EDA draft

2023-05-17

Contents

0.1 Import Data

0.1.1 Import PMD data

Glimpse of PMD dataset

```
##         DBUID DBPOP    DAUID DAPOP  CSDUID    CSDNAME CSDTYPE  CSDPOP CMAUID
## 1 10010165001   160 10010165   506 1001519 St. John's      CY 108,860      1
## 2 10010165002    25 10010165   506 1001519 St. John's      CY 108,860      1
## 3 10010165006   268 10010165   506 1001519 St. John's      CY 108,860      1
##   CMAPUID    CMANAME CMATYPE  CMAPOP PRUID
## 1   10001 St. John's       B 205,955    10
## 2   10001 St. John's       B 205,955    10
## 3   10001 St. John's       B 205,955    10
##                                             PRNAME   PRPOP      lon     lat
## 1 Newfoundland and Labrador / Terre-Neuve-et-Labrador 519,716 -52.7765 47.5300
## 2 Newfoundland and Labrador / Terre-Neuve-et-Labrador 519,716 -52.7793 47.5290
## 3 Newfoundland and Labrador / Terre-Neuve-et-Labrador 519,716 -52.7768 47.5265
##   in_db_emp prox_idx_emp in_db_pharma prox_idx_pharma in_db_childcare
## 1         1       0.0202            0          0.0121               0
## 2         1       0.0193            0           0.014               0
## 3         1       0.0199            1          0.0205               1
##   prox_idx_childcare in_db_health prox_idx_health in_db_grocery
## 1             0.0402            0          0.0069             0
## 2             0.0257            0          0.0028             0
## 3             0.0395            1           0.007             0
##   prox_idx_grocery in_db_educpri prox_idx_educpri in_db_educsec
## 1               ..             0           0.0384             0
## 2               ..             0           0.0562             0
## 3               ..             0           0.0734             0
##   prox_idx_educsec in_db_lib prox_idx_lib in_db_parks prox_idx_parks
```

```
## 1           0.0495         0       0.0486           0         0.0141
## 2           0.0375         0           ..           0             ..
## 3           0.0436         0       0.0545           0             ..
##   in_db_transit prox_idx_transit transit_na amenity_dense suppressed
## 1             1           0.0058          0             0          0
## 2             0           0.0046          0             0          0
## 3             1           0.0101          0             0          0
```

0.2 DATA Summary

Structure of the dataset

```
## 'data.frame':    489676 obs. of  41 variables:
##  $ DBUID             : num  1e+10 1e+10 1e+10 1e+10 1e+10 ...
##  $ DBPOP             : chr  "160" "25" "268" "53" ...
##  $ DAUID             : int  10010165 10010165 10010165 10010165 10010166 10010166 10010166 10010167 1
##  $ DAPOP             : chr  "506" "506" "506" "506" ...
##  $ CSDUID            : int  1001519 1001519 1001519 1001519 1001519 1001519 1001519 1001519 1001519 1
##  $ CSDNAME           : chr  "St. John's" "St. John's" "St. John's" "St. John's" ...
##  $ CSDTYPE           : chr  "CY" "CY" "CY" "CY" ...
##  $ CSDPOP            : chr  "108,860" "108,860" "108,860" "108,860" ...
##  $ CMAUID            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ CMAPUID           : int  10001 10001 10001 10001 10001 10001 10001 10001 10001 10001 ...
##  $ CMANAME           : chr  "St. John's" "St. John's" "St. John's" "St. John's" ...
##  $ CMATYPE           : chr  "B" "B" "B" "B" ...
##  $ CMAPOP            : chr  "205,955" "205,955" "205,955" "205,955" ...
##  $ PRUID             : int  10 10 10 10 10 10 10 10 10 10 ...
##  $ PRNAME            : chr  "Newfoundland and Labrador / Terre-Neuve-et-Labrador" "Newfoundland and L
##  $ PRPOP             : chr  "519,716" "519,716" "519,716" "519,716" ...
##  $ lon               : num  -52.8 -52.8 -52.8 -52.8 -52.8 ...
##  $ lat               : num  47.5 47.5 47.5 47.5 47.5 ...
##  $ in_db_emp         : chr  "1" "1" "1" "1" ...
##  $ prox_idx_emp      : chr  "0.0202" "0.0193" "0.0199" "0.0204" ...
##  $ in_db_pharma      : chr  "0" "0" "1" "0" ...
##  $ prox_idx_pharma   : chr  "0.0121" "0.014" "0.0205" "0.0238" ...
##  $ in_db_childcare   : chr  "0" "0" "1" "0" ...
##  $ prox_idx_childcare: chr  "0.0402" "0.0257" "0.0395" "0.0425" ...
##  $ in_db_health      : chr  "0" "0" "1" "0" ...
##  $ prox_idx_health   : chr  "0.0069" "0.0028" "0.007" "0.0074" ...
##  $ in_db_grocery     : chr  "0" "0" "0" "0" ...
##  $ prox_idx_grocery  : chr  ".." ".." ".." ".." ...
##  $ in_db_educpri     : chr  "0" "0" "0" "0" ...
##  $ prox_idx_educpri  : chr  "0.0384" "0.0562" "0.0734" "0.0733" ...
##  $ in_db_educsec     : chr  "0" "0" "0" "0" ...
##  $ prox_idx_educsec  : chr  "0.0495" "0.0375" "0.0436" "0.0548" ...
##  $ in_db_lib         : chr  "0" "0" "0" "0" ...
##  $ prox_idx_lib      : chr  "0.0486" ".." "0.0545" "0.0796" ...
##  $ in_db_parks       : chr  "0" "0" "0" "0" ...
##  $ prox_idx_parks    : chr  "0.0141" ".." ".." "0.013" ...
##  $ in_db_transit     : chr  "1" "0" "1" "0" ...
##  $ prox_idx_transit  : chr  "0.0058" "0.0046" "0.0101" "0.0098" ...
##  $ transit_na        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ amenity_dense     : chr  "0" "0" "0" "0" ...
```

	DBPOP	DAPOP	CSDPOP	CMAPOP	PRPOP	lon	lat
	Min. : 0.00	Min. : 0.0	Min. : 0	Min. : 10741	Min. : 35874	Min. :-140.91	Min. :
	1st Qu.: 5.00	1st Qu.: 431.0	1st Qu.: 2559	1st Qu.: 103811	1st Qu.: 1278365	1st Qu.:-110.03	1st Qu
	Median : 29.00	Median : 523.0	Median : 13678	Median : 747545	Median : 4648055	Median : -80.53	Media
	Mean : 71.83	Mean : 727.8	Mean : 235937	Mean :1670003	Mean : 6748994	Mean : -90.30	Mean
	3rd Qu.: 81.00	3rd Qu.: 675.0	3rd Qu.: 134413	3rd Qu.:2463431	3rd Qu.:13448494	3rd Qu.: -73.78	3rd Q
	Max. :7607.00	Max. :22077.0	Max. :2731571	Max. :5928040	Max. :13448494	Max. : -52.66	Max.
	NA's :315	NA's :315	NA's :315	NA's :212914	NA	NA	NA

```
##  $ suppressed        : int  0 0 0 0 0 0 0 0 0 0 ...
```

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.

```
length(unique(pmd$DBUID))
```

```
## [1] 489676
```

There are no duplicate rows as the number of unique dissemination block id is same as the number of rows in the dataset.

There are some features which are currently in character type but needs to be numeric such as proximity indices, population. Will convert these features datatypes.

Summary of the dataset

0.3 Missing Values

0.3.1 Missing values percentage

In glimpse of dataset we saw there were missing values in `prox_idx_lib` but the above output suggest there is no missing values. Because Statistics Canada use some specific notation for missing values. The following standard symbols are used in Statistics Canada publications:
..− > not available for a specific reference period
$F-$ > to unreliable to be published By changing these symbols to NA missing percentages are

We can see that the library proximity indicator contains the most missing values, almost 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.

0.3.2 All proximity measures are NA

```
## No. of rows: All proximity measures are NA 64764
```

So, there are 64764 dissemination blocks where none of the proximity measures are available. Let's check the population of those dissemination blocks.

	x
prox_idx_lib	76.9939715
prox_idx_grocery	71.1925845
prox_idx_educsec	71.1619520
prox_idx_pharma	63.5430366
prox_idx_transit	62.9744974
prox_idx_educpri	53.9779364
prox_idx_parks	52.1994135
prox_idx_childcare	50.1784854
CMAUID	43.4805872
CMAPUID	43.4805872
CMAPOP	43.4805872
prox_idx_health	38.6400395
prox_idx_emp	13.4934120
DBPOP	0.0643282
DAPOP	0.0643282
CSDPOP	0.0643282
DBUID	0.0000000
DAUID	0.0000000
CSDUID	0.0000000
CSDNAME	0.0000000
CSDTYPE	0.0000000
CMANAME	0.0000000
CMATYPE	0.0000000
PRUID	0.0000000
PRNAME	0.0000000
PRPOP	0.0000000
lon	0.0000000
lat	0.0000000
in_db_emp	0.0000000
in_db_pharma	0.0000000
in_db_childcare	0.0000000
in_db_health	0.0000000
in_db_grocery	0.0000000
in_db_educpri	0.0000000
in_db_educsec	0.0000000
in_db_lib	0.0000000
in_db_parks	0.0000000
in_db_transit	0.0000000
transit_na	0.0000000
amenity_dense	0.0000000
suppressed	0.0000000

	x
prox_idx_lib	76.9939715
prox_idx_grocery	71.1925845
prox_idx_educsec	71.1619520
prox_idx_pharma	63.5430366
prox_idx_transit	62.9744974
prox_idx_educpri	53.9779364
prox_idx_parks	52.1994135
prox_idx_childcare	50.1784854
CMAUID	43.4805872
CMAPUID	43.4805872
CMAPOP	43.4805872
prox_idx_health	38.6400395
prox_idx_emp	13.4934120
in_db_emp	1.0960308
in_db_pharma	1.0960308
in_db_childcare	1.0960308
in_db_health	1.0960308
in_db_grocery	1.0960308
in_db_educpri	1.0960308
in_db_educsec	1.0960308
in_db_lib	1.0960308
in_db_parks	1.0960308
in_db_transit	1.0960308
amenity_dense	1.0960308
DBPOP	0.0643282
DAPOP	0.0643282
CSDPOP	0.0643282
DBUID	0.0000000
DAUID	0.0000000
CSDUID	0.0000000
CSDNAME	0.0000000
CSDTYPE	0.0000000
CMANAME	0.0000000
CMATYPE	0.0000000
PRUID	0.0000000
PRNAME	0.0000000
PRPOP	0.0000000
lon	0.0000000
lat	0.0000000
transit_na	0.0000000
suppressed	0.0000000

```
unique(all_prox_na$DBPOP)[1:5]
```

```
## [1]  0 33 78 30 19
```

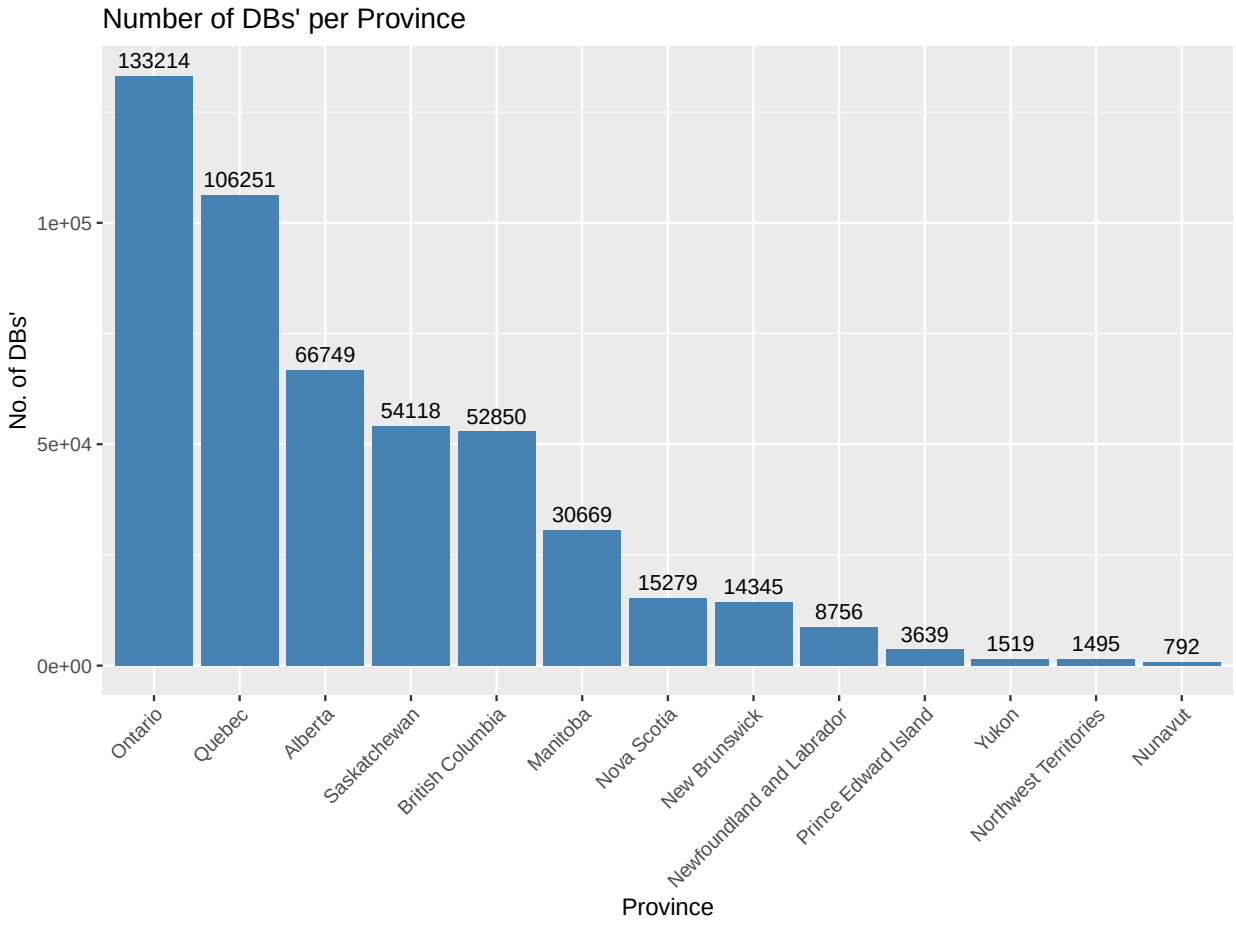
```
sort(unique(all_prox_na$DBPOP), decreasing = TRUE)[1:5]
```

```
## [1] 1858 1522 1501 1404 1265
```

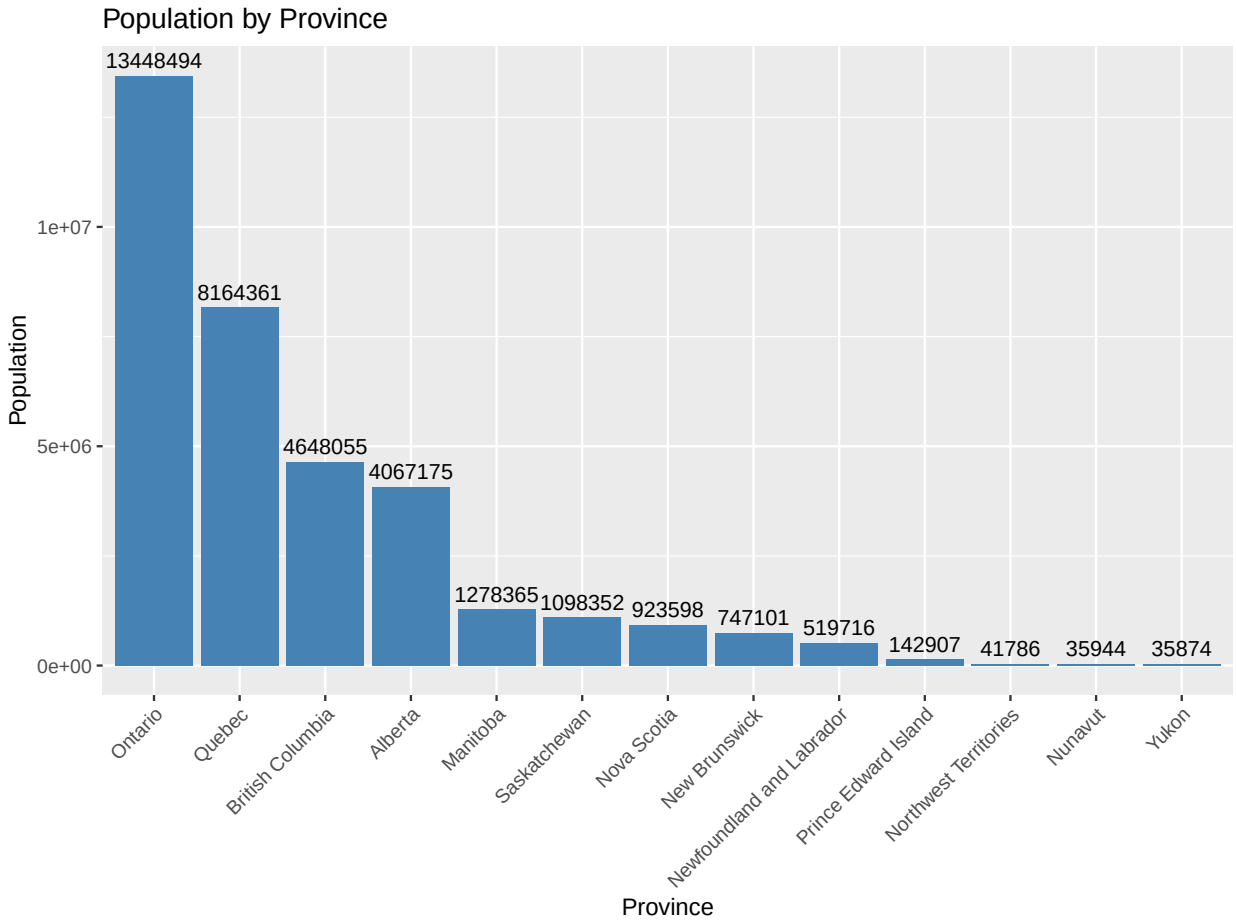
There are DB's with large population that has no proximity measures. The aminities might be not in range that's why its not calculated.

