Identifying Food and Healthcare Deserts and their Impact

Brad Mills

August 1, 2020

1. Introduction

1.1 Background

The concept of food deserts is a growing problem that has been highlighted by the United States Department of Agriculture (USDA) and can be defined as geographic areas where a resident's access to affordable, healthy food options, especially fresh fruits and vegetables, is limited or nonexistent due to the absence of grocery stores within convenient travelling distance.[1] Similarly, medical or healthcare deserts are described as regions where there is inadequate access to one or more kinds of medical services.[2] Beyond the inconvenience factor of not having readily available access to basic human needs, the absence of healthy food and/or healthcare can have a broad negative impact on a neighborhood and those living in it.

1.2 Problem/Idea

Identifying food deserts and healthcare deserts by using live location data like that available through the Foursquare API should be achievable with a little creativity and elbow grease, aka data wrangling. This neighborhood data could further be linked to educational data from a different source to explore whether it is indicative of other issues, e.g. lower graduation rates, so that potential problems can be identified early and addressed through corrective interventions.

1.3 Interests

Local governments, grocers, healthcare institutions, and educators should all be interested in the timely identification of food and healthcare deserts within their communities. Local governments could use the information to get out ahead of urban planning and seek to recruit new businesses to their area. Grocery stores could use the data to look for new areas of possible expansion, as could healthcare providers on the medical side. And if a correlation can be seen between food or healthcare deserts and low graduation rates, school administrators and educators would be interested in early detection of problem areas and the clustering of similar neighborhoods within their school district to help identify students that might need additional assistance due to the other challenges that often accompany living in a food and healthcare desert.

2. Data

2.1 Data sources

The source of location data for this capstone project will be calls to the Foursquare API. Based on an initial review of the approximately 330 unique Venue categories observed in the NYC and Toronto lab examples, the following categories could prove useful in identifying food deserts:

- Discount Store*
- Deli / Bodega*
- Supermarket
- Organic Grocery
- Grocery Store
- Market
- Farmers Market
- Health Food Store
- Cheese Shop
- Fish Market
- Convenience Store*

The categories footnoted with '*' above will require a bit more investigation. In small neighborhoods these are often a source of food, but not necessarily the healthy kind, so I may in the end come up with some kind of formula or ratio that would actually penalize for their presence.

To identify healthcare deserts, again based on initial research, the presence or absence of the following categories could be used:

- Pharmacy
- Optical Shop
- Doctor's Office
- Drugstore
- Medical Center

And to link to a second data source to evaluate the potential correlation between food and healthcare deserts to school graduation rates, I will be using one or both of the following categories:

- High School
- School

In the assessment of graduation rates and their correlation to food and healthcare deserts, I will initially be concentrating on Indiana High Schools located primarily in the Indianapolis metropolitan area. Graduation rate data for 2019 is currently available in Excel spreadsheet format on the Indiana Department of Education's website.[3]

2.2 Data cleaning

While graduation data and statistics for Indiana and the Indianapolis area High Schools is available online, initial inspection indicates that, as usual, it will take a bit of data wrangling to get it into a usable format. Use of the pandas library and methods should help, but some potential issues to address include:

- The Excel spreadsheet contains multiple tabs, two of which relate specifically to graduation rates at the individual school-level.
- Of the two sheets related to school-level graduation rates, one waives the need for students to pass exit exams and potentially hides some of the challenges of coming from a disadvantaged area, so I'll need to decide whether to use one or both of the graduation rates.
- The sheet that distinguishes between NonWaiver students (students that passed an exit requirements exam) is not particularly well formatted for parsing, e.g. column headers appear at different levels and not in the first row, so I'll need to leverage pandas read options or other methods.
- Some schools are of such a small size that their calculations are not statistically significant, so those rows will need to be filtered out.
- I'll be focusing on Indianapolis area schools for the purpose of mapping and clustering, so will need to filter on a list of desired school corporations or within a specific distance from downtown Indianapolis.
- When measuring Venues within a certain radius of a central neighborhood point, which in this case will be based on school addresses, it may be necessary to factor in urban vs rural locations. In the city, at least in Indianapolis where public transportation exists but is less than ideal, reasonable travel may be by foot or bus and somewhat more restricted in distance than in the outer regions of the city where car transportation is the norm. I will likely need to infer community type and add an indicator column for this as I pre-process the data.

