Manual cutoffs

Ricky Heinrich

2023-05-30

Introduction

The Proximity Measures Database contains continuous measures for 10 amenities for a number of DB within a specific threshold. The distribution of these proximity measures is heavily right skewed, and there are for the most part no discernible clusters. The density distribution of each amenity is shown in Figure 1.

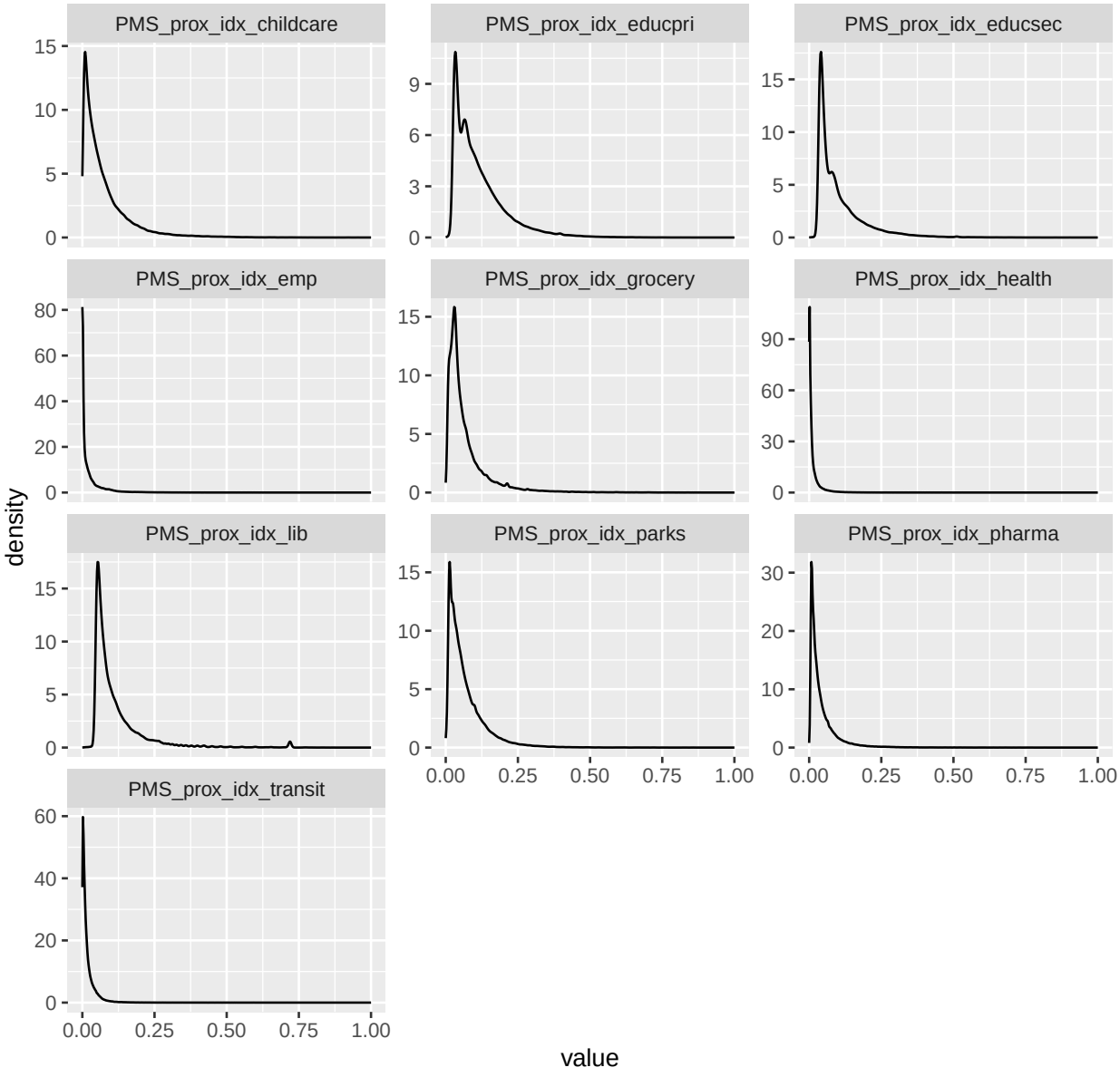

Figure 1: Distribution of proximity measures by amenity

When transforming the data, the inherent relationship between data points remain the same, but the new structure may reveal new insights. The most 'famous' transformation available is the log transform. It "can be used to make highly skewed distributions less skewed". It may help "make patterns more visible". A consideration to be aware of is that the log of 0 is -Inf. To account for proximity values of 0 in our dataset, we shift the distribution by +0.0001. This avoids the problem of -Inf whilst maintaining the original distances