0.4 Distributions

0.4.0.1 Dissemination block per province.

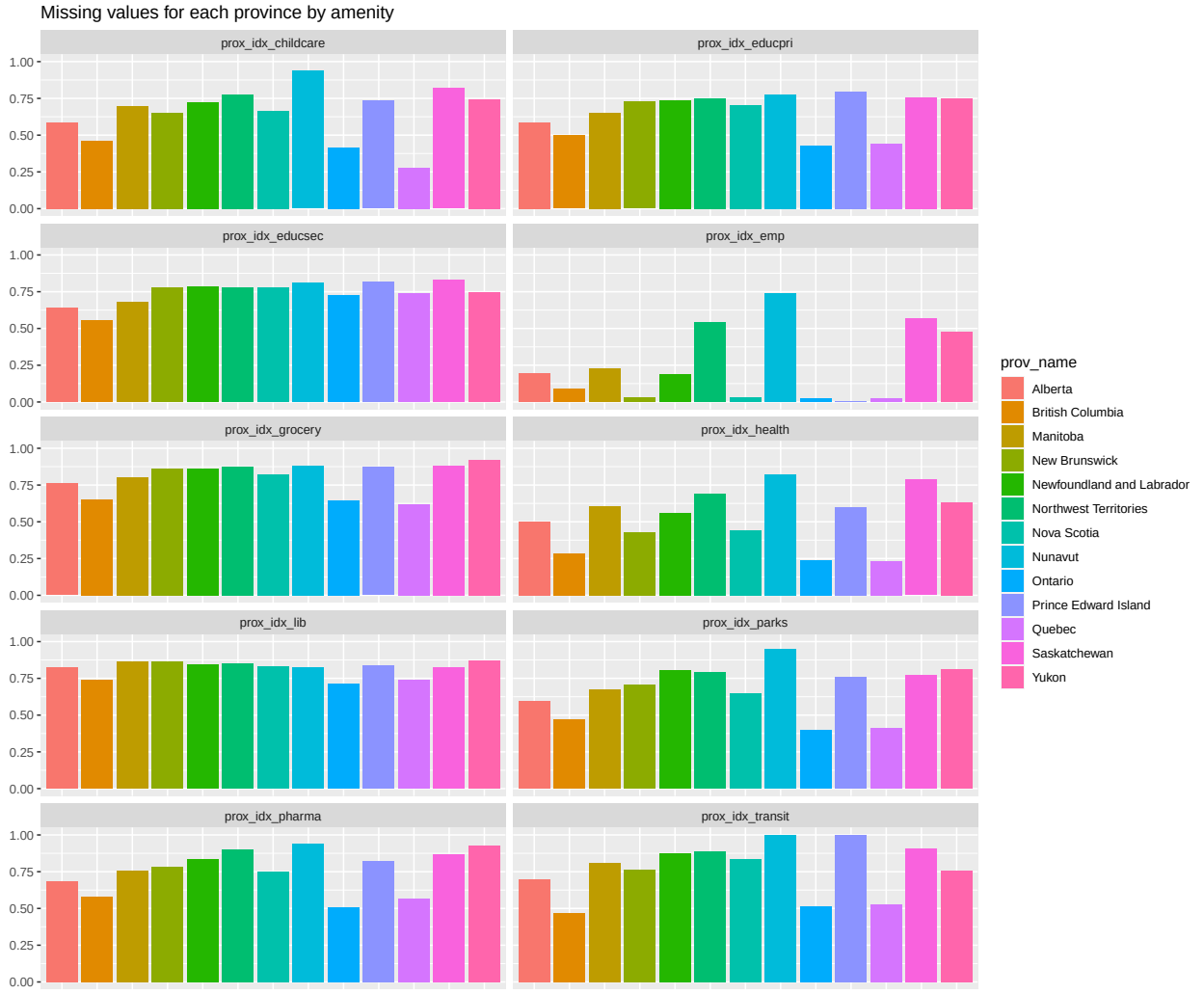


0.4.0.2 Population per Province



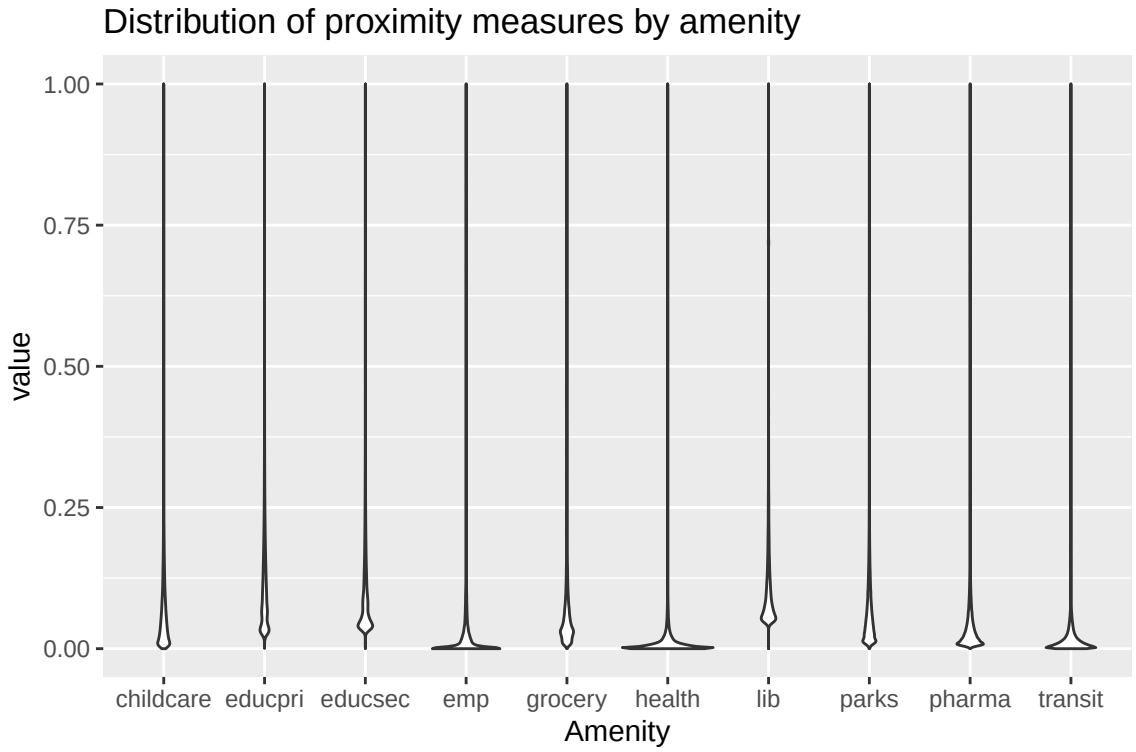
There's a large population gap between BC and Saskatchewan but both has approximately same no. of DBs'.

0.4.0.3 Proportion of missing values for each amenity by province.



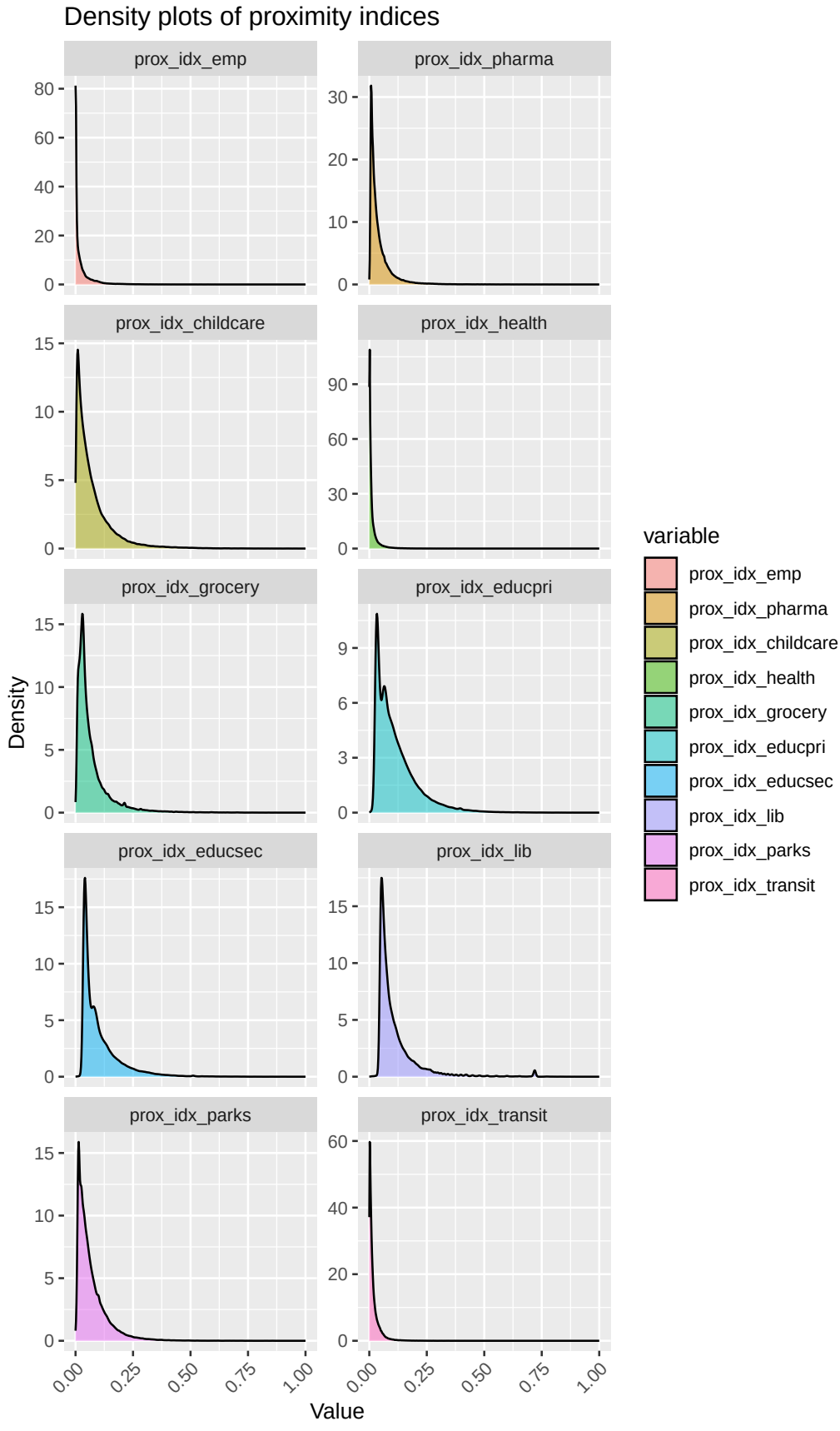
We see that overall, employment has the lowest rates of missing values, but has also more range depending on the province. Ontario and Quebec seems to have the least missing values for most amenities relative to the other regions, whereas Nunavut usually has the most. It seems like the amount of proximity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

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



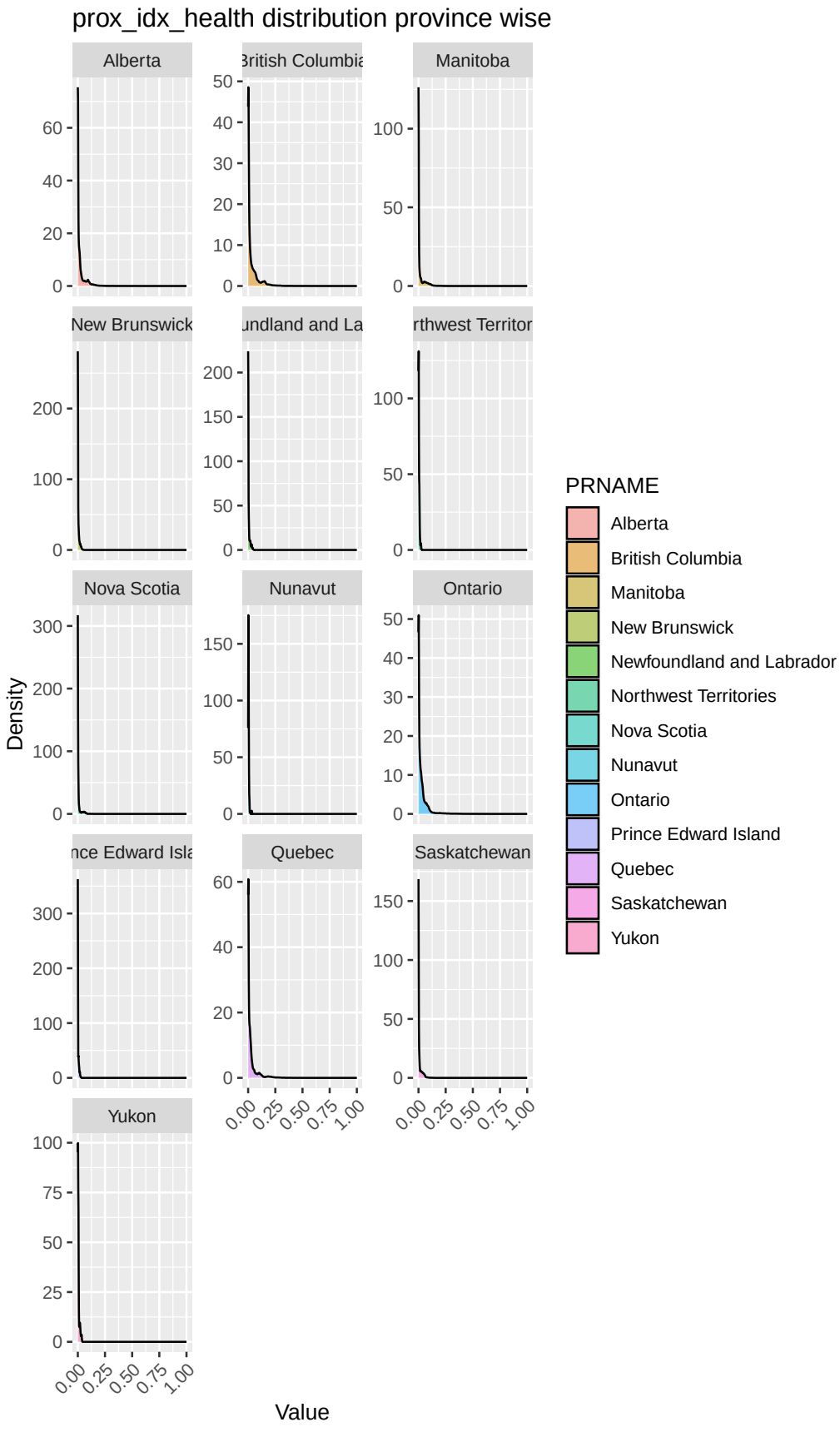
In this violin plot, we see that the highest densities of proximity values lie below 0.12 for all amenities. We see that the amenities with the highest distribution density closer to 0 are health, then employment, then transit. Library has the lowest density right around zero, and 'starts' a bit later.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

0.4.0.5 Kernel Density plots of Proximity Measures Next we see the kernel densities of proximity measures for each amenity.

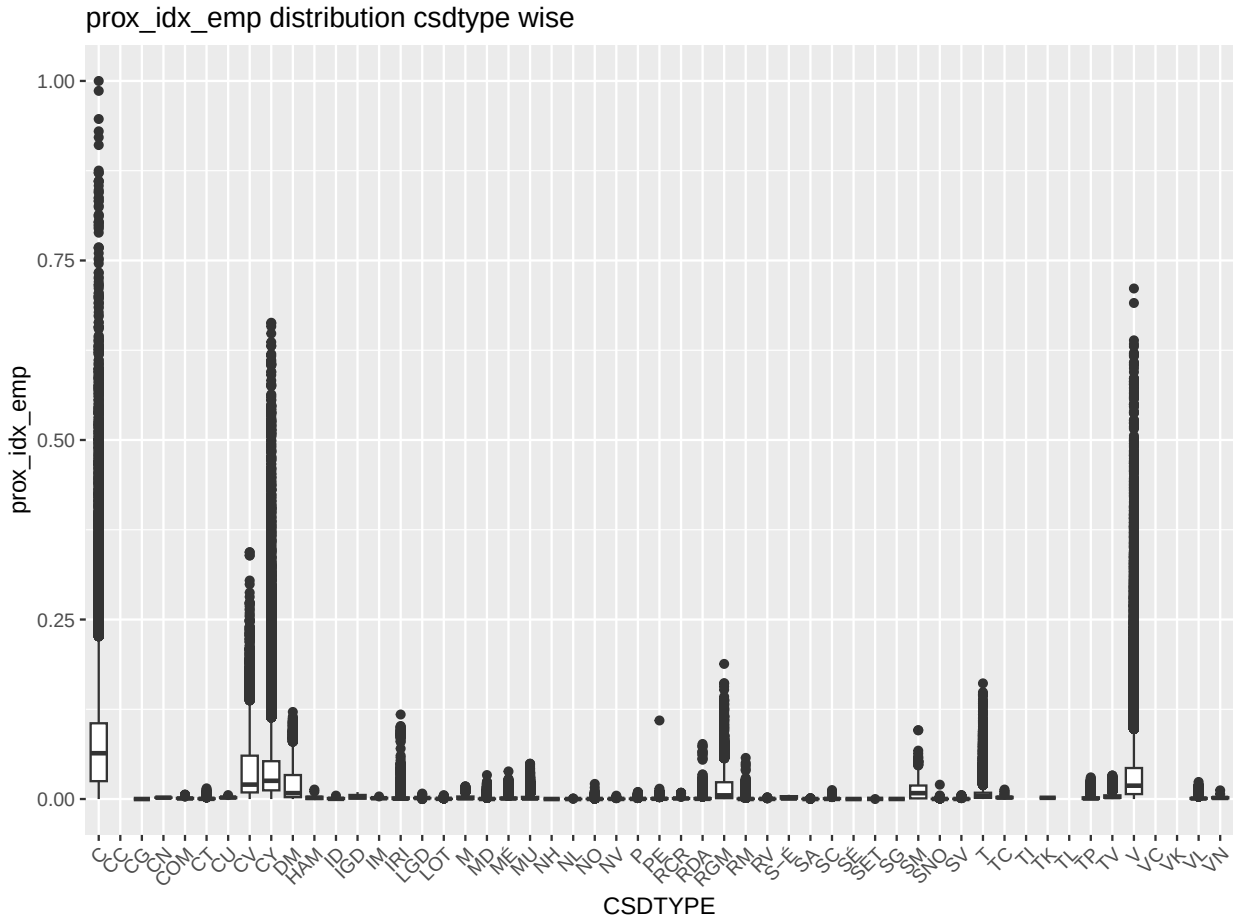

Density plots of proximity indices
prox_idx_emp
prox_idx_pharma
prox_idx_childcare
prox_idx_health
prox_idx_grocery
prox_idx_educpri
prox_idx_educsec
prox_idx_lib
prox_idx_parks
prox_idx_transit
Density
Value
0.00
0.25
0.50
0.75
1.00
variable
prox_idx_emp
prox_idx_pharma
prox_idx_childcare
prox_idx_health
prox_idx_grocery
prox_idx_educpri
prox_idx_educsec
prox_idx_lib
prox_idx_parks
prox_idx_transit

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.

Kernel density plot of prox_idx_health by provinces

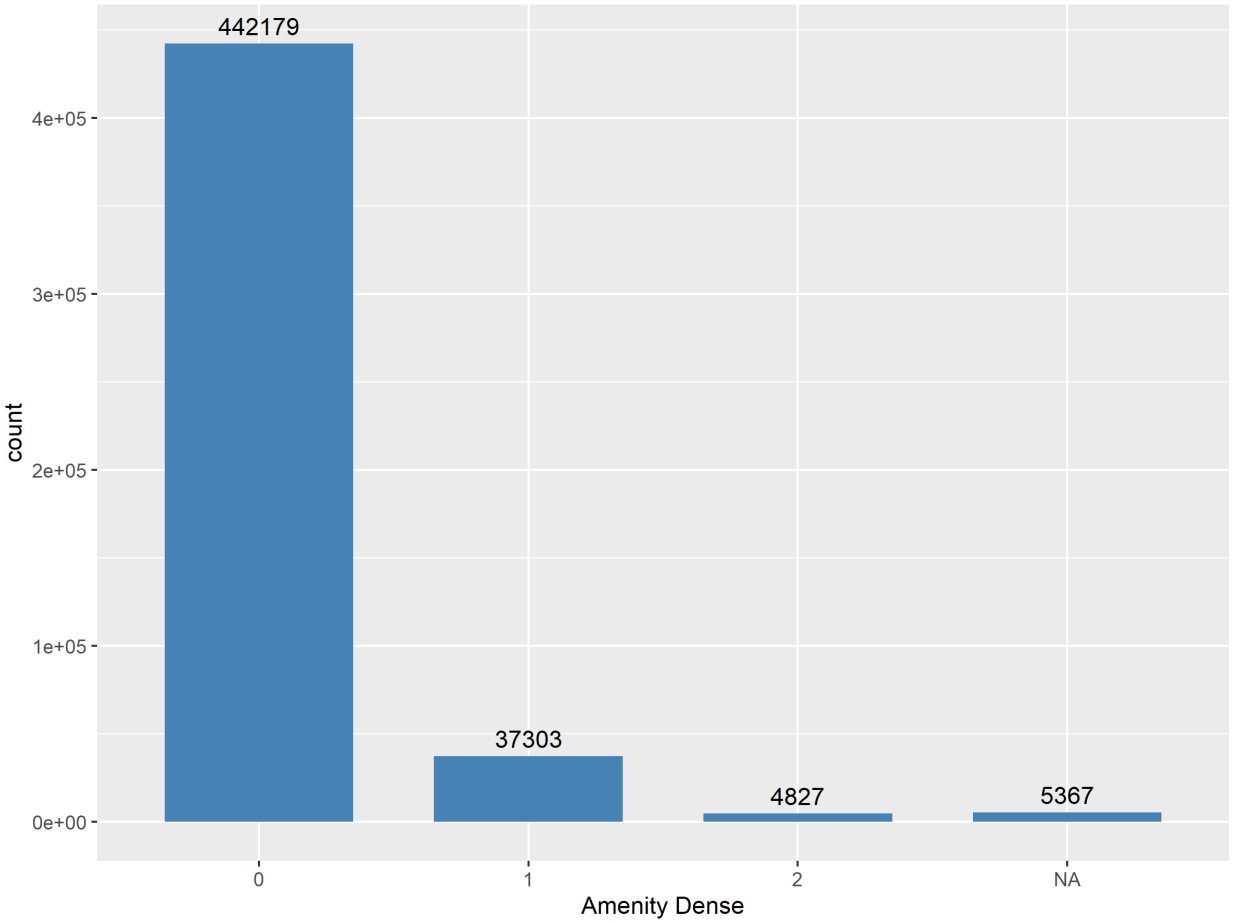

prox_idx_health distribution province wise
Alberta
British Columbia
Manitoba
New Brunswick
Newfoundland and La
rthwest Territor
Nova Scotia
Nunavut
Ontario
nce Edward Isla
Quebec
Saskatchewan
Yukon
Density
Value
PRNAME
Alberta
British Columbia
Manitoba
New Brunswick
Newfoundland and Labrador
Northwest Territories
Nova Scotia
Nunavut
Ontario
Prince Edward Island
Quebec
Saskatchewan
Yukon

prox_idx_emp distribution csdtype wise



On average, the proximity of employees is lower in most of the CSDs. However, cities, towns, and district municipalities have higher average proximity measures of employee. This observation is understandable as these areas tend to offer more job opportunities.