3. Methodology

3.1 Exploratory Analysis

As anticipated in the initial data write-up above, some of my exploratory analysis effort involved identifying how best to identify and separate urban from rural school locations. I found several good articles[4,5] that confirmed my suspicion that I'd want to distinguish travel/search distance based on this distinction due to different modes of available travel. Section 5 of the Jupyter notebook code includes a function called travDist() that implements this concept without relying on a check of school corporations that I had originally thought might be necessary. This general approach has the potential additional advantage of making the code flexible for use in other metropolitan areas without relying on specific knowledge around local school systems.

One initially unanticipated challenge in linking the two data sources, Foursquare's API with Indiana's Department of Education website, was that they frequently referred to the same school by slightly different names. Further detailed analysis revealed some common patterns though, and I was able to programmatically standardize the names through data cleaning routines in Section 2 of the code (grtStandardize, fsqStandardize) rather than curate manually.

A final area of exploratory analysis involved how best to leverage the Foursquare API, including some limitations to the number of results returned for the free version. Switching from search? to explore? mode and limiting the query to specific venue categories, anticipated in the Data write-up above and finalized in Section 5 of the Jupyter notebook, led to good results that required only a little bit of additional cleaning and filtering.

3.2 Inferential Statistics

To demonstrate a correlation between food and/or medical desert risk based solely on area business types and the graduation rate of neighboring schools, I leveraged primarily visualization methods for descriptive statistics studied in the Coursera courses, e.g. the Box plots generated in Sections 6 and 8 of the code and captured in the presentation slides as well.

Slide 12 in the presentation does include an analysis of variance (ANOVA) that shows a significant difference among the 3 Food desert categories based on graduation rate. I calculated this particular ANOVA by hand while I was tutoring my daughter for her college Stats class, and the next logical step might be a Partial Correlation analysis to factor in neighborhood population size as a potential confounder. I did extract graduation class size from the Department of Education spreadsheet as a potential proxy for population size, but these advanced analyses techniques were not covered in our coursework.

3.3 Machine Learning

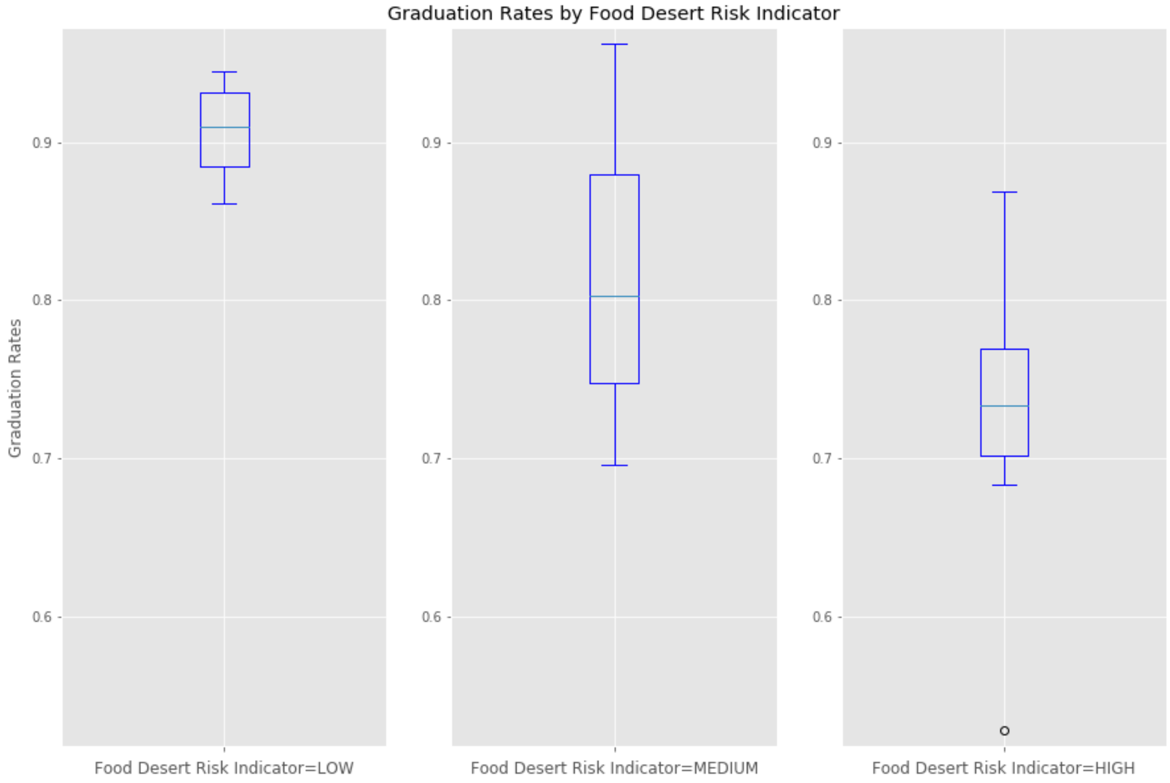
To augment and compare to the Food and Medical desert categories that were created in Section 5 of the code using statistical methods, I leveraged unsupervised K-means Clustering to look at two additional approaches to identifying Food deserts.