between all values. The downsides of using a log transformation are [DOWNSIDES]. Figure 2 demonstrates the distribution of the log transformed proximity measures, where all the amenities' distributions were shifted by +0.0001. We can already visually identify more possible clusters.

In Figure 3, we only shifted the distribution by +0.0001 of the amenities that had a minimum value of 0. Grocery, educpri, educsec, and lib did not have values of 0 in their distribution and such were not shifted. The visual difference of the distributions between when +0.0001 is applied vs when it is not are imperceptible. For simplification in reproducibility, we will just apply the distribution shift to all amenities.

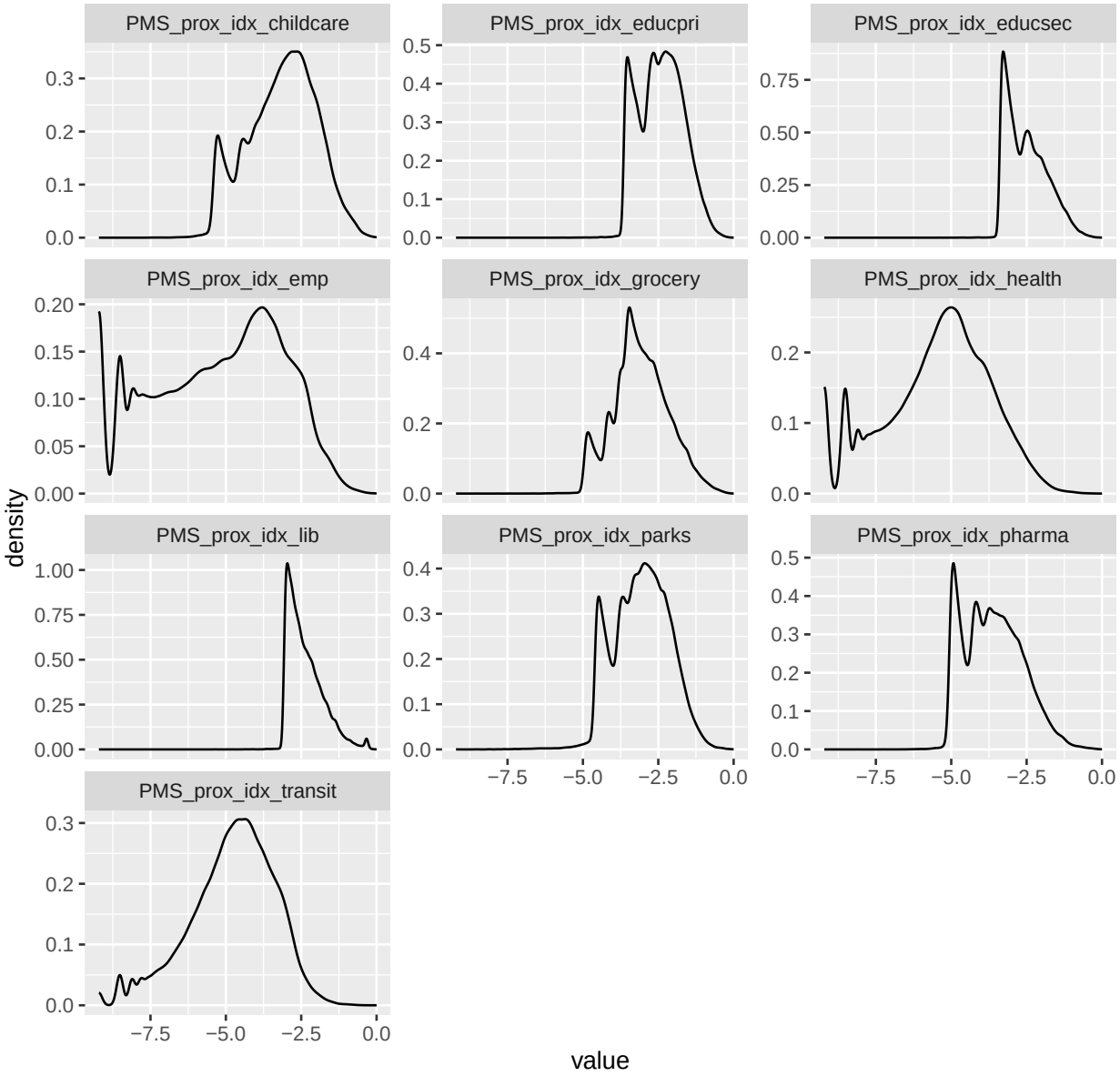

Figure 2: LOG TRANSFORMED(0.0001): Distribution of proximity measures by amenity

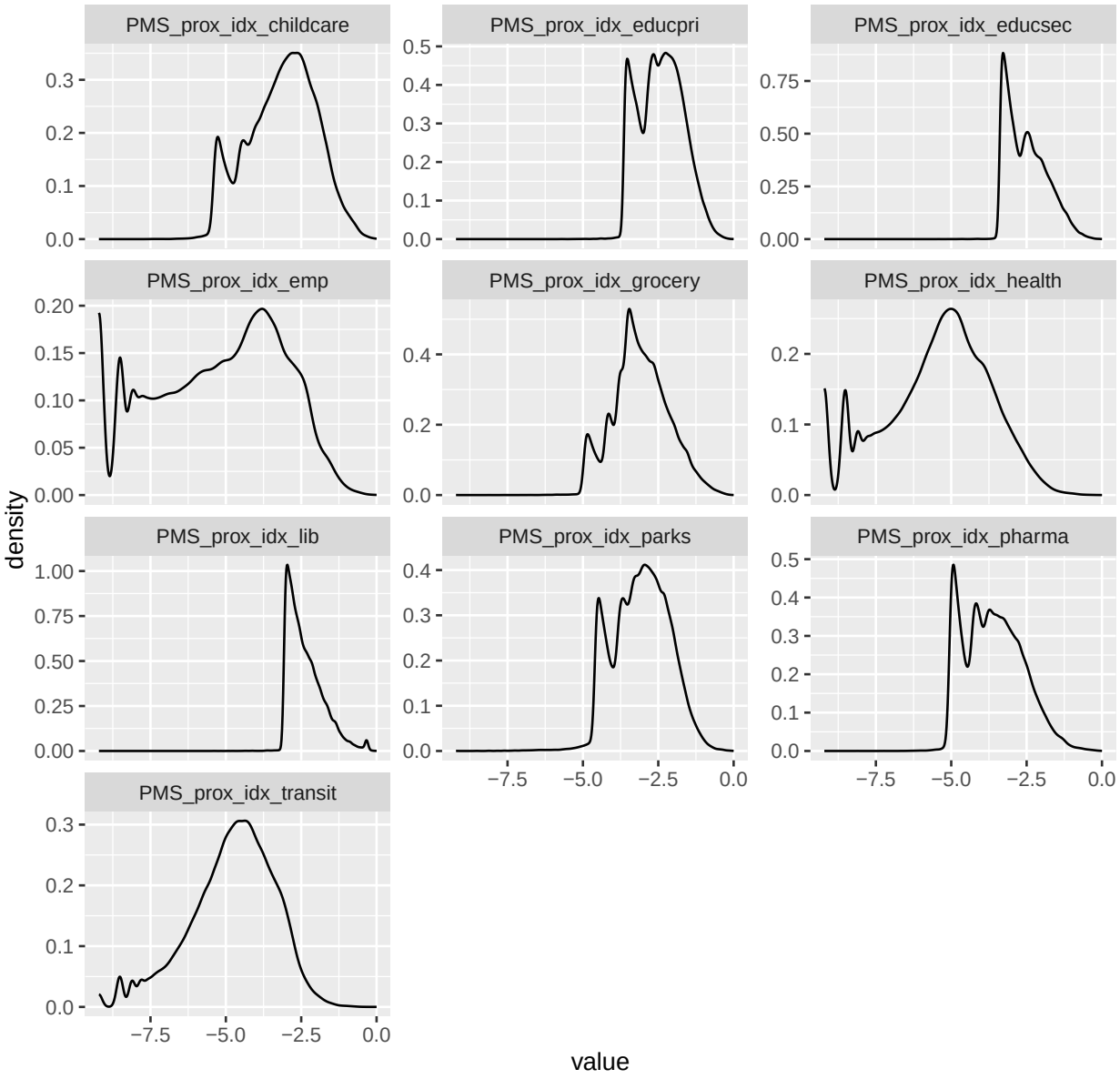

Figure 3: LOG TRANSFORMED(0.0001 in some): Distribution of proximity measures by amenity

Segmenting via minima

A segmentation technique is to segment the distribution at select minima of the density distribution. Each minimum in the density curves represents a density sparse region, which may be a 'natural' break in the continuous measures. Figure 4 provides an overview of where maxima and minima are located in the density curves of every amenity. We see that there are a lot of points that are by definition local minima, but are not fully indicative of density sparse regions. We can limit which minima are representative of density sparse regions by only including those who have a threshold difference between themselves and surrounding maxima. We will conduct an indepth analysis of which minima should intuitively represent a cutoff value for each amenity.