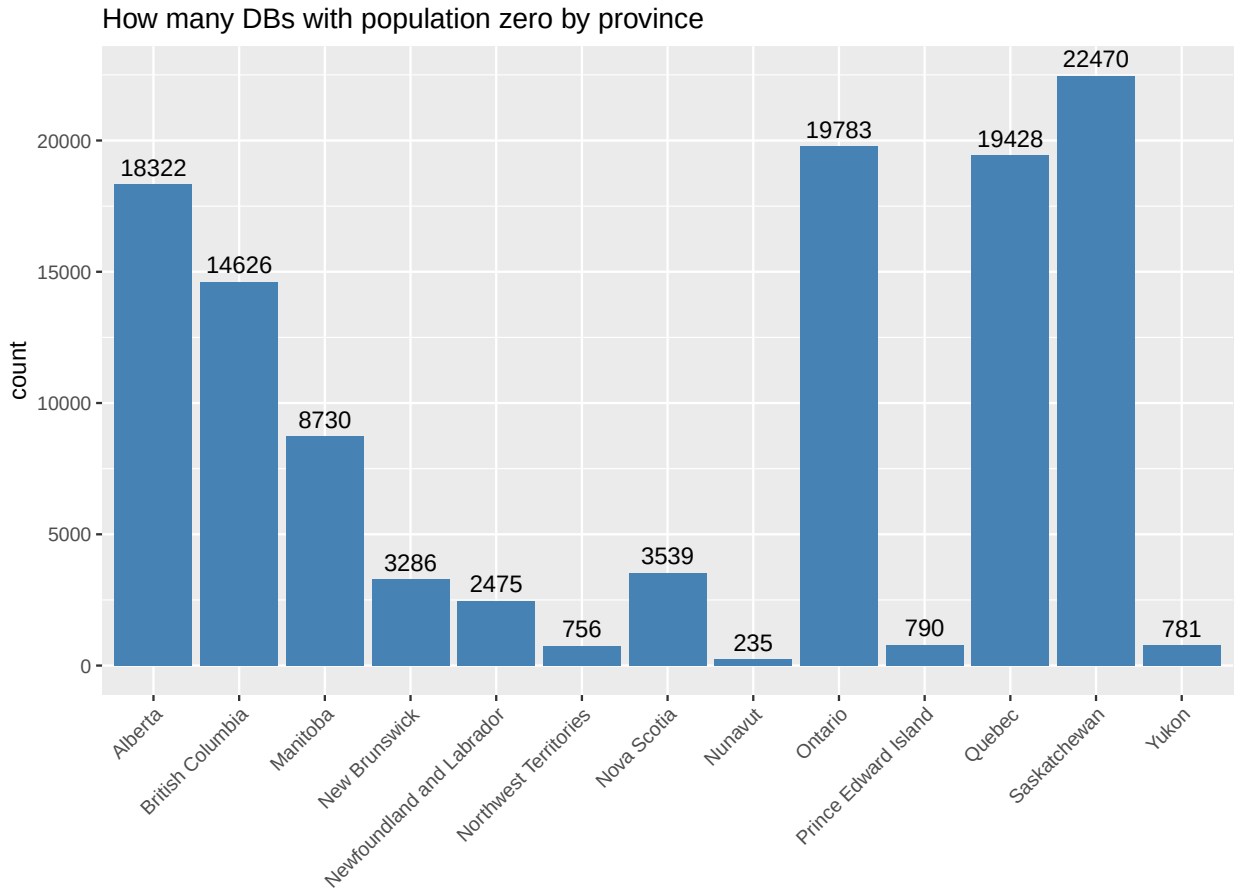
0.4.0.6 Distribution of amenitie_dense



37303 DBs' are in an amenity dense neighbourhood, 4827 DBs' are in a high amenity density neighbourhood, 442179 DBs' are in a non-amenity dense neighbourhood.

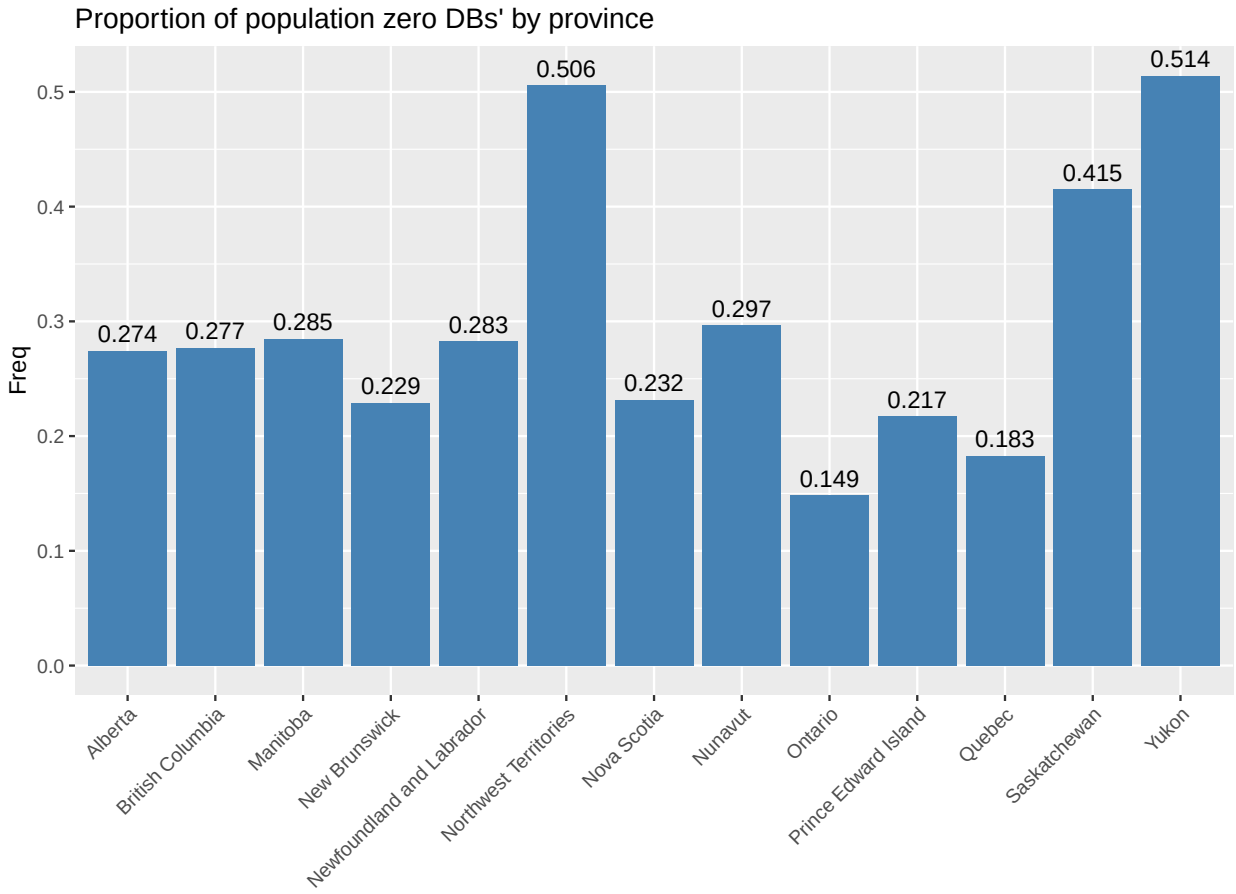
0.4.0.7 Population zero About 24% of the DBs in Canada (in 2016) have a population of 0. It could be reasonable to expect that if the population of a DB is 0, then the proximity measure are also near 0: it is intuitive that for the most part, amenities are further away from areas with no populations. 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.

DBs with a population of zero by Province



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

Proportion of population zero DBs' by province



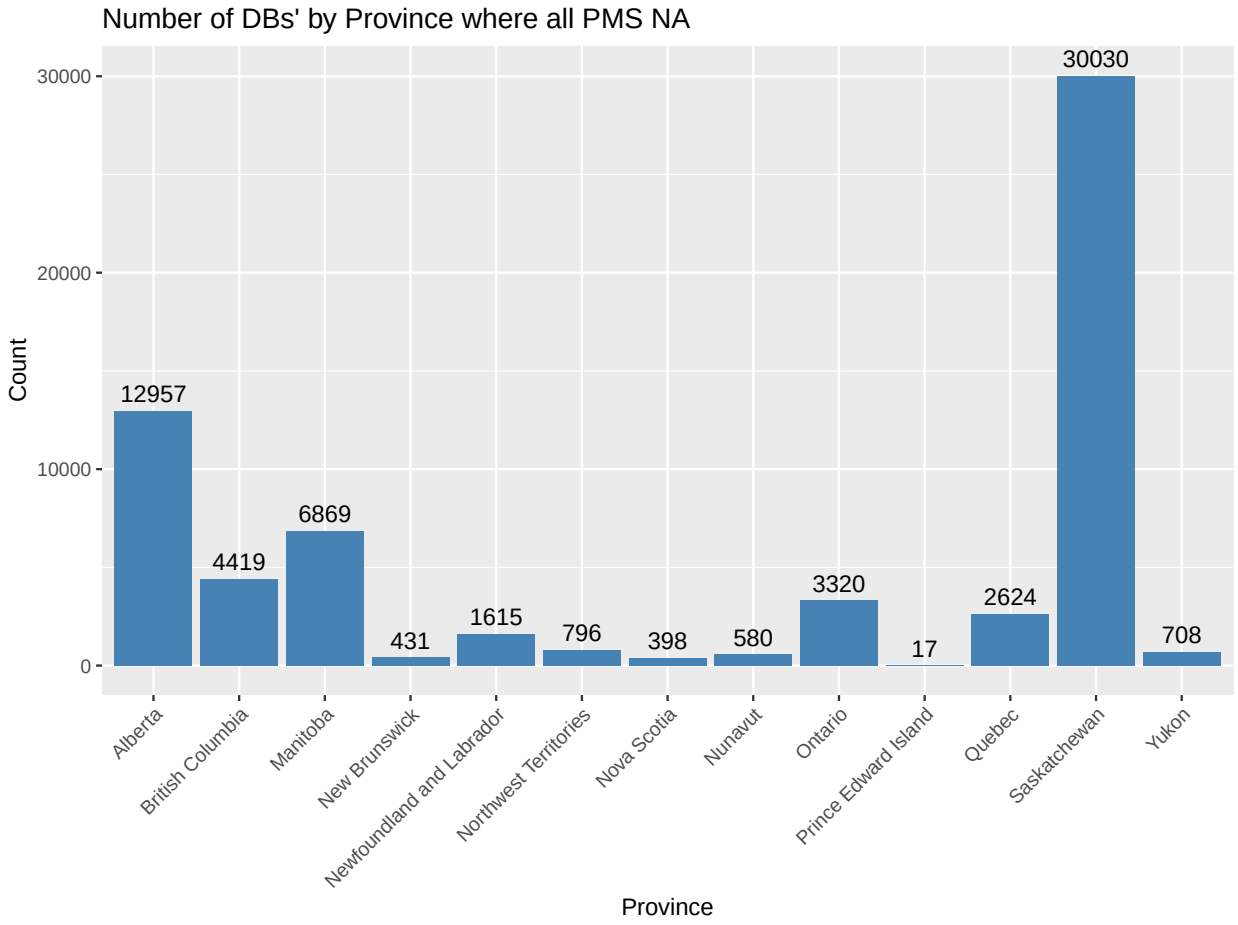
Taking the proportions however, we see that over 50% of Yukon and NWT's DBs have a population of 0, and Saskatchewan has over 40%. Ontario has the lowest at around 15%, followed by Quebec at around 18%.

DB counts of zero populations by CSDTYPE

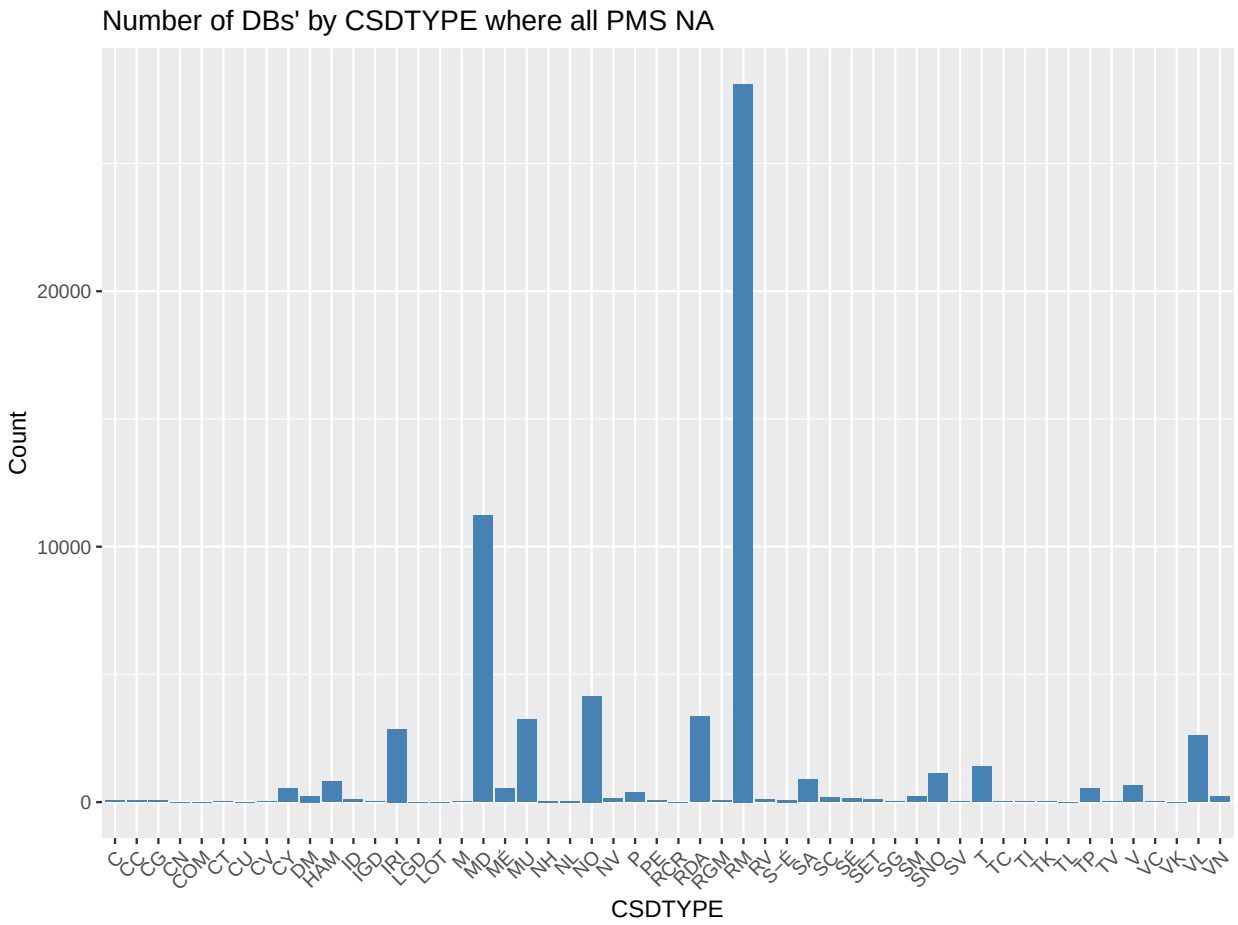
```
##
##    RM    CY     V    MD     T   RDA    MU    MÉ    NO    TP     C   IRI    VL
## 20926 18638 12454  9859  8423  7077  4723  4568  4013  3569  3023  2613  2425
##     P    DM    CV    SC   RGM   SNO    SM    PE    SA   LOT    TV   HAM    CT
##  1634  1582  1309  1277  1231  1127   701   671   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
```

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.

DBs' by Province where all PMS NA



DBs' by CSDTYPE where all PMS NA



We can see that These DBs are from different province and different csd. If we can plot them on map then it may make sense! Majority portions of db where all proximity measures are null from Regional municipality and Municipal district.

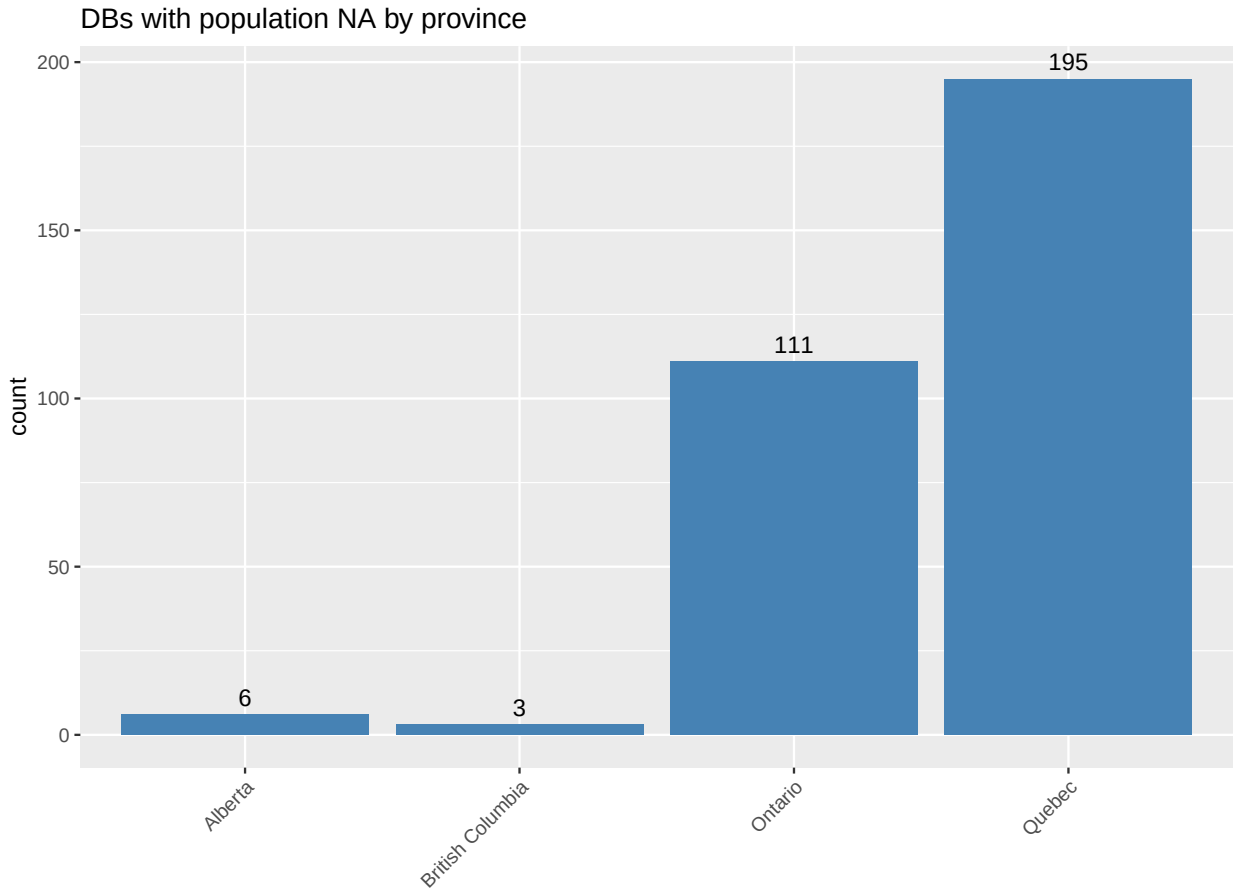
DBs' for which all the proximity measures of amenities are 0

```
# Check if there are any dbs where all proximity measures are 0
all_prox_0 <- pmd[rowSums(pmd[, prox_cols] == 0, na.rm = TRUE) == length(prox_cols), ]
nrow(all_prox_0)
```

```
## [1] 0
```

There are no dissemination blocks for which all the proximity measures of amenities are 0. So, dissemination blocks with no populations still has the proximity measures of amenities may be those dissemination blocks are close from other populated dissemination blocks, or the dissemination blocks contain parks, office buildings, industrial area, under construction etc.

DBs with population NA by province



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.