Section 7 of the Jupyter notebook details the basis for picking k and the resulting clusters when using a) Total grocery-related venues and b) leveraging further grocery subcategories. This machine learning approach was selected to explore and hopefully optimize the best way to group similar neighborhoods without explicitly looking at each of the possible permutations of features. The final section of the notebook, Section 8, assesses the correlation of the resulting clusters with school graduation rates using visual Box plots as was done previously.

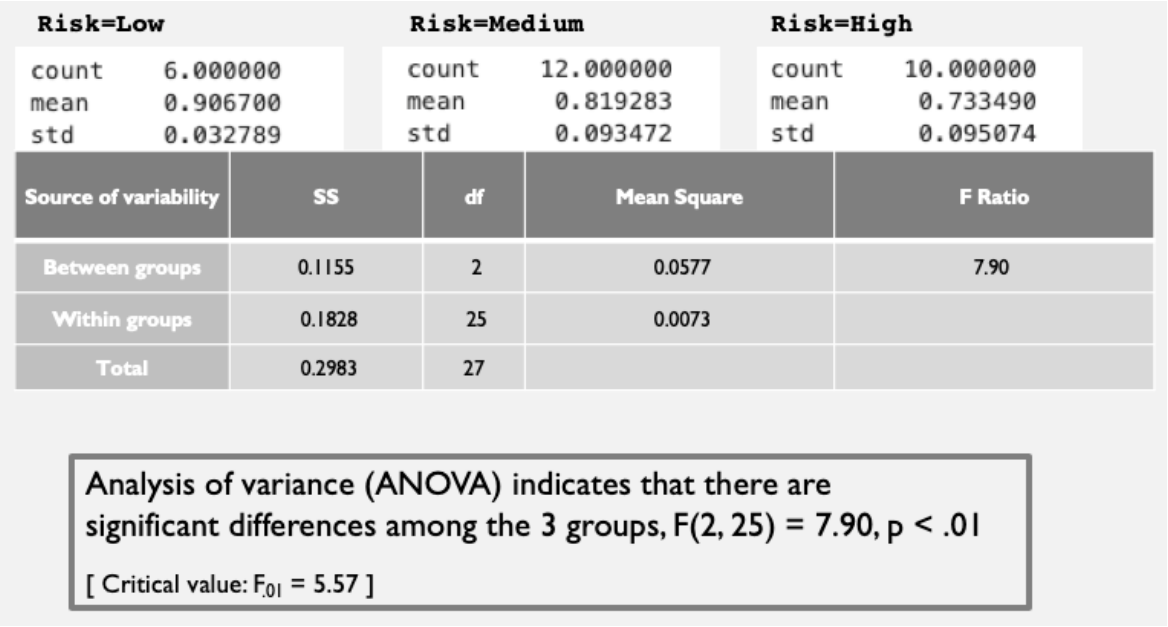
4. Results

It is clear, both visually through the box plots from notebook sections 6 and 8 (partially repeated below), and from the analysis of variance (ANOVA) from presentation slide 12 (also repeated below), that there is evidence of a strong correlation between lack of access to adequate food or medical facilities and lower graduation rates in neighboring High Schools.