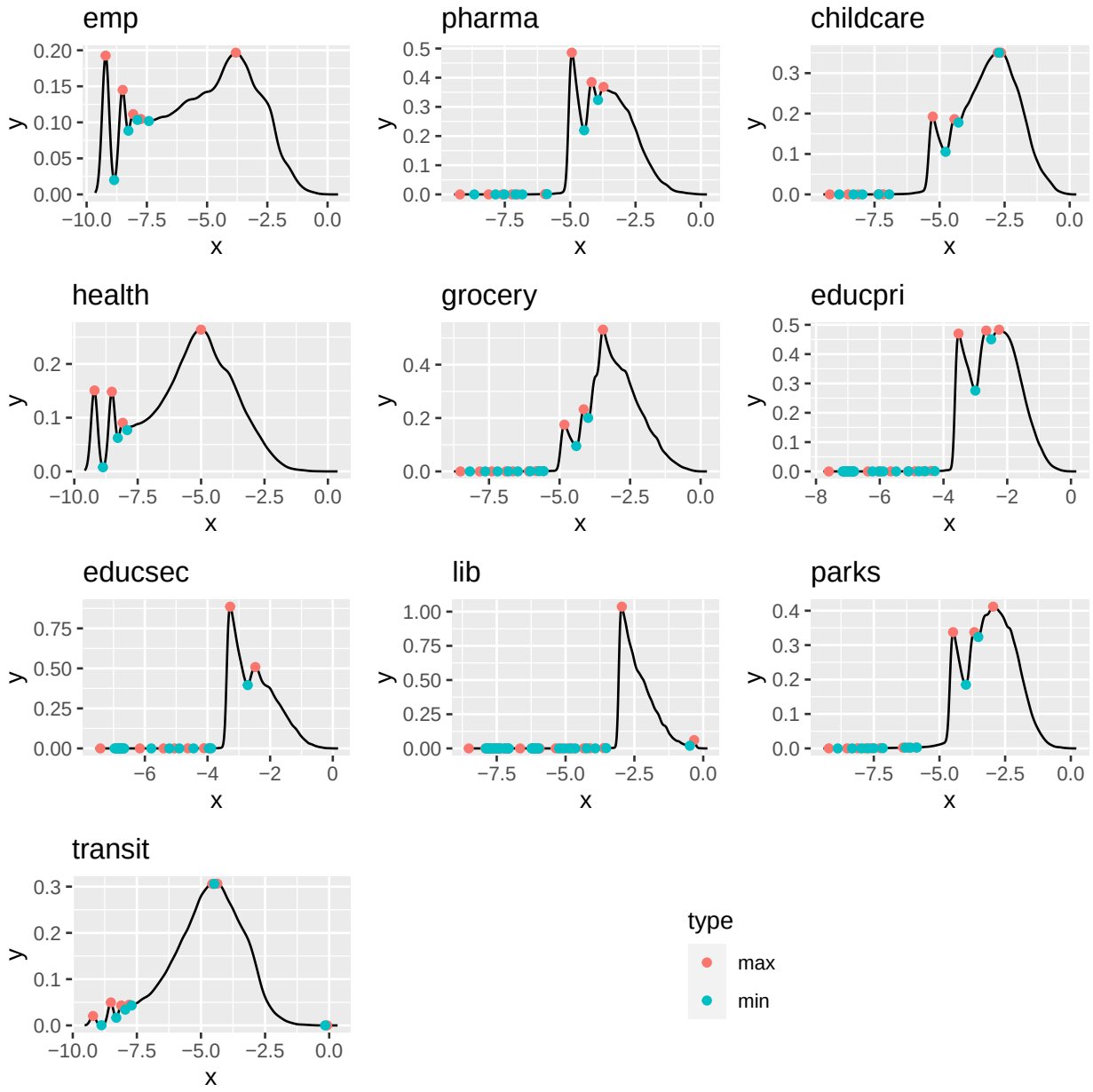

Figure 4: Location of minima and maxima

Employment

The Employment density distribution contains 4 minima. The following figure illustrates the density distribution with the minima ploted in blue and the maxima in red. Visually, we may not construe the third or the fourth minima as a cutoff value, as the peak in between is fairly small. As well, there are other areas in the curve that seem to plateau, and may be visually decent places for a cutoff value, but are not technically places where a minima is present.

As is, there would be 5 groups, corresponding to 4 cutoffs.