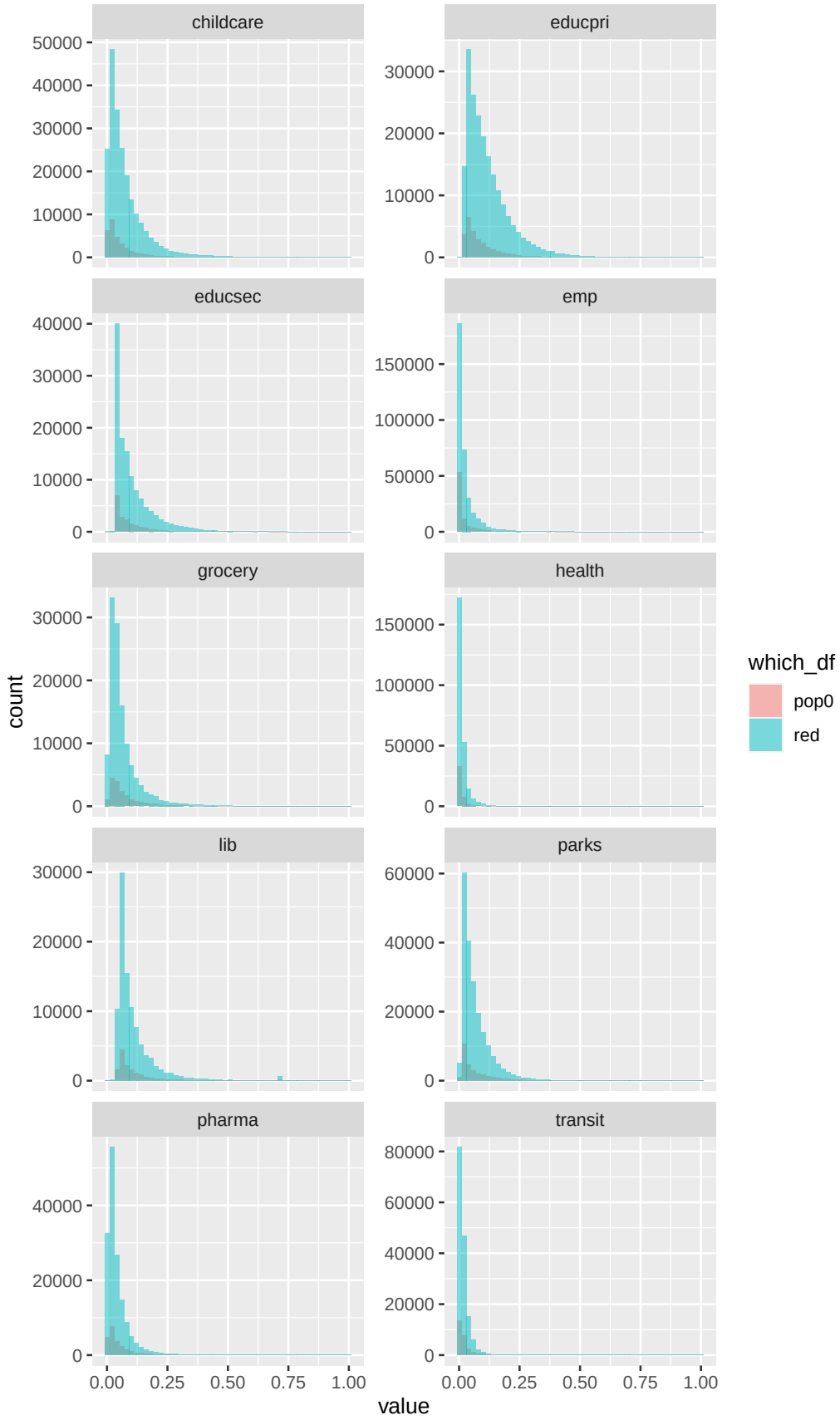
0.4.0.8 Effect of removing population = 0

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 ?)

```
##       prox_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"           "1.44686656480264e-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

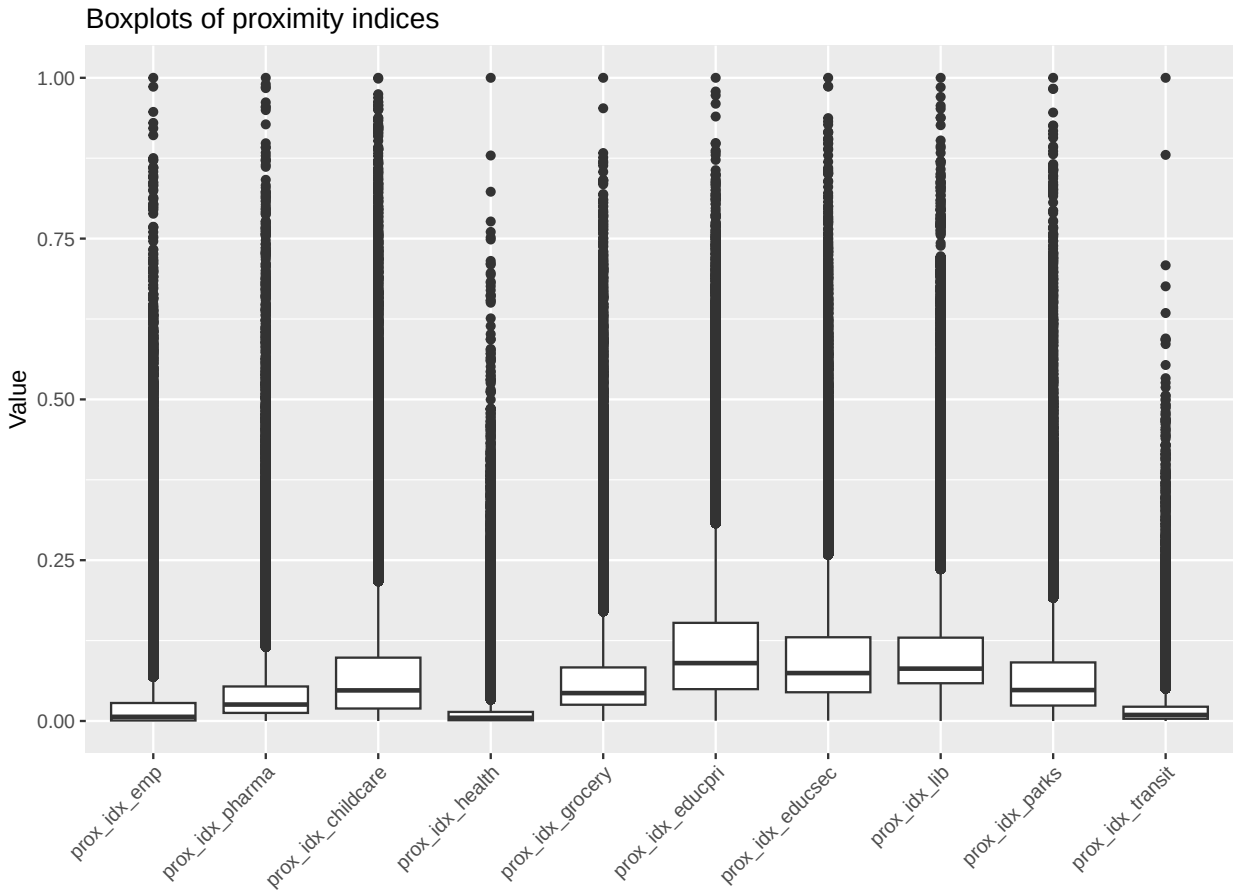
From this following table, we see that 72% of the proximity measure values where population = 0 are NA, compared to 50% of those where population !=0.

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

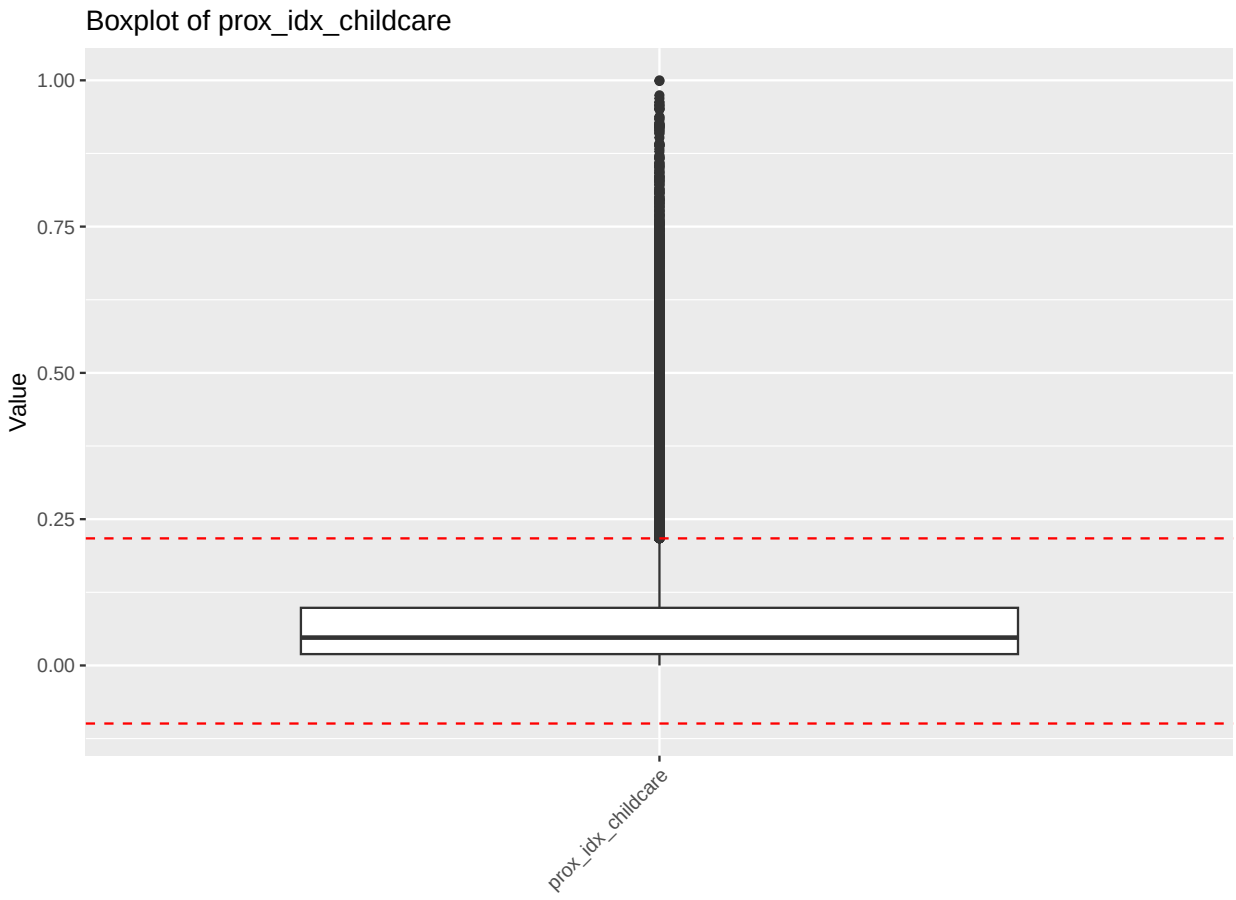
0.4.1 Outliers

Outliers can have a significant impact on the clustering results by pulling the centroids towards themselves, creating biased clusters, and reducing the effectiveness of the clustering algorithm.

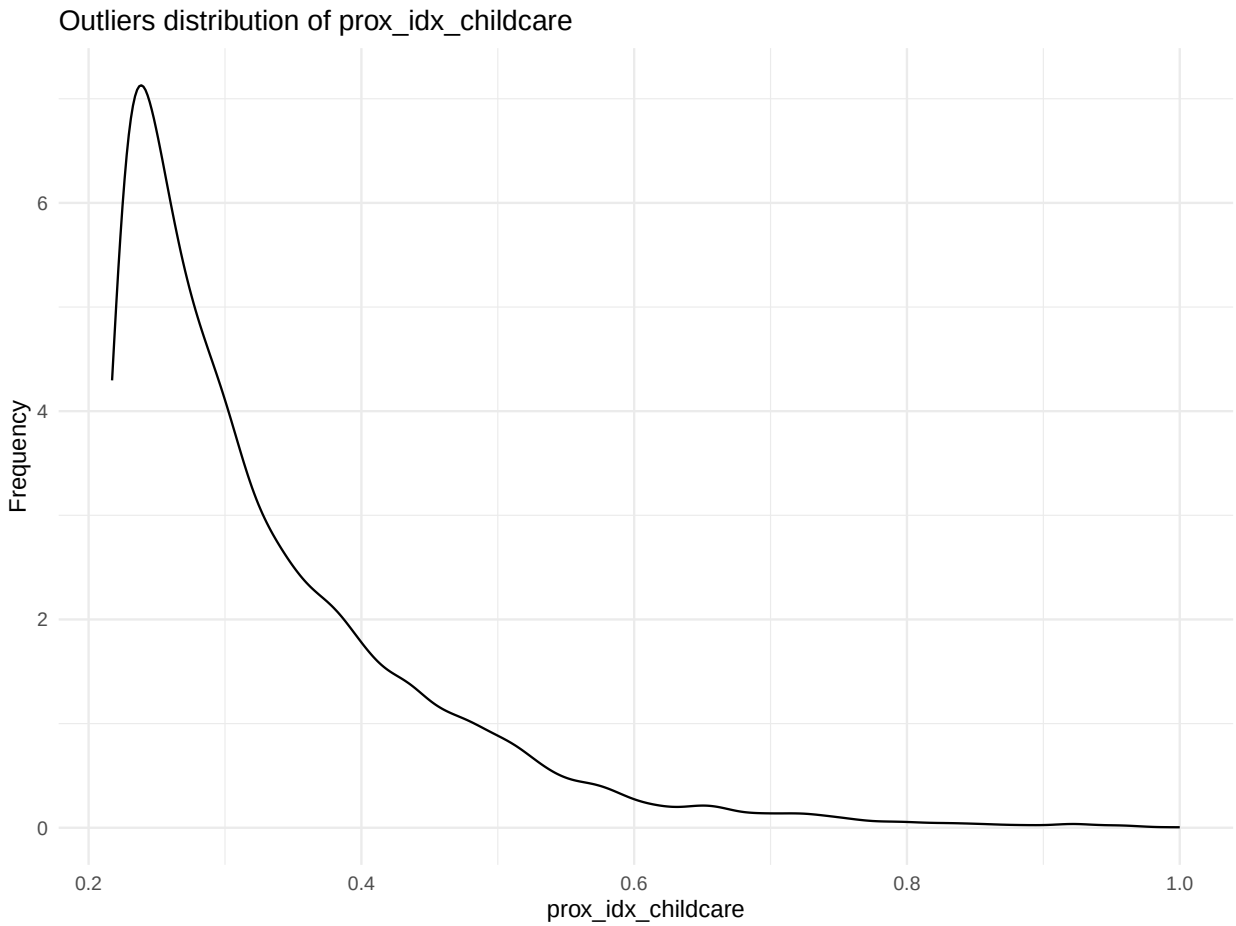
Boxplots of proximity indices



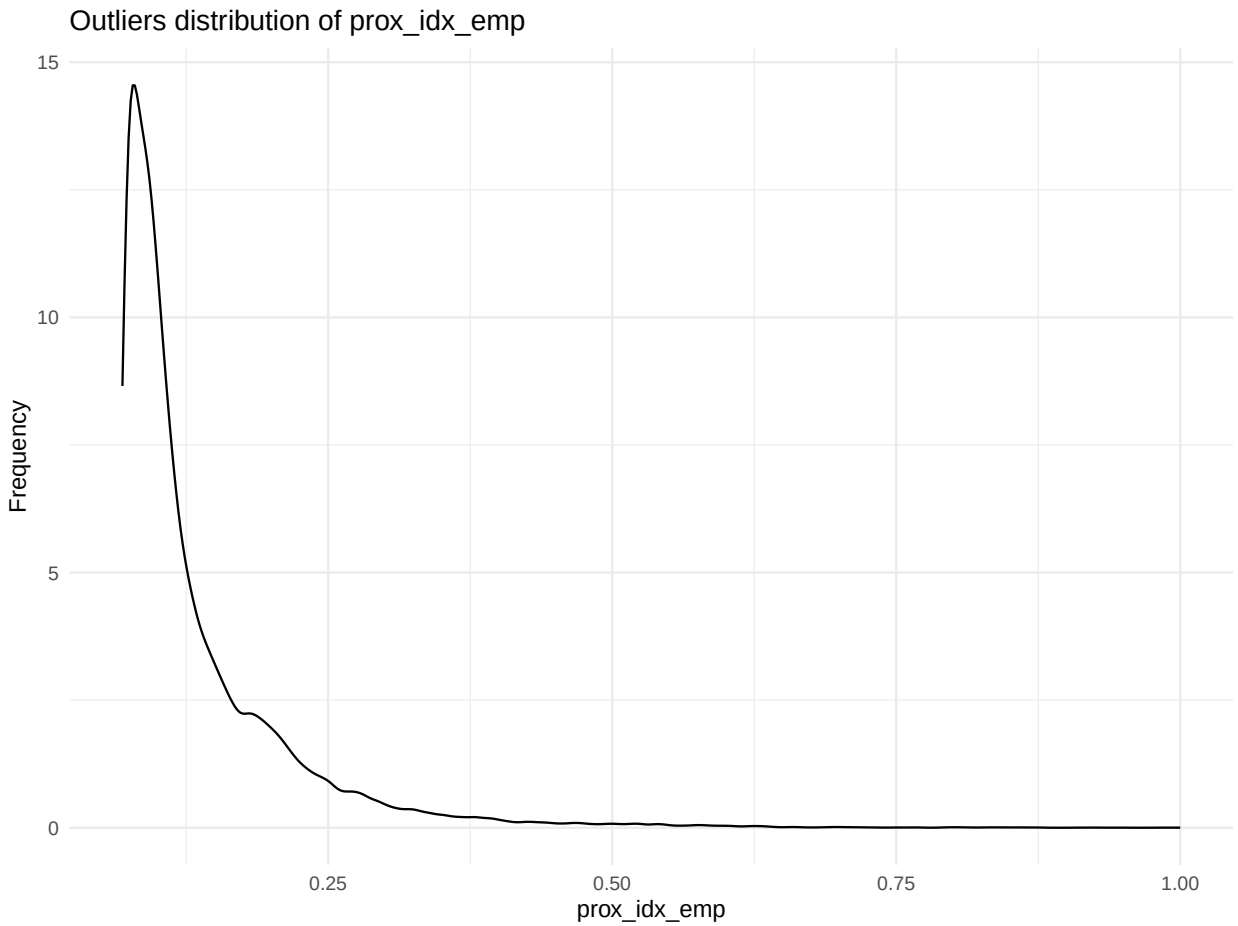
Boxplot of prox_idx_childcare with outlier indication

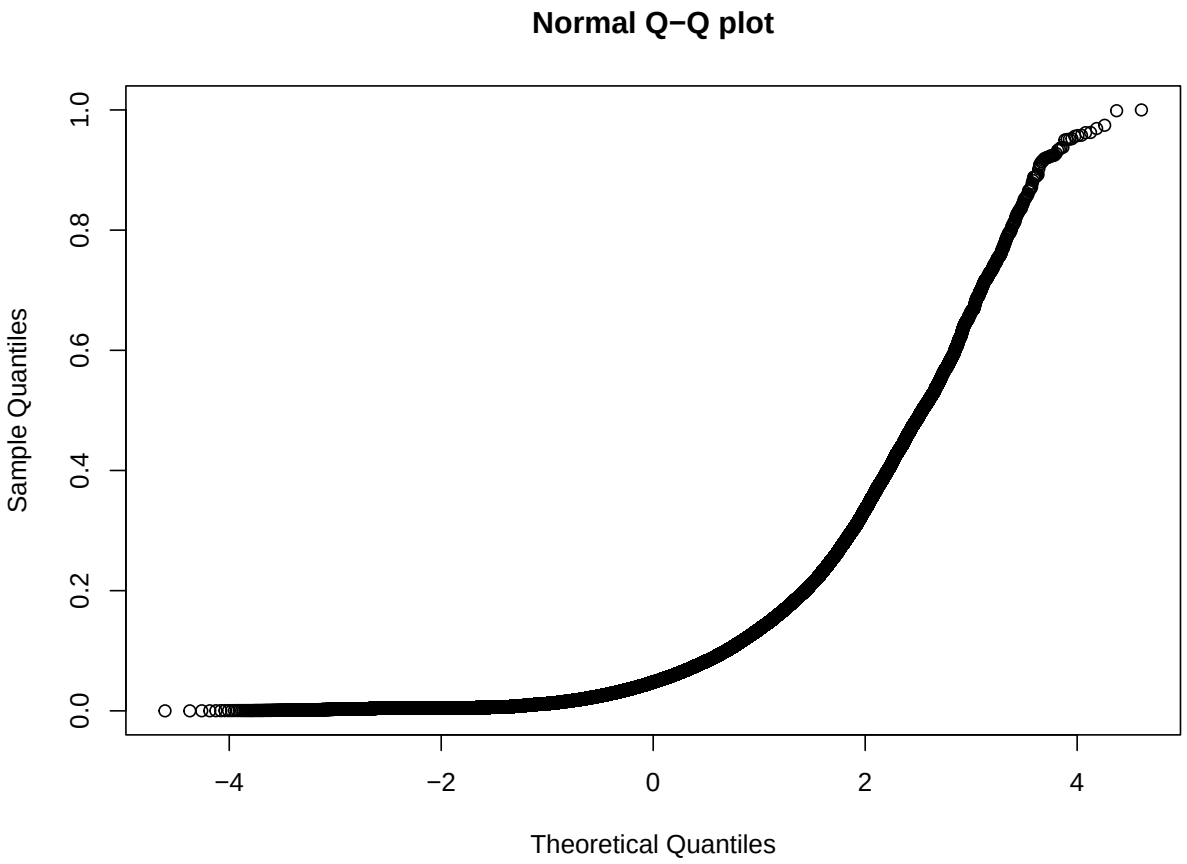


Outliers distribution of prox_idx_childcare



Outliers distribution of prox_idx_emp


Outliers distribution of prox_idx_emp
15
10
5
0
Frequency
0.25
0.50
0.75
1.00
prox_idx_emp



It's hard to detect the outliers from Normal Q-Q plot.

0.4.1.1 Rosner's test for Outliers

Let's do the Rosner's test.

Rosner's test for multiple outliers is used by VSP to detect up to 10 outliers among the selected data values. This test will detect outliers that are either much smaller or much larger than the rest of the data. Rosner's approach is designed to avoid the problem of masking, where an outlier that is close in value to another outlier can go undetected.

Rosner's test is appropriate only when the data, excluding the suspected outliers, are approximately normally distributed, and when the sample size is greater than or equal to 25. (https://vsp.pnnl.gov/help/vsample/rosners_outlier_test.htm)

Just to check we set k=5000, so it will look for 5000 outliers in the data.

```
# Rosner's test
library(EnvStats)
test <- rosnerTest(pmd$prox_idx_emp, k = 5000, warn = TRUE)
```

```
length(test$all.stats$Outlier == TRUE)
```

```
## [1] 5000
```

As in the resnor's test output we find 5000 outlier so there might be more in our data. It's a proof that there are too many outliers in the data.

```
library(outliers)
chisq.out.test(pmd$prox_idx_emp)
```

0.4.1.2 Chi-squared test for Outliers

```
##
##  chi-squared test for outlier
##
## data:  pmd$prox_idx_emp
## X-squared = 394.08, p-value < 2.2e-16
## alternative hypothesis: highest value 1 is an outlier
```

So the max proximity measure of grocery is an outlier.