Neighborhood Graduation Rate Descriptive Statistics
when Risk is *Low* (high number of grocery store availability),
Risk is *Medium* (average number of grocery store availability), or
Risk is *High* (little to no grocery store availability)



Analysis of Variance (ANOVA) for the same samples



Risk=Low

```
count     6.000000
mean      0.906700
std       0.032789
```

Risk=Medium

```
count    12.000000
mean      0.819283
std       0.093472
```

Risk=High

```
count    10.000000
mean      0.733490
std       0.095074
```

Source of variability	SS	df	Mean Square	F Ratio
Between groups	0.1155	2	0.0577	7.90
Within groups	0.1828	25	0.0073	
Total	0.2983	27		

Analysis of variance (ANOVA) indicates that there are significant differences among the 3 groups, F(2, 25) = 7.90, $p < .01$

[Critical value: $F_{.01}$ = 5.57]

This same visual indication of correlation was seen for the Medical Desert Risk Indicator that I calculated in notebook Section 6, although definitely with more overlap of the Interquartile ranges of the groups than in the case of the Food Desert Risk Indicator.

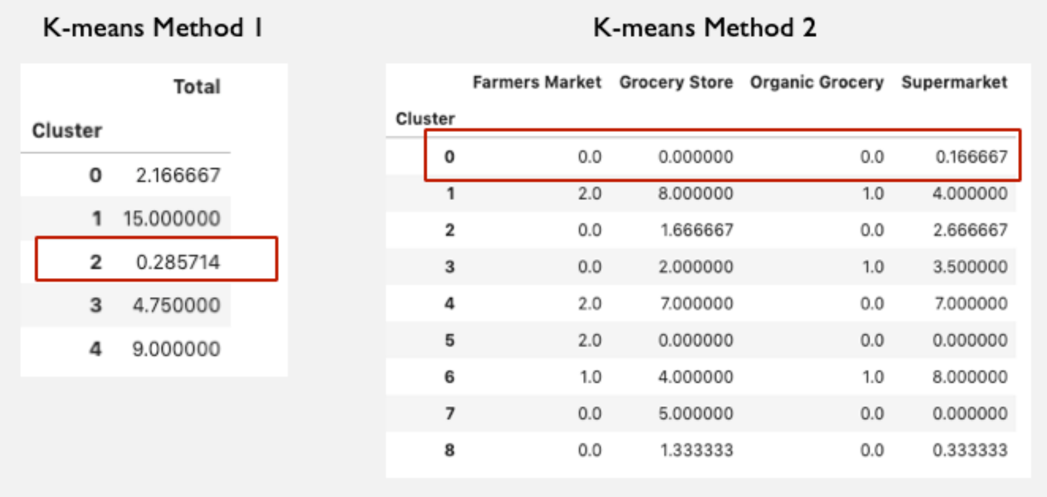
K-Means Clustering also led to good separation of groups that corresponded to having high or low mean graduation rates, both when using a total grocery store count to define the cluster and grocery subtypes, but again the overlap of Interquartile ranges appears to be a potential issue, particularly for those clusters with lower mean graduation rates.

Beyond the correlation question, it is interesting to look at the ability to identify Food Deserts from live location data. Slide 9 from the presentation deck shows a map generated in Section 6 of the notebook, with neighborhoods colored by Food Desert Risk Indicator. Alongside this map is a 2018 map generated by the Polis Center at IUPUI[5]. As indicated on the slide, the concentration of Food Deserts highlighted by the project's automated method aligns with many of the areas that their research also called out.

5. Discussion

I've already indicated in the Results section that the correlation between the initial Food Desert Risk Indicator and neighboring school graduation rates was shown to be statistically significant by ANOVA, but how does it compare to the other cluster-based solutions and which method appears to be the most robust? I would love to be able to say that machine learning was the answer, and granted I only tried a single machine learning approach in this project, but as a life-long fan of Occam's razor, here is another example where a simple approach appears to have some distinct advantages.