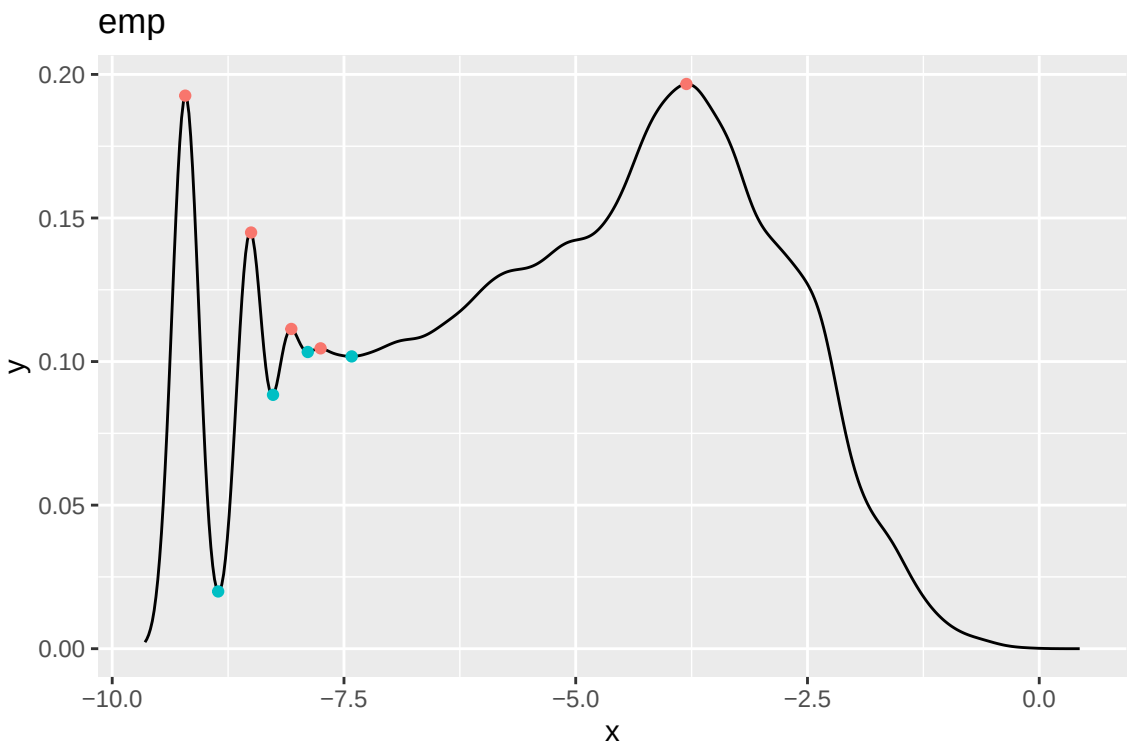

Figure 5: Employment density curve with minima and maxima

Pharmacies

In the case of Pharmacies, shown in Figure 6, there are many technical minima and maxima in an area that visually seems flat and have overall very low density. There is no doubt that these areas are not indicative specifically of density sparse regions, as the whole area is density sparse. The following plot, Figure 7, shows the difference between the density value of maxima-minima pairs (unidirectional). We see that for Pharmacies, the difference in the first 6 pairs is very small, as we can tell from the previous plot. The difference in density between the first maxima and the first minima, for example, is 2.8577488×10^{-5}, which is very small compared to the 8th (the 2nd visually discernible peak in Figure 6): 0.0612174. This may suggest that we should only use as cutoffs the minima that have a threshold difference with the neighbouring maxima. An appropriate threshold for Pharmacies may be a difference of 0.001.