```
chisq.out.test(pmd$prox_idx_emp, opposite = TRUE)
```

```
##
##  chi-squared test for outlier
##
## data:  pmd$prox_idx_emp
## X-squared = 0.26787, p-value = 0.6048
## alternative hypothesis: lowest value 0 is an outlier
```

So the min proximity measure of grocery is an outlier.

The proximity measures are already normalized but we can still see outliers in these. So, we should use clustering algorithms that can handle outliers.
For example DBSCAN clustering is robust against outliers when we choose minimum number of points (minPts) - (a threshold) large enough.
Ordering points to identify the clustering structure (OPTICS) is an algorithm for finding density-based clusters in spatial data which is also robust against outliers.
https://en.wikipedia.org/wiki/OPTICS_algorithm#cite_note-1 But we can't use general k-means: the squared error approach is sensitive to outliers. But there are variants such as k-medians for handling outliers.
(https://www.researchgate.net/publication/220490566_A_review_of_robust_clustering_methods)

Another approach is to apply a transformation to the data that can reduce the impact of outliers. For example, we could apply a log transformation or a power transformation to the data. These transformations can help to reduce the influence of extreme values and make the data more symmetric.

0.5 Conclusion

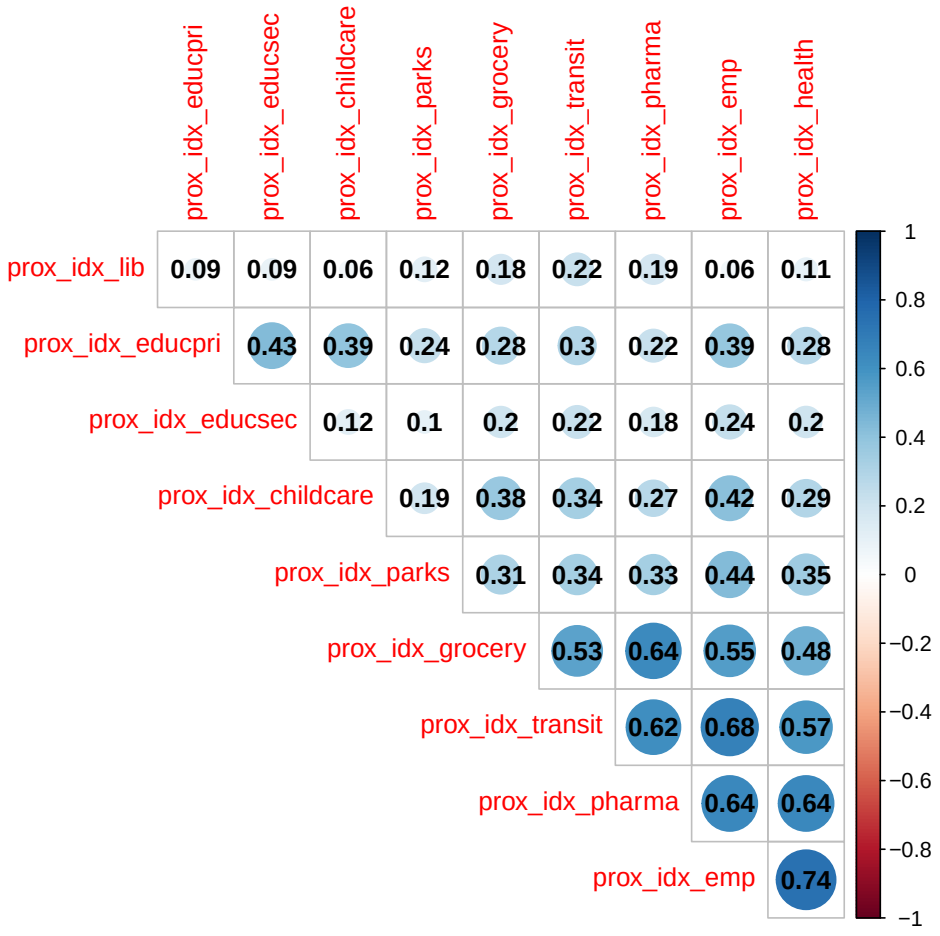
There are no obvious clusters in the proximity measures to the naked eye.

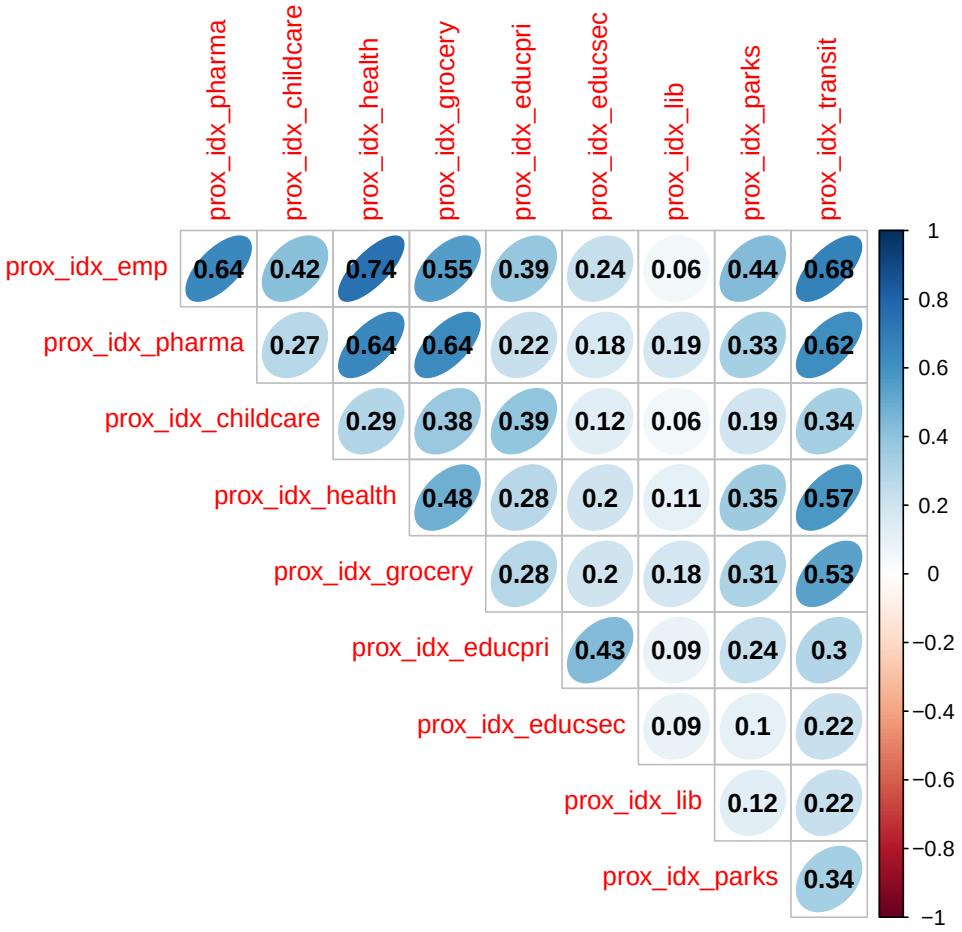
The distribution of missing values is not the same across amenities nor provinces.

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.

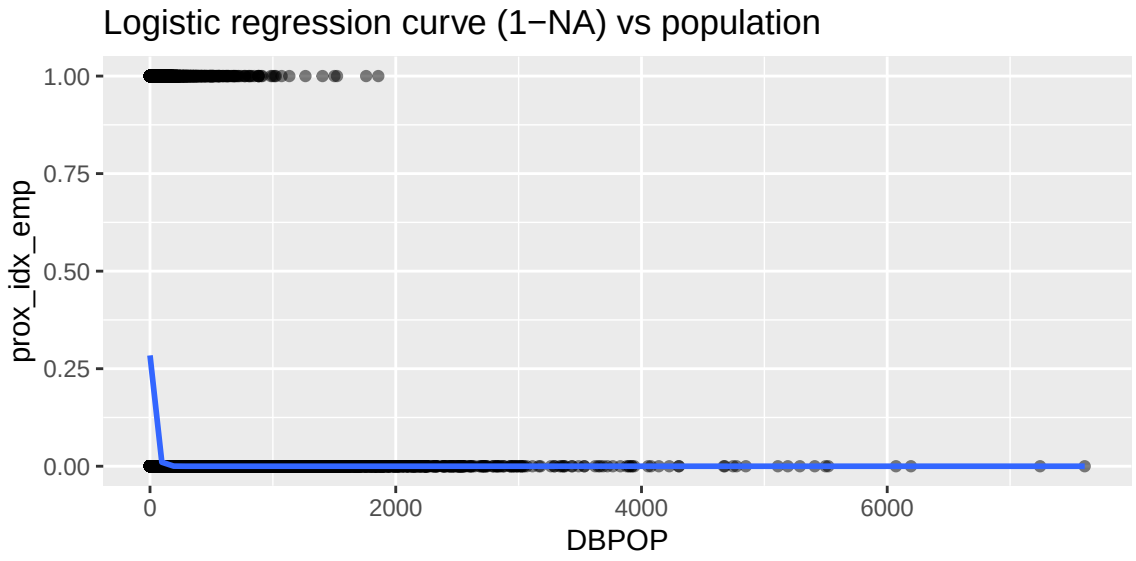
It doesn't appear that population of a DB is the only factor affecting whether a proximity measure is missing (NA); it was the only one tested as it was the only one included in this dataset. According to StatsCan's definition of a DB however, "only population and dwelling counts are disseminated at 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?).

0.5.1 Correlation

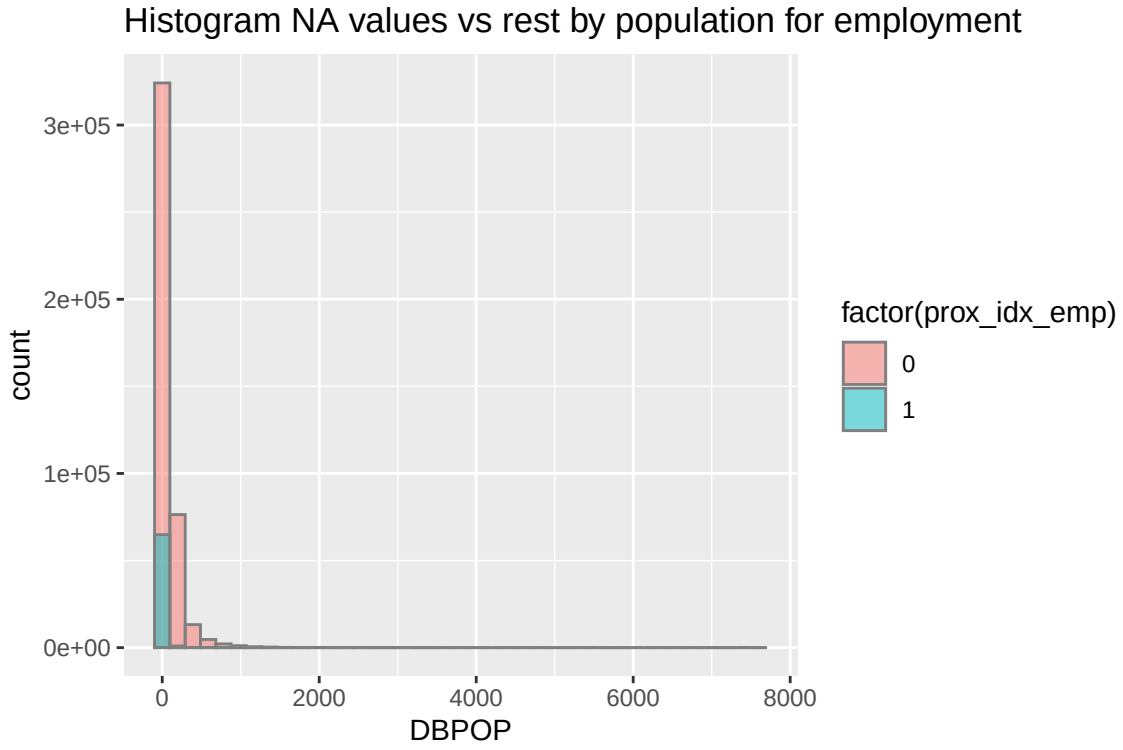



prox_idx_pharma
prox_idx_childcare
prox_idx_health
prox_idx_grocery
prox_idx_educpri