Cluster Characteristics from K-means



K-means Method I

Cluster	Total
0	2.166667
1	15.000000
2	0.285714
3	4.750000
4	9.000000

K-means Method 2

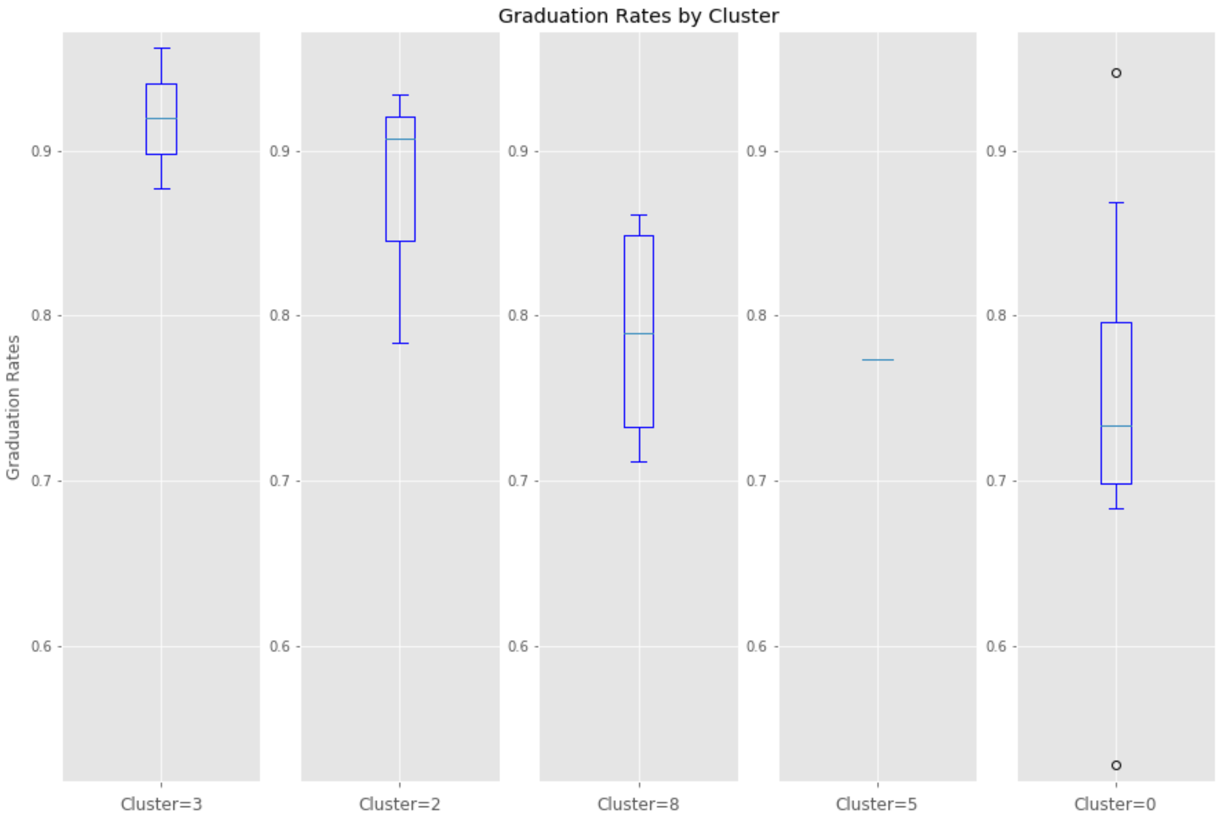
Cluster	Farmers Market	Grocery Store	Organic Grocery	Supermarket
0	0.0	0.000000	0.0	0.166667
1	2.0	8.000000	1.0	4.000000
2	0.0	1.666667	0.0	2.666667
3	0.0	2.000000	1.0	3.500000
4	2.0	7.000000	0.0	7.000000
5	2.0	0.000000	0.0	0.000000
6	1.0	4.000000	1.0	8.000000
7	0.0	5.000000	0.0	0.000000
8	0.0	1.333333	0.0	0.333333

If one examines either of the K-means based approaches, the resulting clusters as described on slides 14 and 15 in the presentation deck (and isolated above) both have the same potential critical flaw: they do not make a distinction between having a very low grocery store presence and having no grocery presence at all. Cluster 2 from Method 1 and Cluster 0 from Method 2 have *near-zero* means, but must include at least a neighborhood or two that has some grocery presence. Not that you would expect an unsupervised algorithm to factor this in, but in the case of Food deserts, it does seem to make sense that having even 1 convenient food source in a neighborhood would be a significant improvement and attractant over having none. The 'High' risk category of our initial statistical approach it turns out factors this in and contains only the neighborhoods with 0 stores.

Utility: the correlation of Food and Medical deserts to lower graduation rates could allow the risk indicators to serve as a proxy and early warning system. At a minimum, educators could use the measurements as hotspot indicators for areas where students might need a little extra attention due to the challenges of living in an area that is losing businesses and perhaps struggling to remain vibrant.