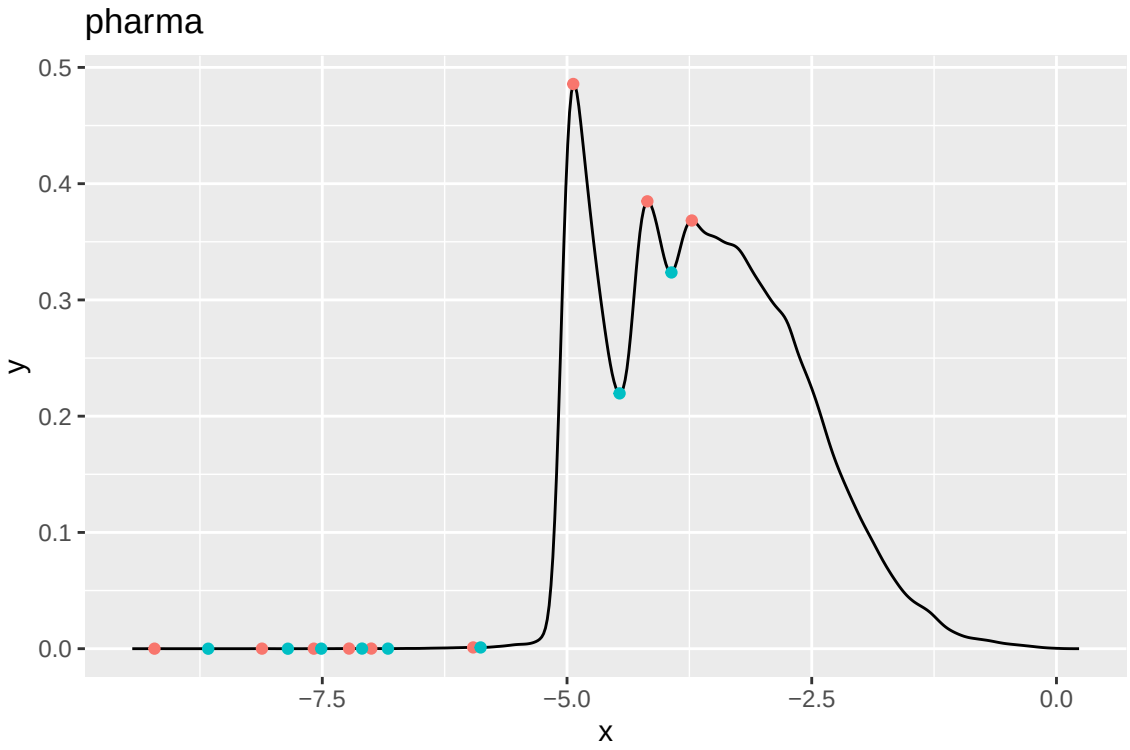

Figure 6: Pharmacies density curve with minima and maxima

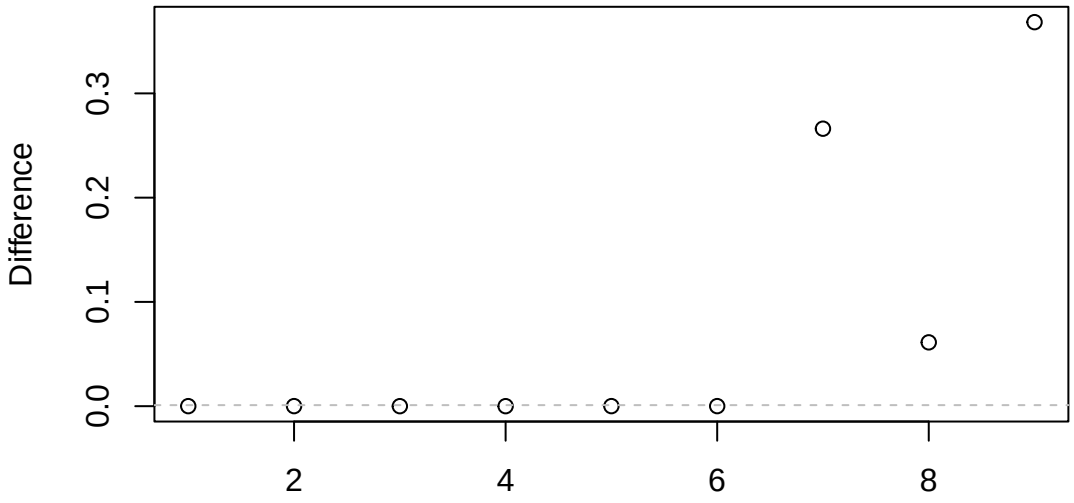

Figure 7: Difference between density value of a maxima-minima pairs, with suggested threshold = 0.001