prox_idx_educsec
prox_idx_lib
prox_idx_parks
prox_idx_transit
prox_idx_emp 0.64 0.42 0.74 0.55 0.39 0.24 0.06 0.44 0.68
prox_idx_pharma 0.27 0.64 0.64 0.22 0.18 0.19 0.33 0.62
prox_idx_childcare 0.29 0.38 0.39 0.12 0.06 0.19 0.34
prox_idx_health 0.48 0.28 0.2 0.11 0.35 0.57
prox_idx_grocery 0.28 0.2 0.18 0.31 0.53
prox_idx_educpri 0.43 0.09 0.24 0.3
prox_idx_educsec 0.09 0.1 0.22
prox_idx_lib 0.12 0.22
prox_idx_parks 0.34
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

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.



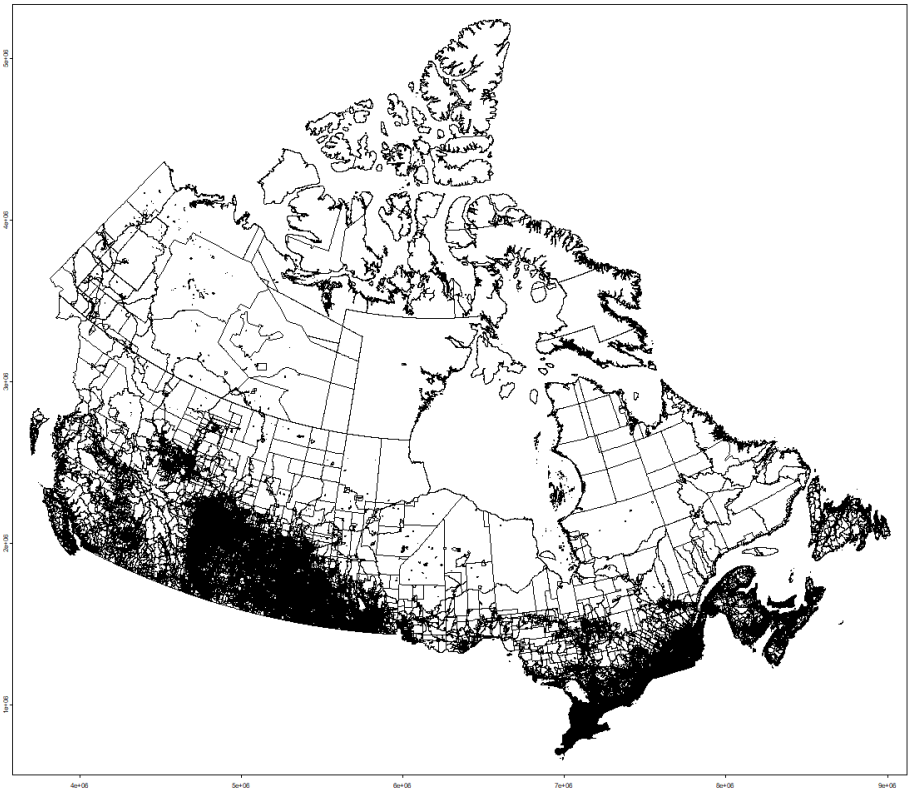
We can also plot the histograms of missing values vs populations for each amenity, where '1' (blue) is a missing value and '0' (pink) is a value not missing. Again, the remainder are found in the appendix.



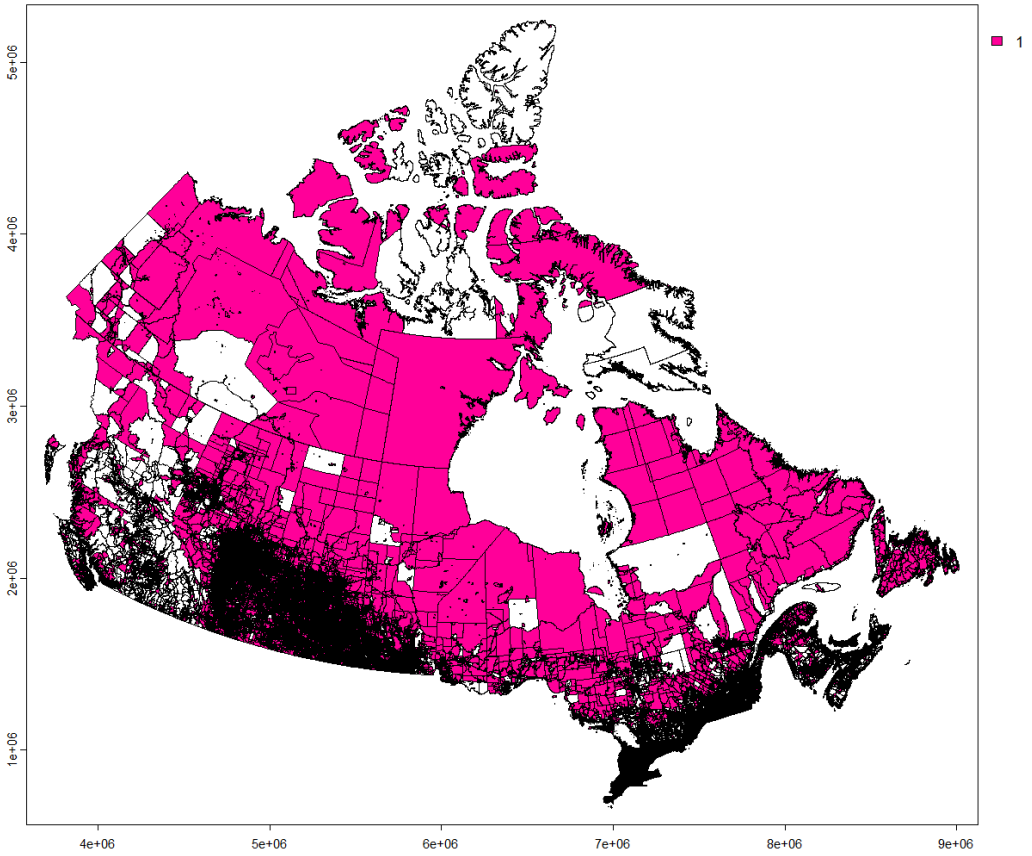
Overall, there are more DBs with lower populations than larger populations. We see that for some amenities, at smaller populations, there are a lot more missing values. Again, employment and health are the only two where there are always more actual values than missing values at every population bin.

0.5.2 Spat

DB plot on Map of Canada

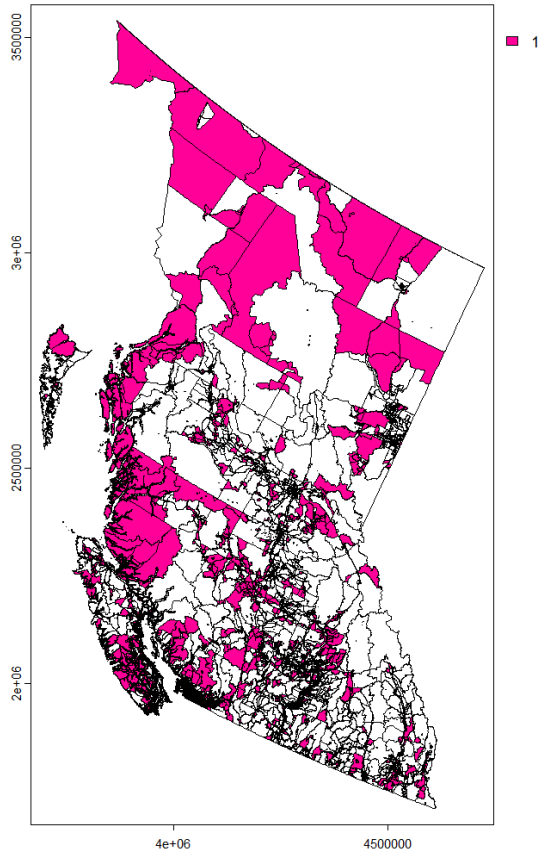


Marked DBs where all proximity measures are null.

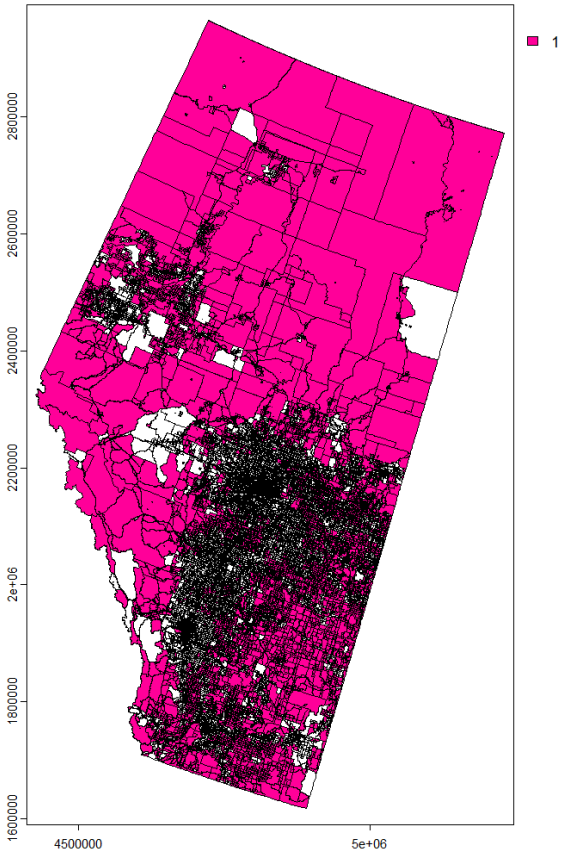


So where all proximity measures are missing are not randomly distributed. We have seen before that these areas have population. But our main target is to cluster the proximity measures. As all the proximity measures are missing here we can delete these rows from our database then do the clustering and also we can do clustering by keeping them aswell and see the difference.

Closer look to all null proximity measures of BC.

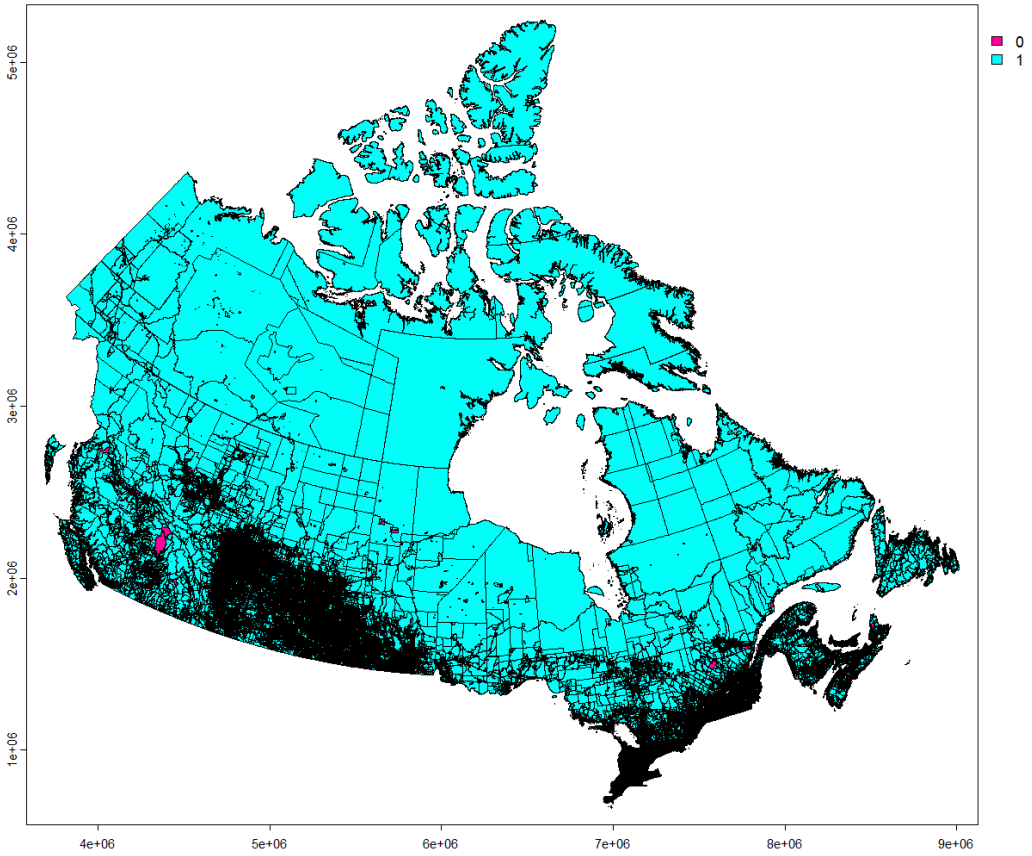


Closer look to all null proximity measures of Alberta.

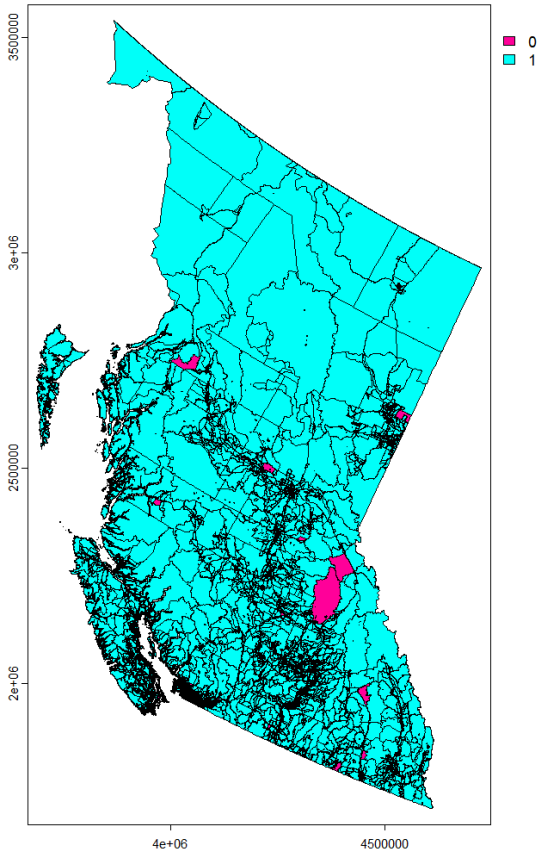


Now let's look at the proximity measures na values aminity wise.

Let's look where grocery proximity measures are missing all over Canada.

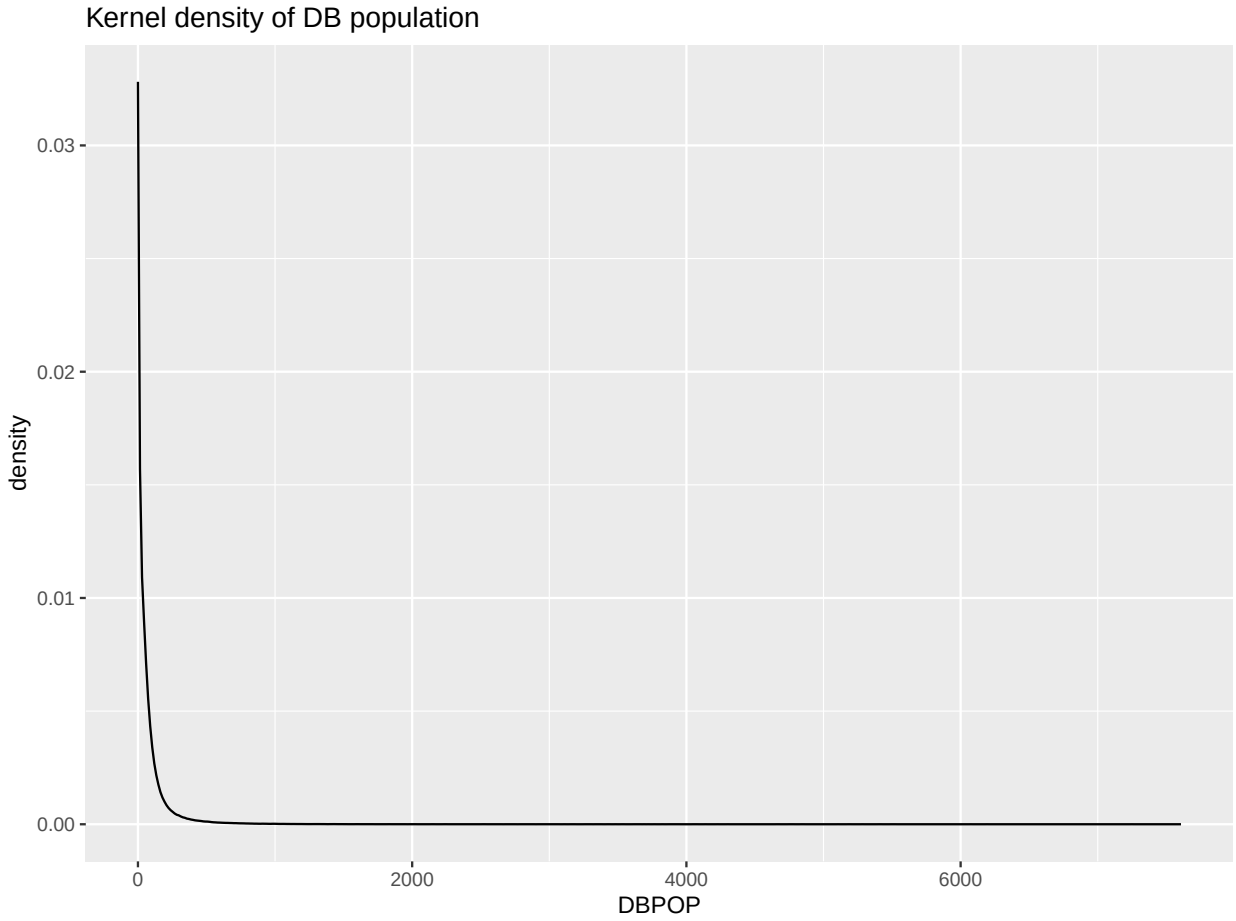


Closer look to null grocery proximity measures of BC

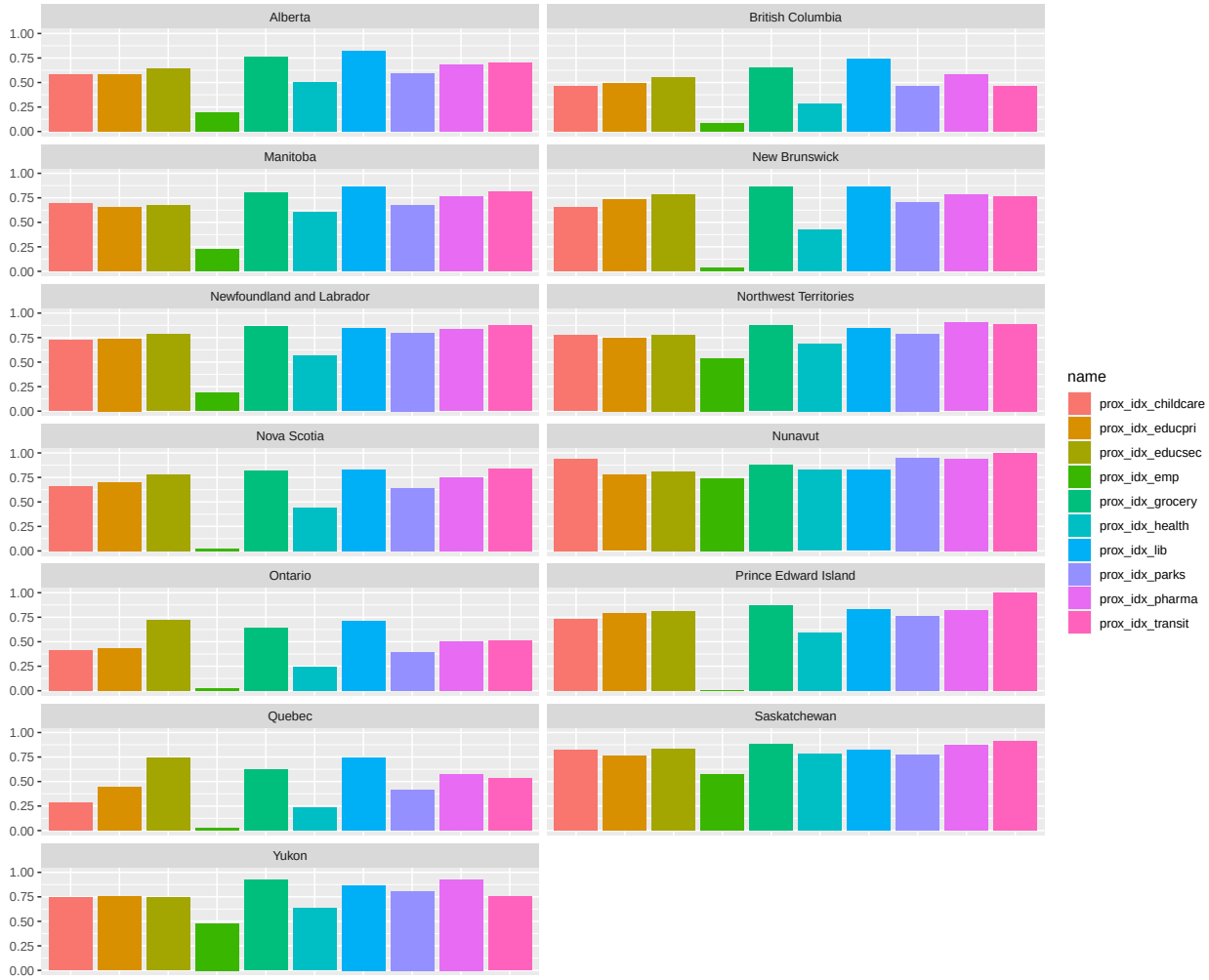


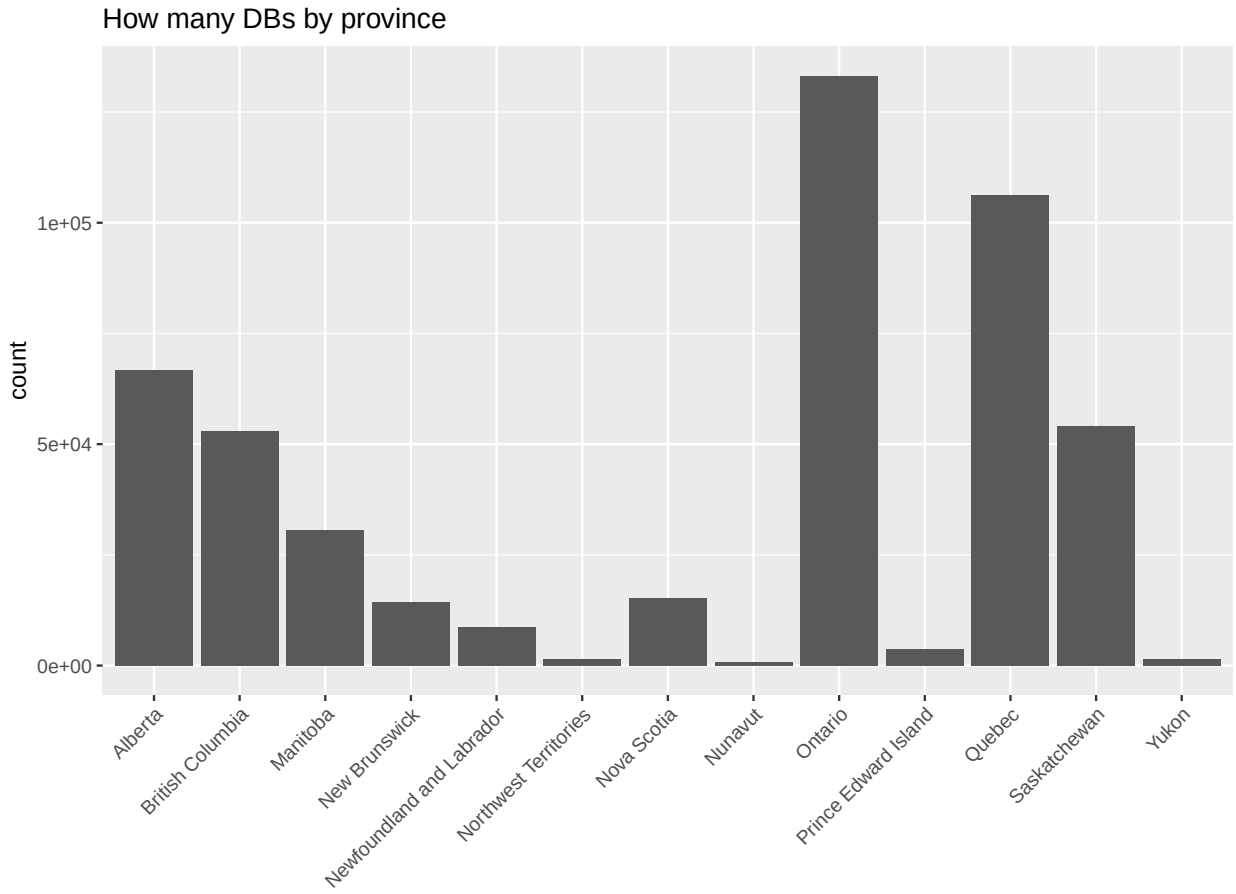