The potential of K-means Method 2:

Neighborhood Graduation Rates by Cluster (K-Means Method 2)



I had high hopes for the K-Means Clustering approach of using business subtypes, e.g. Grocery store vs Farmer's market, or Doctor's office vs Hospital, and perhaps if I had pursued the thread towards Medical deserts the scaling that I added would have had more dramatic effects on correcting for the disproportionate number of Doctor's offices that seemed to dominate the medical counts, but even on the Food desert side I noticed several interesting observations:

1. Cluster 5 seen above and on slide 15 in the presentation is a singleton cluster containing only seasonal Farmer's markets. If it weren't for that distinction, the neighborhood would have likely been placed in Cluster 0 to its right. The graduation rate for the sole school in Cluster 5 is several points higher than the mean for Cluster 0. Farmer's markets are often driven by community interest rather than business-driven, so perhaps that community interest also contributes to the higher graduation rates seen for that school.
2. Cluster 0 has two *outliers*, as indicated by the circles at the bottom and top outside of the whisker regions. The school at the bottom with a graduation rate approaching only 50% did not appear in the Foursquare data for the first few days that I worked on this project, so it was nice to see the advantage of working with live data that can change over time (although as an outlier it definitely stretched out the box plots when it started showing up). It also turns out that this school is no longer a traditional public school, with the state having recently taken it over, presumably in an effort to address low performance. On the other end of the spectrum is the school from this cluster with a greater than 95% graduation rate. That school happens to be one of only two in this cluster that does at least have a single grocery store in its proximity, and it too is no longer a traditional public school. It was originally the only charter school that was getting past my coding filters, but since it was a school that I happen to be familiar with and have had a summer intern from at the non-profit organization I work for, I decided to leave it in place to see how it affected the algorithms.

It's hard to talk about correlation without at least mentioning causation, so I hope it is clear that I'm not suggesting the latter relationship here. The solution to reversing low graduation rates isn't to simply add another grocery store or doctor's office into a neighborhood, put perhaps it's not far from that. Life is complicated enough without having to travel great distances for groceries and medical services, so areas with low numbers of critical amenities are not going to be attractive to families and could be at the risk of deteriorating. It takes a village to raise a child, as the saying goes, and that village should include adequate access to food and medical necessities.

6. Conclusion

I enjoyed this project, in fact wanted to savor it a bit by diving deeper into some areas than was perhaps necessary for a capstone project because I found the problem interesting and wanted to leverage as much from the series of IBM courses as I could. Despite what I think are interesting and compelling results, there are definitely some limitations to this approach that would need to be addressed to take it further towards real world application. I'll address a few of these below as potential next steps.

The use of Indianapolis as a test case benefited from the city's shape, namely an evenly developed city branching out almost equally in all directions and where even its surrounding beltway is fairly circular. Searching with the Foursquare API using a radius therefore worked well, but it remains to be tested whether other cities would be amenable to this method. Cities near water or other natural barriers might not cause great issue if the initially identified central point is close to downtown and on the water as it might be in Chicago, but unnatural borders such as state lines or twin and quad cities could be more challenging.

The linking of data to a secondary source, in this case school graduation rates, ended up having dual benefits. First, the intended use was an interesting study and the goal for this project, but it also helped to verify if the Foursquare data was up to date. Schools in the Indianapolis area have undergone a number of changes recently, with some High Schools being closed, others being repurposed for a different age group, and others becoming charter schools or focused on the arts or sciences. Since I wanted to make equal comparisons between traditional public schools, using the current Department of Education spreadsheet, which separated public schools from private schools and didn't report data for previously closed schools, created one-sided links that I could automatically ignore with no additional coding logic necessary.

The travel distinction between urban and rural areas still needs further investigation. I came up with some formulas that seemed to work well for Indianapolis based on my familiarity (or bias), but again this would need to be tested on other cities. The difference in magnitude of medical facilities versus grocery stores also calls for them to be approached differently. My interest was primarily in Food deserts, which didn't stretch the system, but I was pushing the upper bound on capturing the number of doctor's offices that could be captured with a free Foursquare account.

I still like the second K-Means Clustering approach that I tried and think that it logically has some merit over simple Total counts, so I might try a hybrid approach with it sometime that treats zero Total counts as a special case, given this problem space.

7. References

1. https://foodispower.org/access-health/food-deserts/
2. https://en.wikipedia.org/wiki/Medical_deserts_in_the_United_States
3. https://www.doe.in.gov/sites/default/files/accountability/2019-state-grad-rate-data-20191231.xlsx
4. http://indyfoodcouncil.org/wp-content/uploads/2015/09/Food_Deserts_w_Groc.pdf
5. http://www.savi.org/feature_report/getting-groceries-food-access-across-groups-neighborhoods-and-time/