Removing the pairs of where the difference is below the threshold values give us the following plot. We see that in this case, there would be 2 cutoff points giving 3 groups.

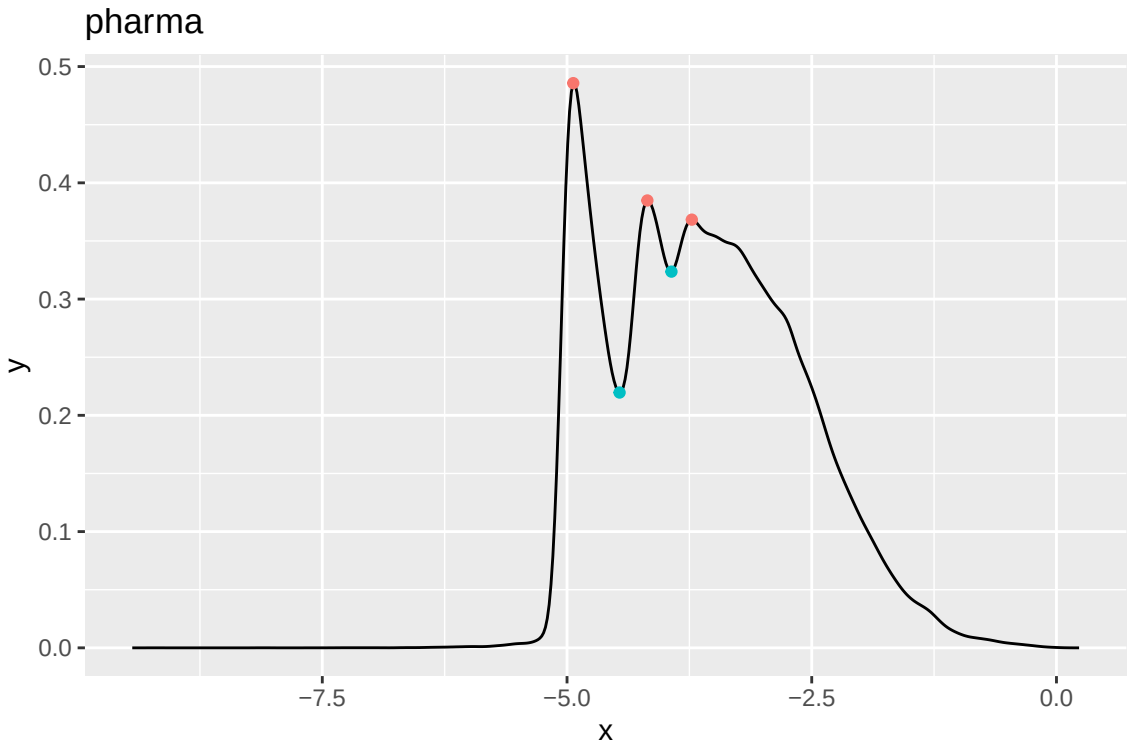

Figure 8: Density plot with suggested cutoff points in blue

Childcare

etc

Extra code