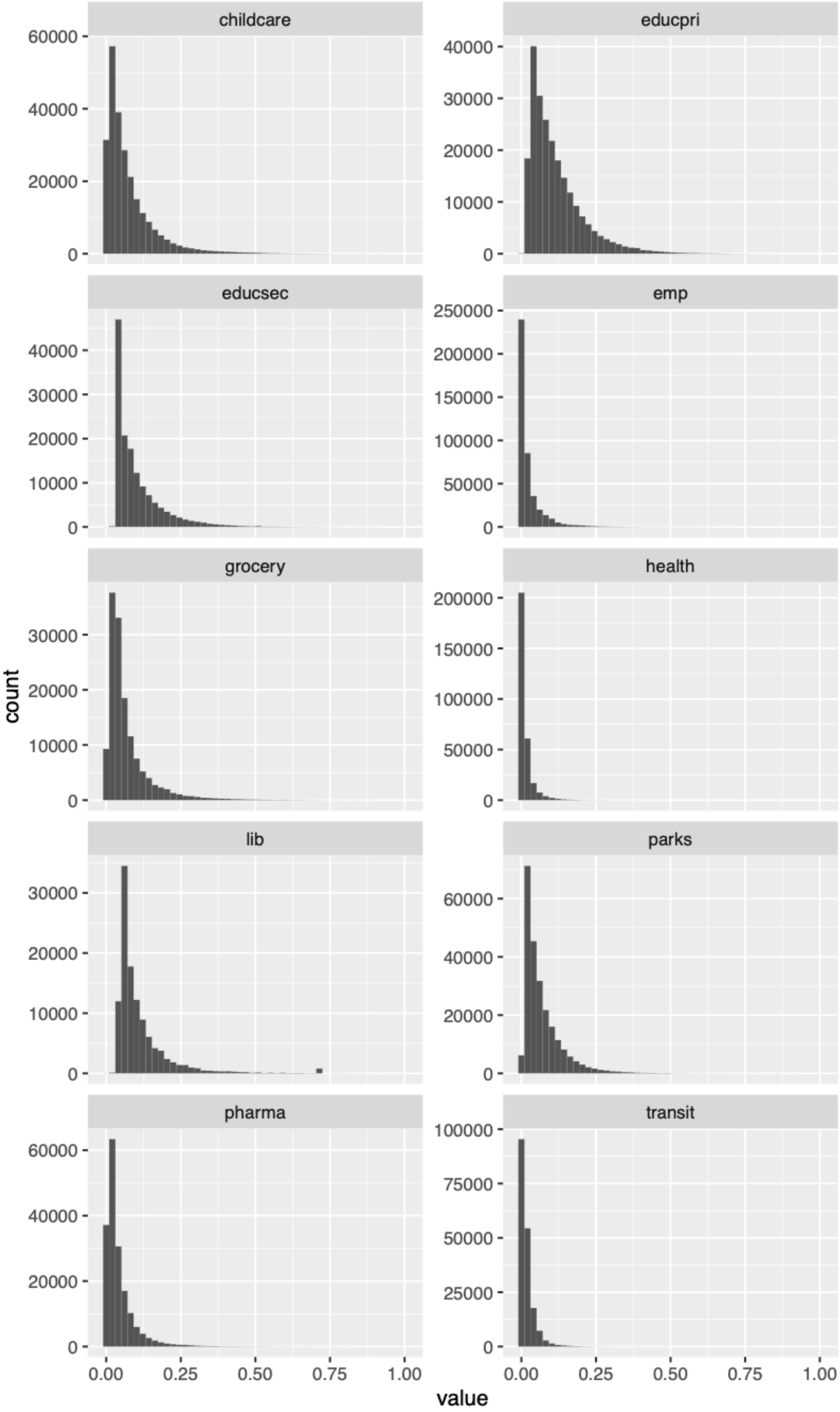
Data Distributions

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

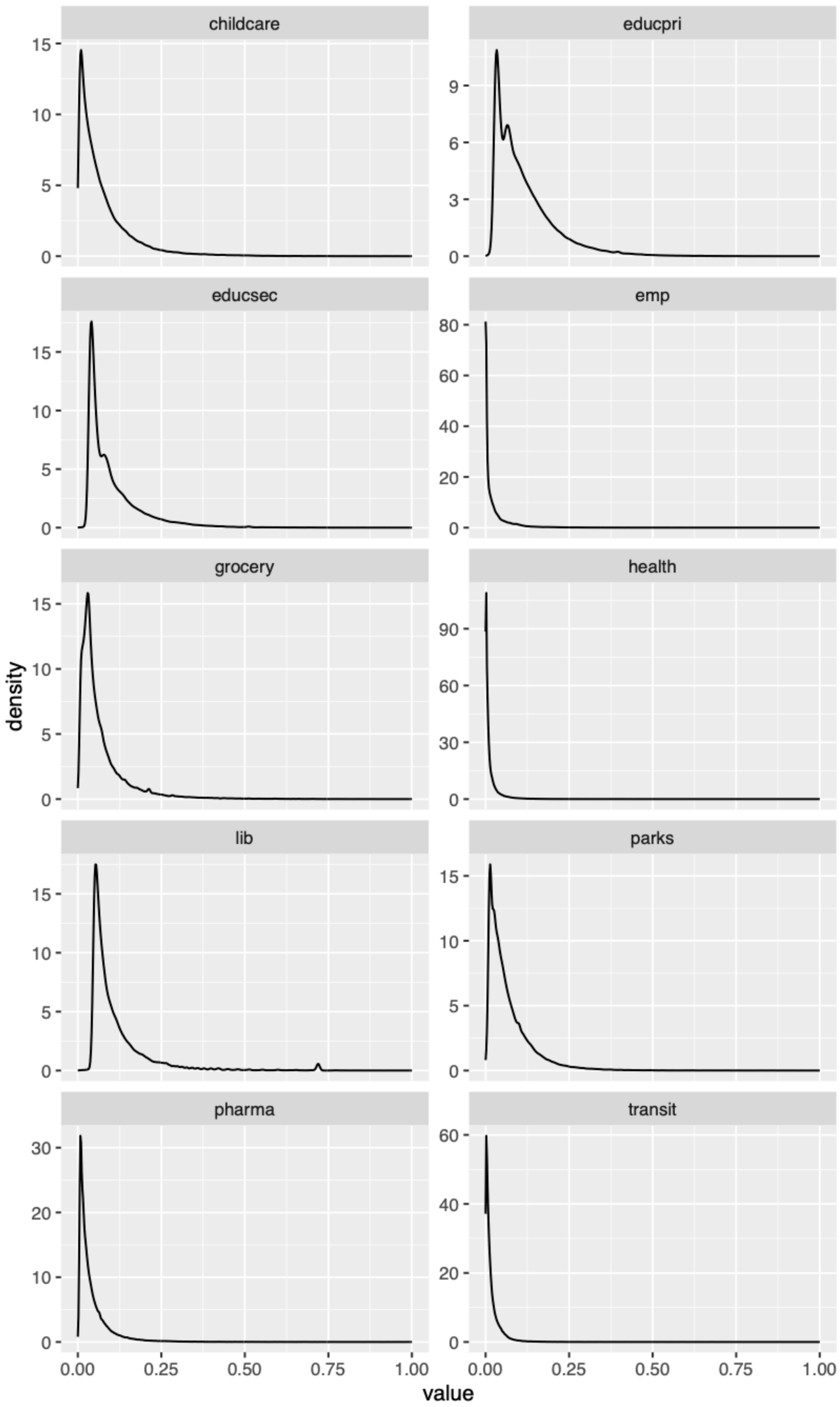
In this violin plot, we see that the highest densities of proximity values lie below 0.12 for all amenities. We see that the amenities with the highest distribution density closer to 0 are health, then employment, then transit. Health and employment have the least amount of missing values, and some conclusion could be made out of that.



Next we see the histograms of proximity values for each amenity. This gives us an idea of the counts for each bin of values. We see that libraries have the least amount of proximity values that are near 0, but we saw above that they also had the greatest amount of missing values.


childcare
educpri
educsec
emp
grocery
health
lib
parks
pharma
transit
count
value
0
20000
40000
60000
10000
30000
50000
100000
150000
200000
250000
25000
75000
0.00
0.25
0.50
0.75
1.00

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 robust clusters, but we will explore if clustering algorithms will.


childcare
educpri
educsec
emp
grocery
health
lib
parks
pharma
transit
density
value
0.00
0.25
0.50
0.75
1.00
0
5
10
15
3
6
9
20
40
60
80
30
90