```
## No. of DBs in BC which prox_idx_grocery is NA 34526

## No. of DBs in BC which prox_idx_grocery is not NA 18324
```

0.6 Appendix

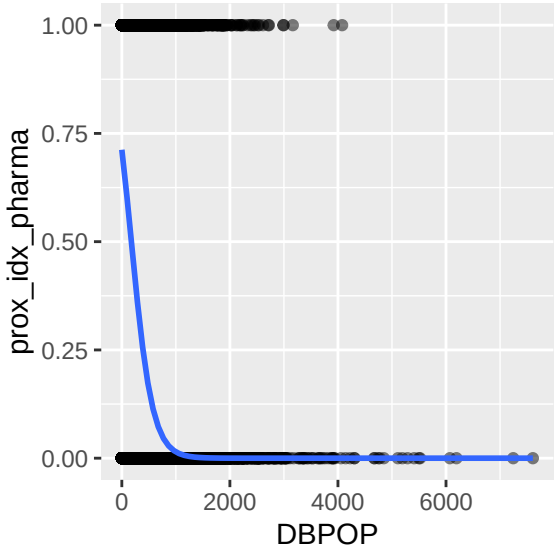

Kernel density of DB population
density
0.03
0.02
0.01
0.00
0
2000
4000
6000
DBPOP

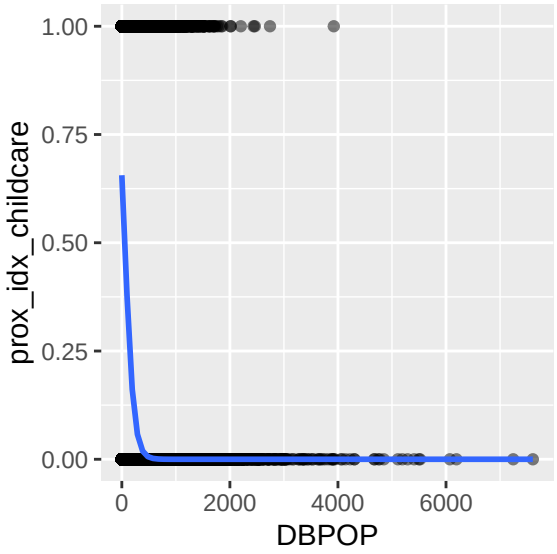
Missing values for each amenity by province

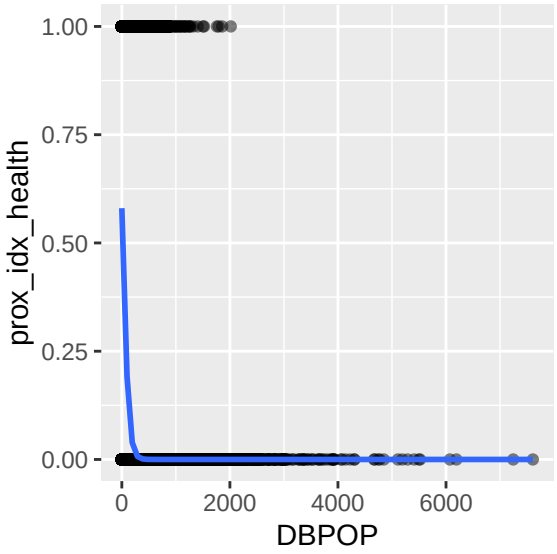


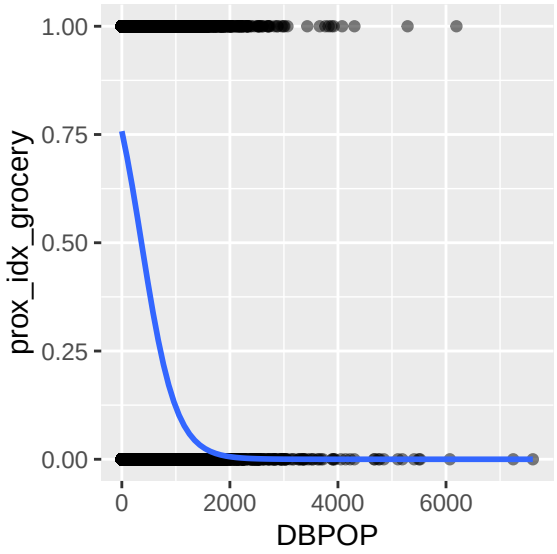


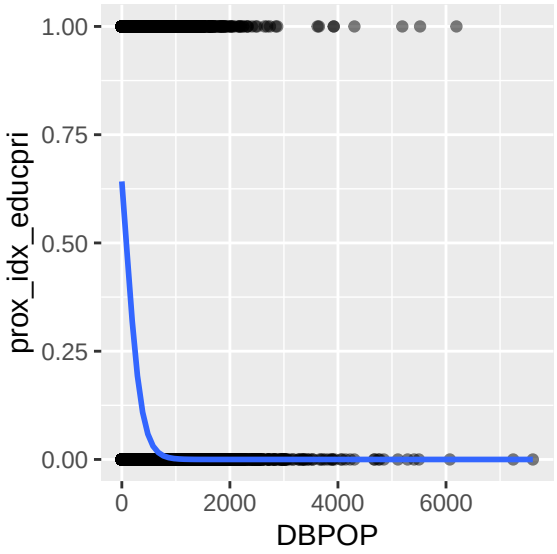
0.6.1 Logistics curves NAs

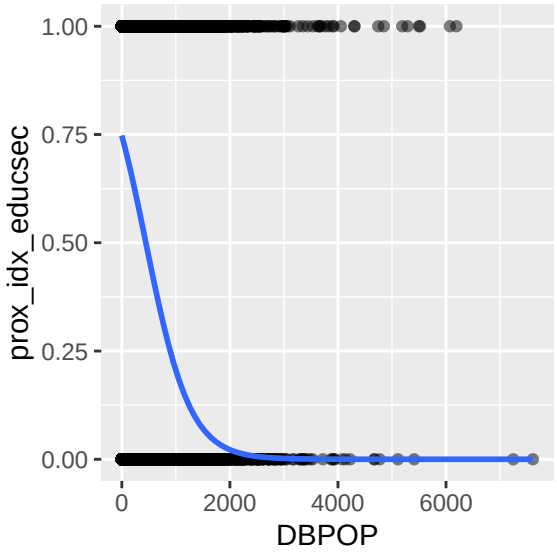


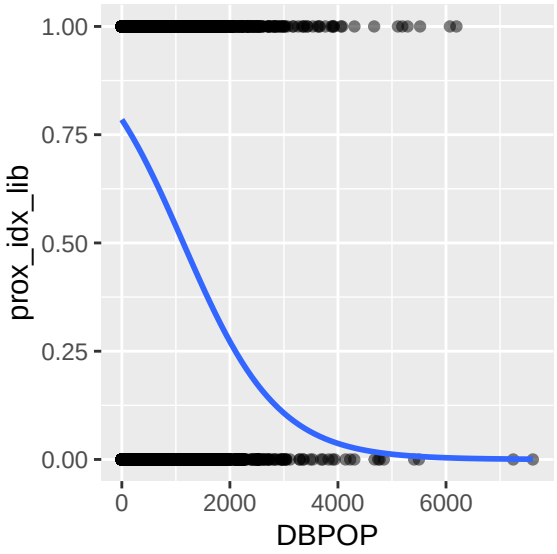

1.00
0.75
0.50
0.25
0.00
prox_idx_childcare
0
2000
4000
6000
DBPOP

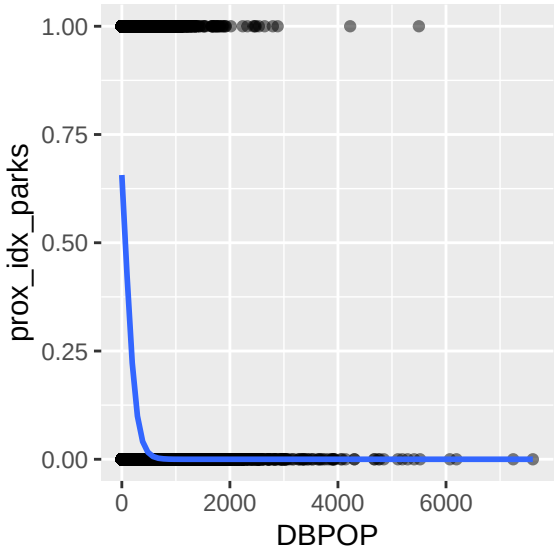

1.00
0.75
0.50
0.25
0.00
prox_idx_health
0
2000
4000
6000
DBPOP

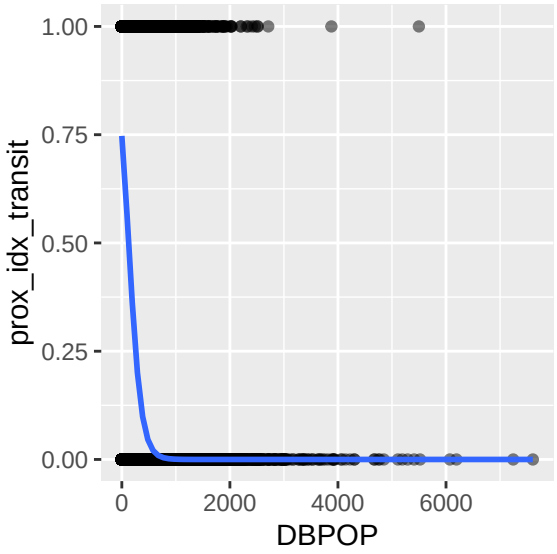

1.00
0.75
0.50
0.25
0.00
prox_idx_grocery
0
2000
4000
6000
DBPOP


1.00
0.75
0.50
0.25
0.00
prox_idx_educpri
0
2000
4000
6000
DBPOP

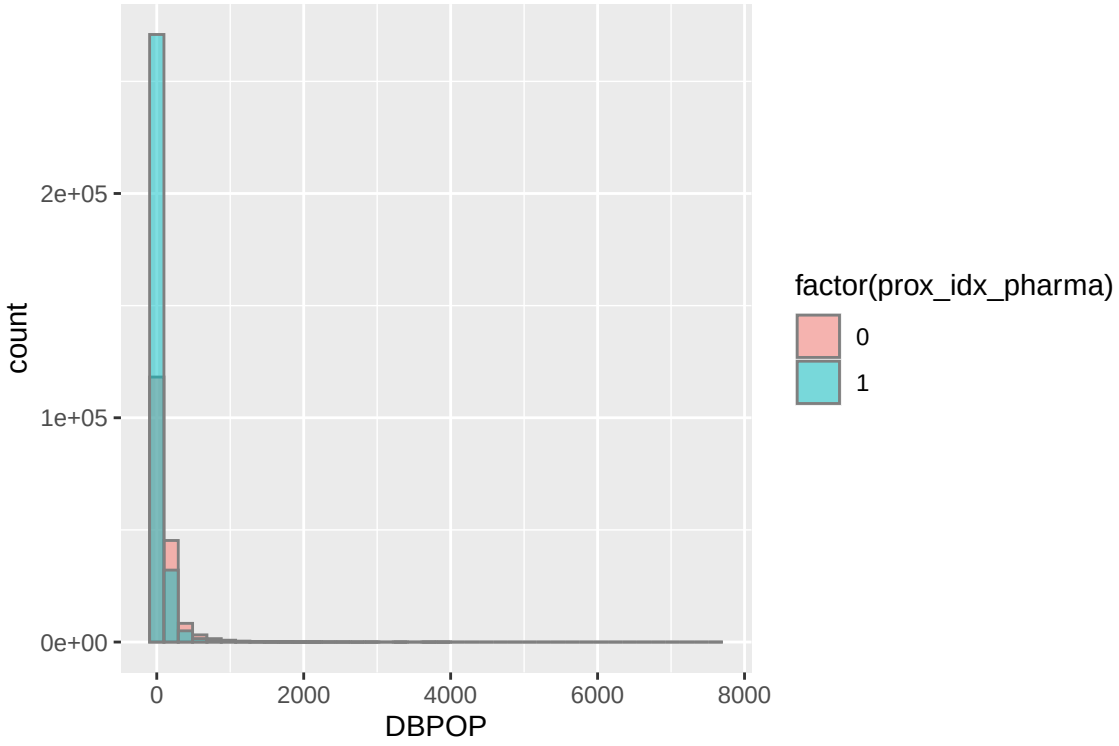

1.00
0.75
0.50
0.25
0.00
prox_idx_educsec
0
2000
4000
6000
DBPOP

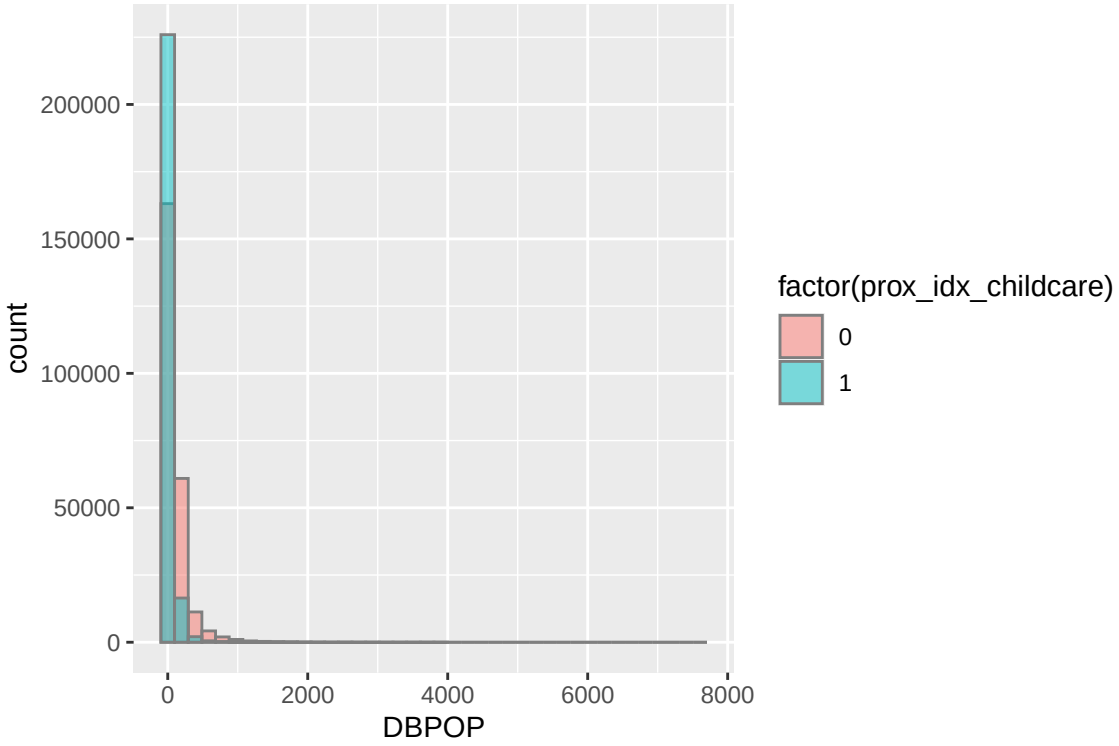

1.00
0.75
0.50
0.25
0.00
prox_idx_lib
0
2000
4000
6000
DBPOP

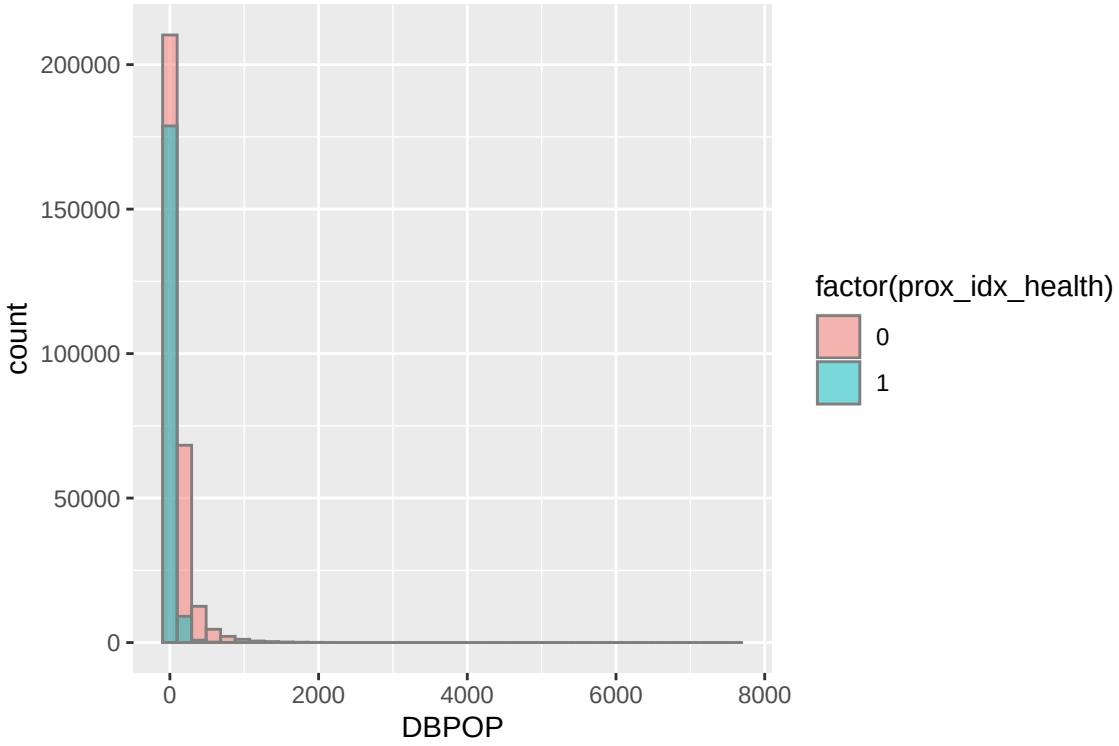

1.00
0.75
0.50
0.25
0.00
prox_idx_parks
0
2000
4000
6000
DBPOP

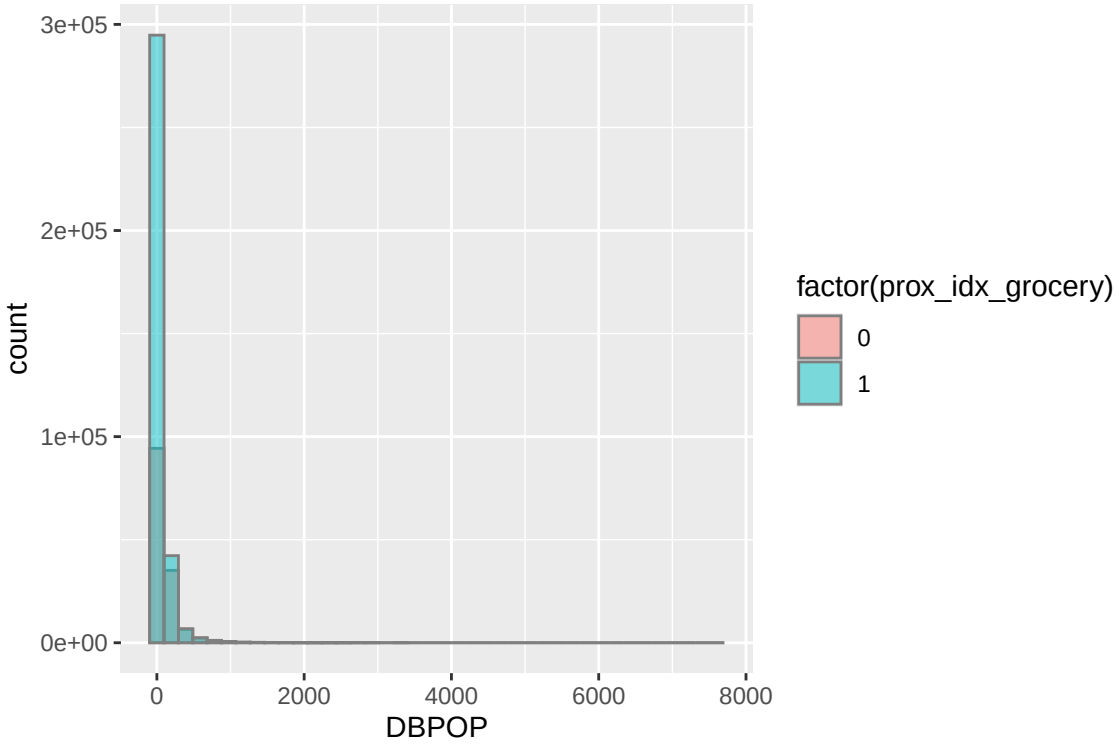

1.00
0.75
0.50
0.25
0.00
prox_idx_transit
0
2000
4000
6000
DBPOP

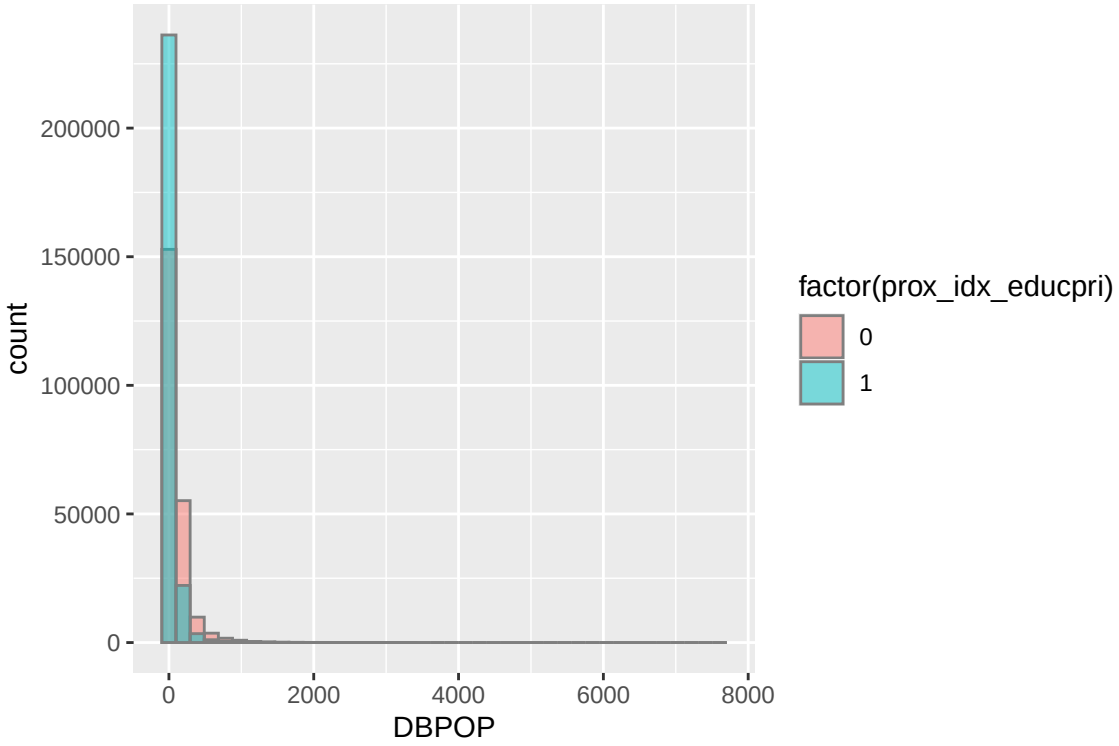
0.6.2 Histograms NAs

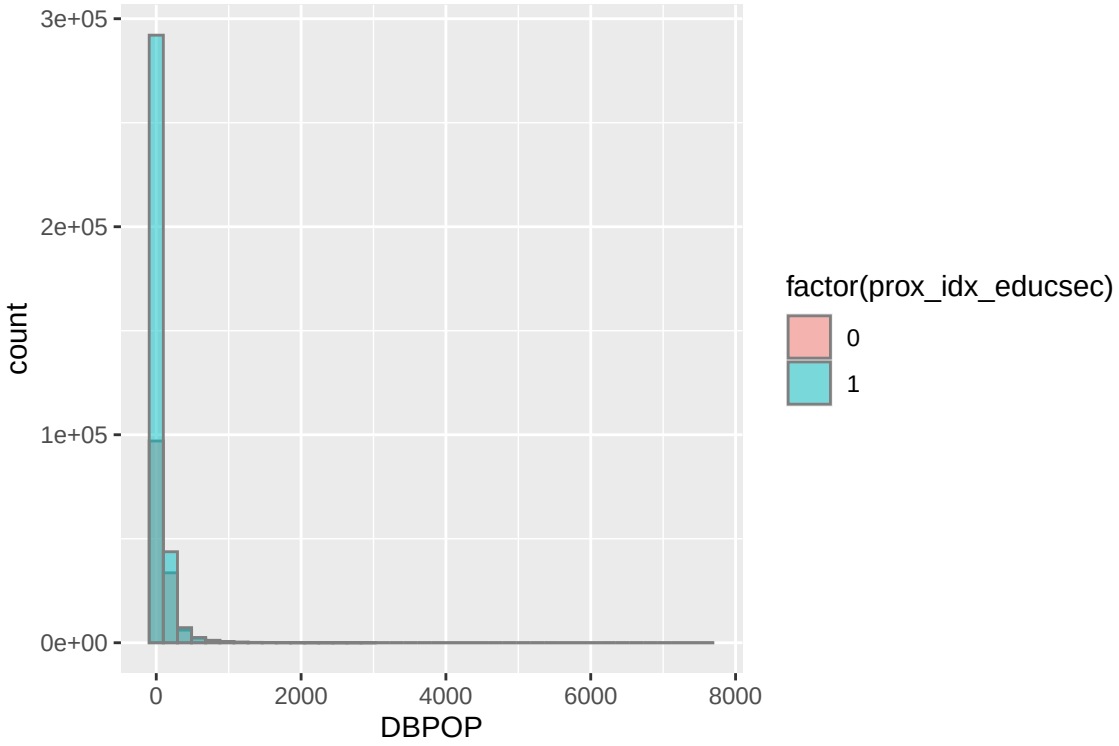


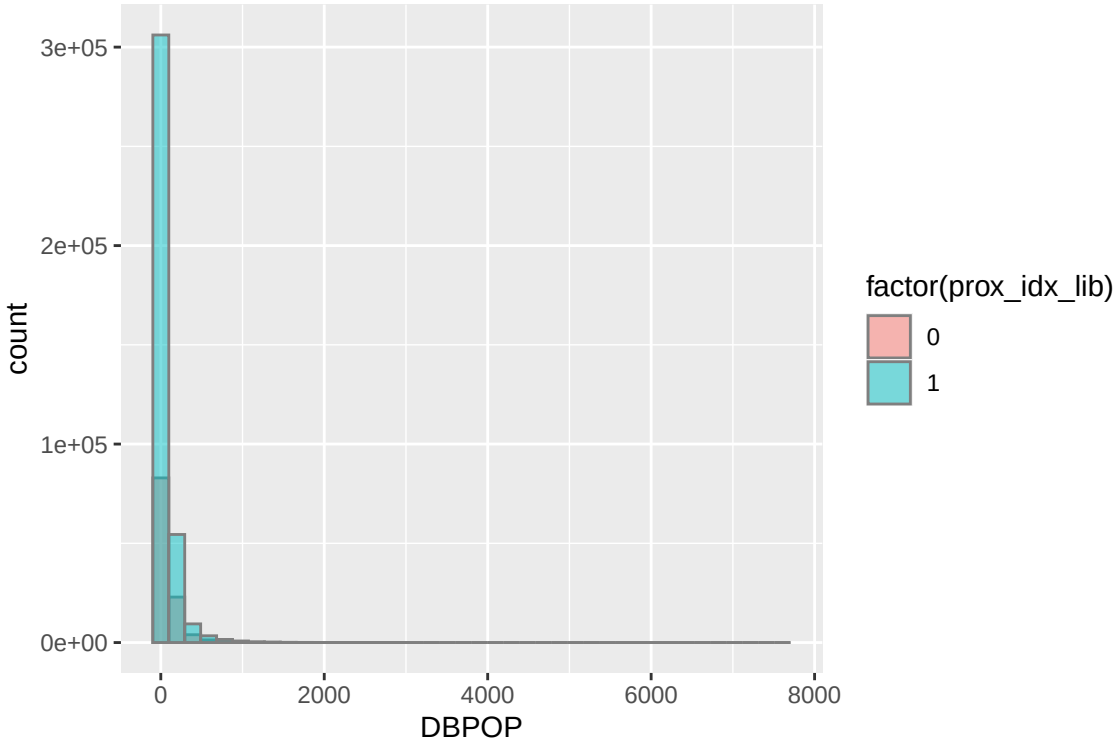


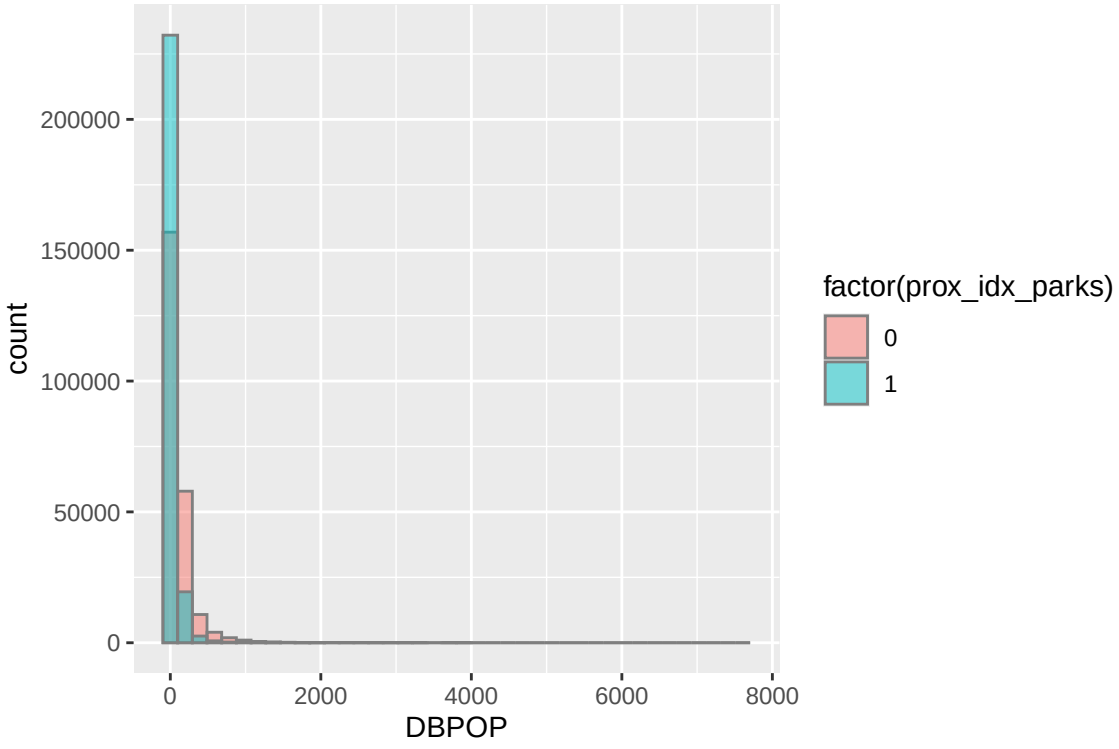

200000
150000
100000
50000
0
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_health)
0
1

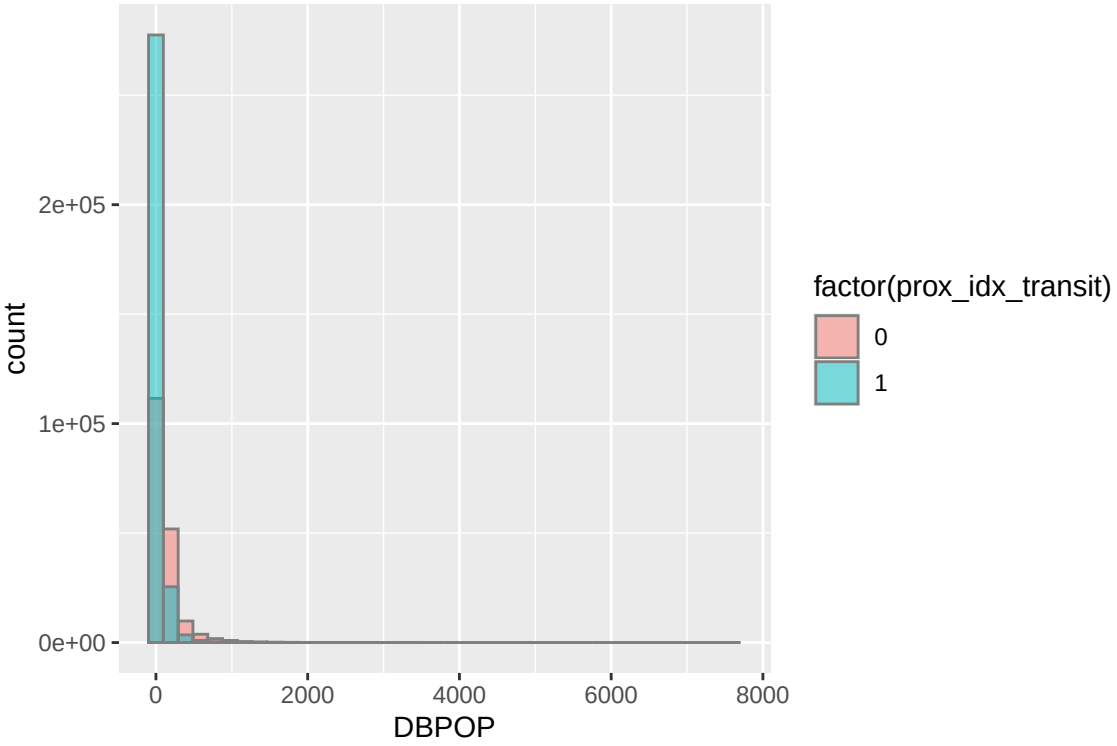

3e+05
2e+05
1e+05
0e+00
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_grocery)
0
1


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

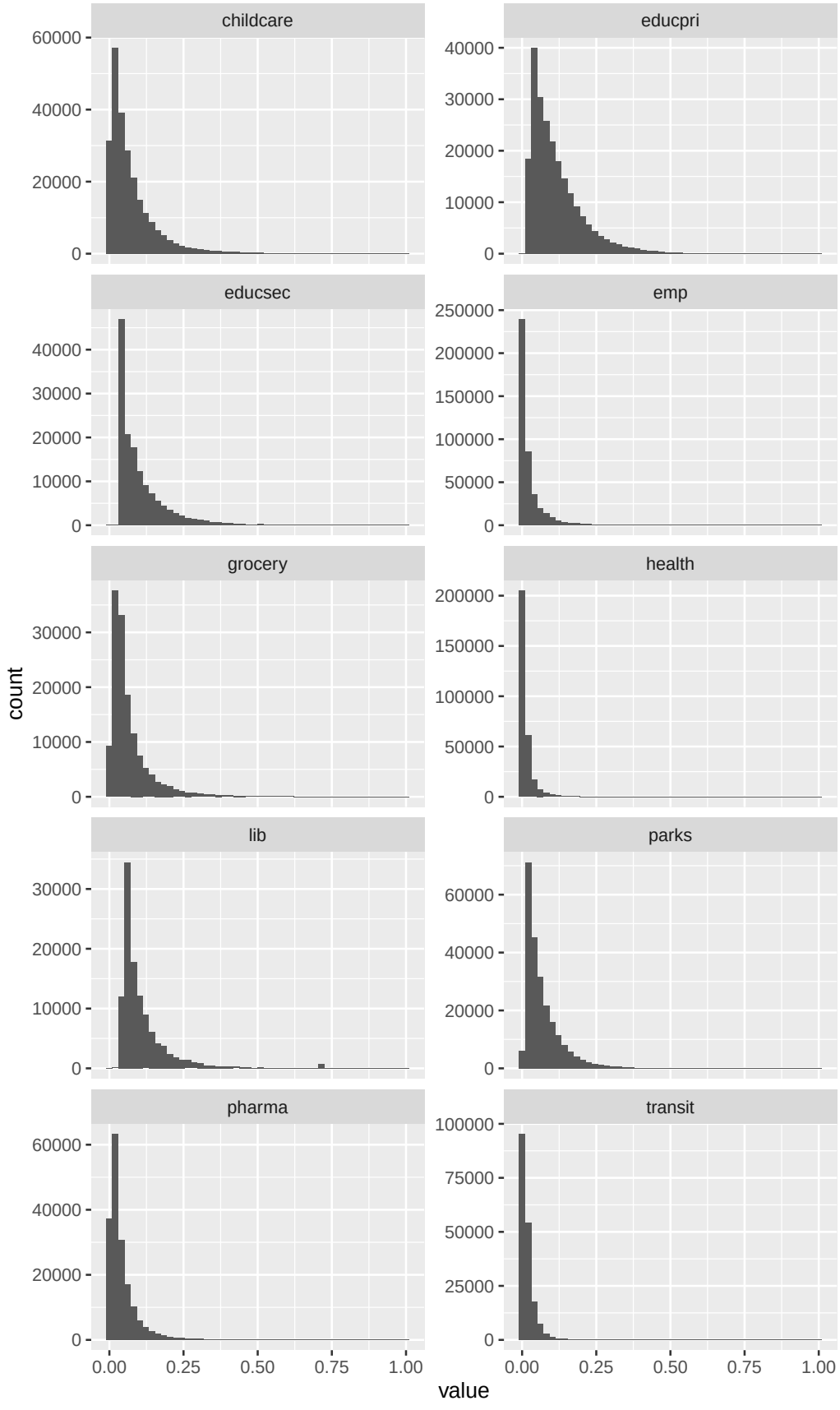

count
3e+05
2e+05
1e+05
0e+00
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_educsec)
0
1


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

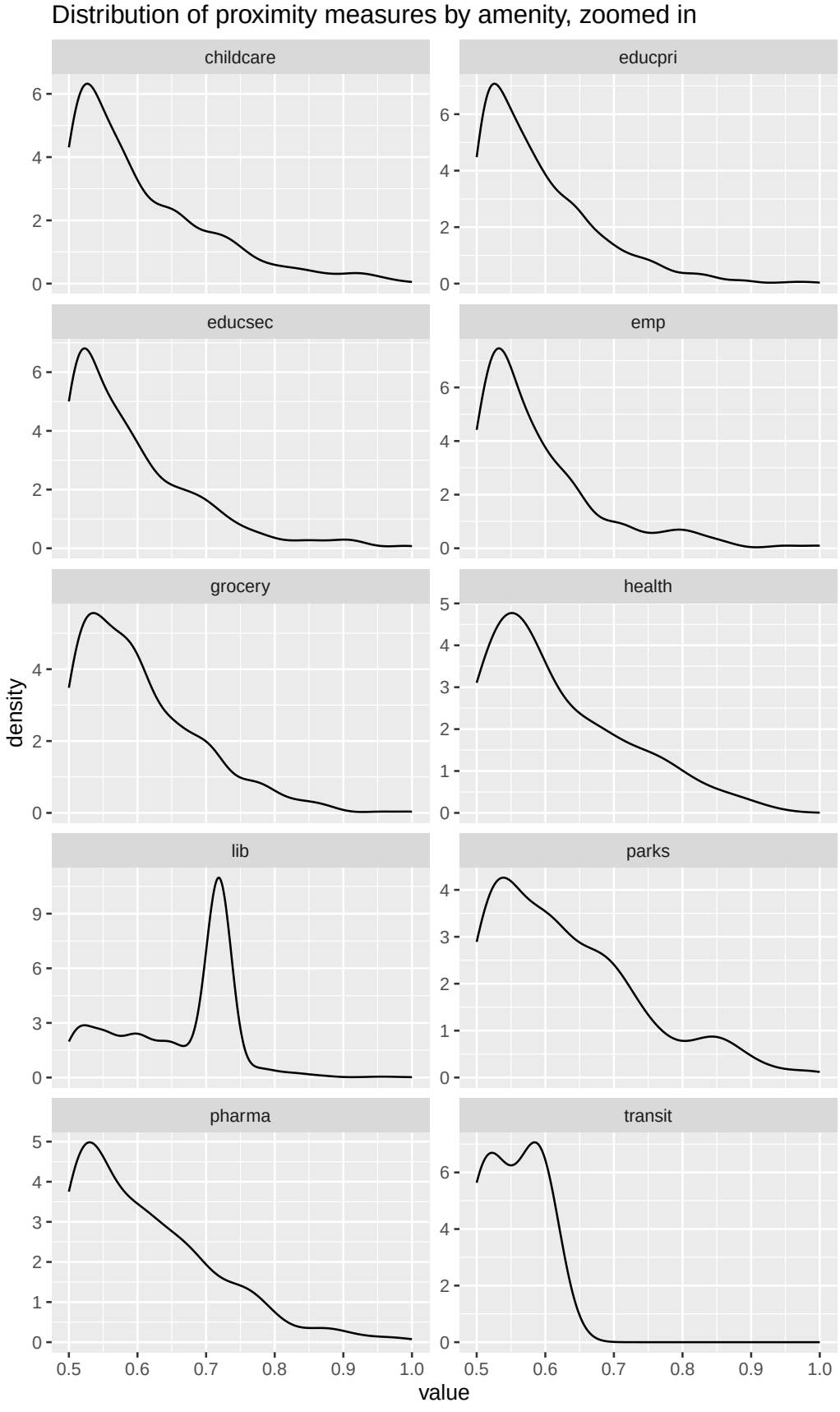

200000
150000
100000
50000
0
count
0
2000
4000
6000
8000
DBPOP
factor(prox_idx_parks)
0
1


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

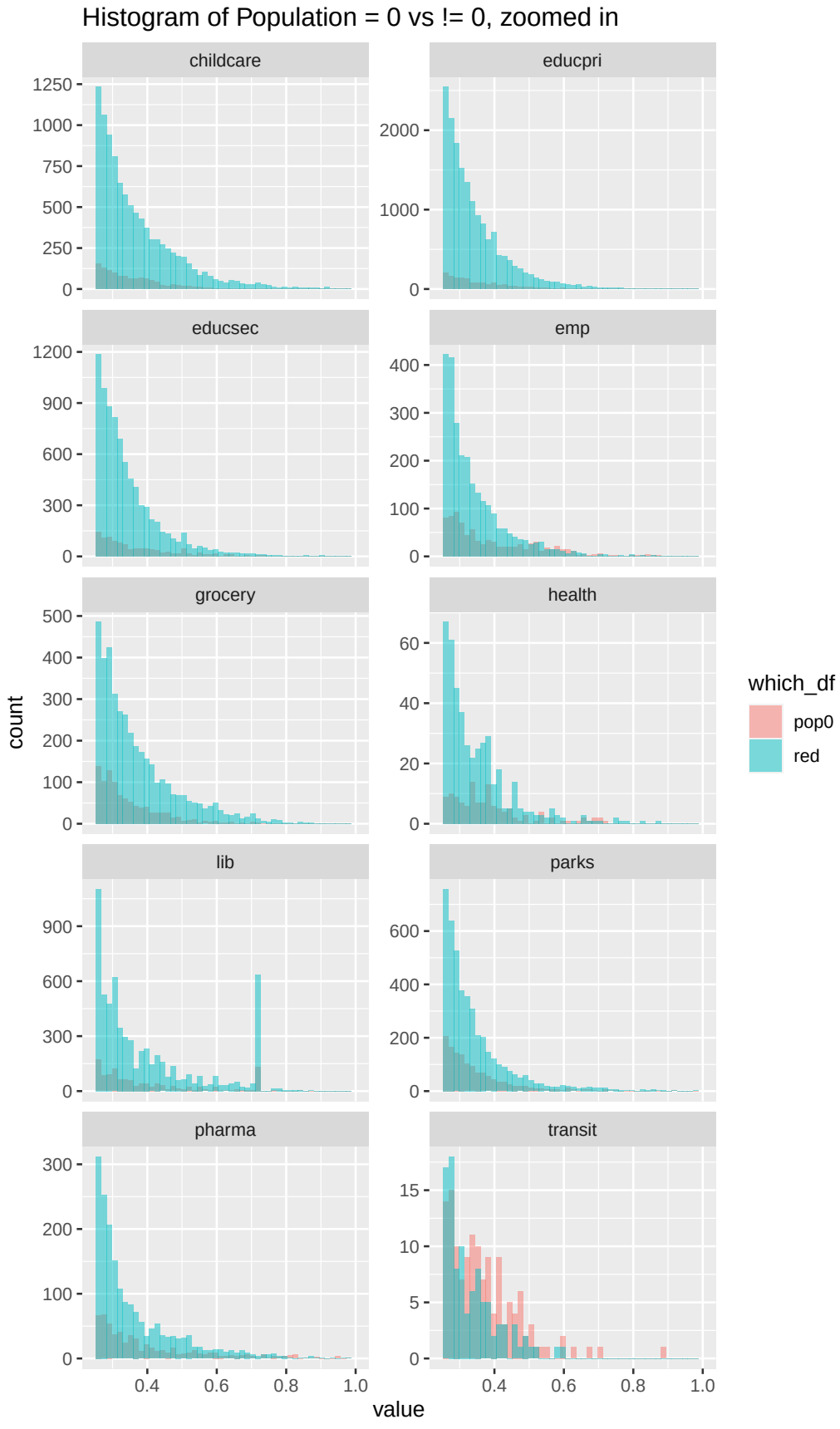
0.6.3 Proximity measures histograms

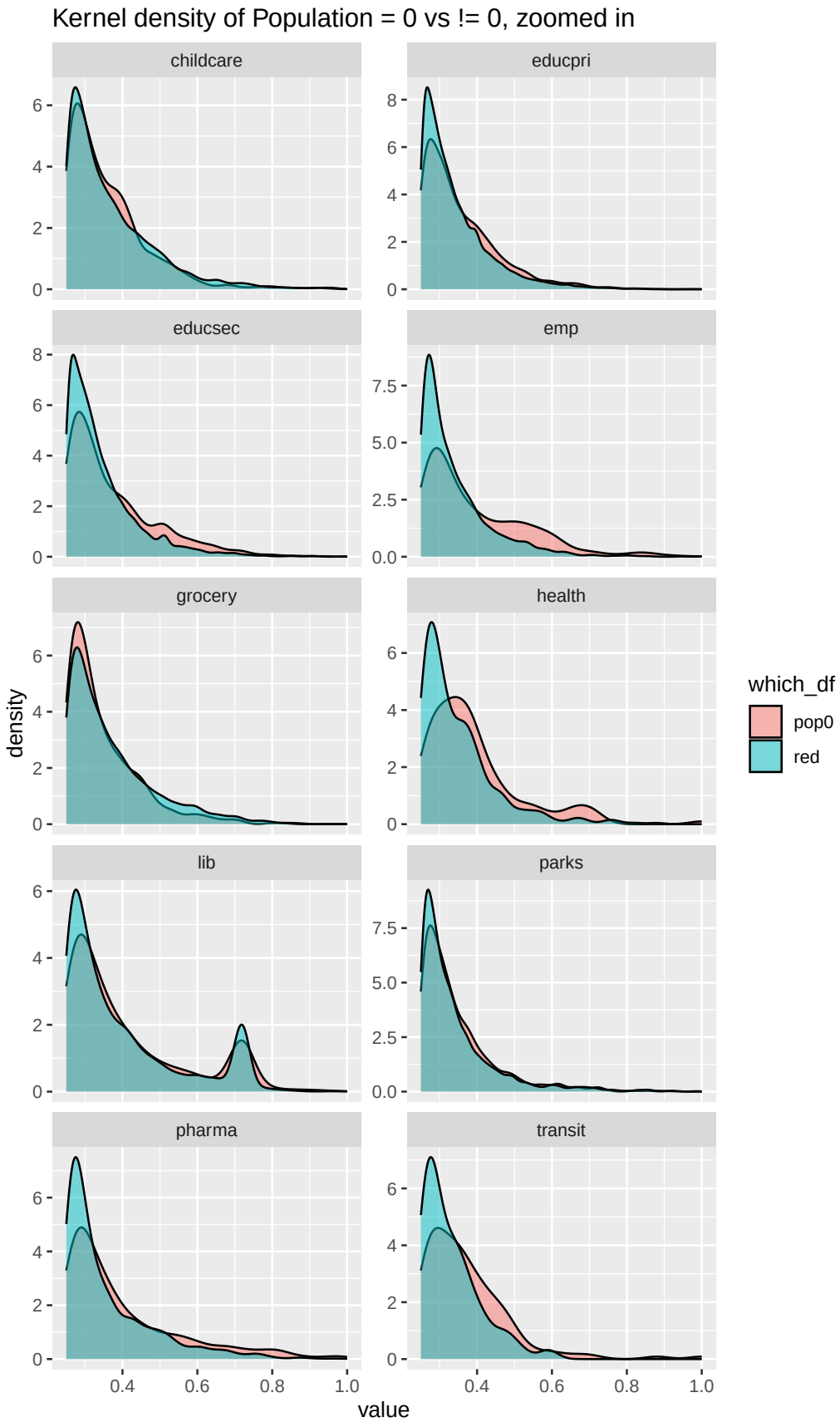


0.6.4 Proximity measures density zoomed in ? I think something is off



0.6.5 Population 0




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

0.6.6 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
```