Come 2023, Lake County, Illinois will be annexing new land into the county borders for the first time since the early 1900s[1]. The new community – located on previously-unincorporated land lying just beyond the current county borders – is represented by the ZIP code 60100[2], and is slated to join the 27 ZIP codes that are currently in the county. Although the move is a welcome one for Lake County, as the added residents should bolster the local tax base and help the county cross a new population threshold that will qualify it for increased federal funding, the annexation comes not without its' problems to sort out. One such problem has been identified by members of the Lake County Department of Health (LCDH), which is set to administer and disburse funds for a new state program for cancer prevention that was just recently passed into law through the 2022 state budget. The program has awarded *just enough* funds for the LCDH to target a handful of communities that collectively are at the highest risk for cancer development, and the department had already started to tackle the problem of which three communities would most benefit from receiving those funds. The addition of the 60100 community in 2023, however, upends things. How do cancer rates in the new community compare to the 27 communities already in the county? How should the LCDH choose the communities that will receive the new state funds for cancer prevention? And will 60100 be one of those communities chosen? The LCDH now finds itself having to tackle all of these questions.

As the Director of Analytics for Lake County, the LCDH has engaged me to help solve these problems using machine learning techniques. Reaching into my machine learning tool-belt – which includes algorithmic tools such as *k-nearest neighbors*, the *Naïve Bayes Classifier, density-based spatial clustering (DBSCAN)*, and h*ierarchical clustering* – it is my thought that the best tool for tackling the DOH's problem is *k-means clustering*. My reasoning behind this is that I can solve the LCDH's problem by grouping the existing ZIP codes within Lake County together based on similar characteristics – in this case, their varying rates of different types of cancer. Not only can I then use those groups to provide guidance to the LCDH on which communities are at the most risk for developing cancer, but I can also use those existing groups to predict where the new 60100 community will fall in relation to the other communities' cancer rates (and whether or not it should be one of the beneficiaries of the new state funds). What I've described here is a clear example where *clustering* would be of use, as, according to Shmueli's *Data Mining for Business Analytics*, clustering's primary goal is to "segment the data into a set of homogenous clusters of records" where "objects within a cluster are similar", and "objects from different clusters are dissimilar".

Furthermore, after viewing the LCDH's dataset (publicly available through catalog.data.gov[3] - see Figure 1), I can see that the numerical columns in the data are all already on the same scale – number of cases per 100,000 people – and that while there are certainly varying rates of all cancers among the communities, all of the rates do exist within a relatively static range (it would be highly unusual, for example, to have one community with a cancer rate in the thousands per 100,000, and then have that community be right next-door to another community with a cancer rate in the tens of thousands per 100,000). Knowing this, I can tell right away that *Euclidean distance* would be an

[1] Note: this is a hypothetical, fictional scenario created specifically for the purposes of this assignment. As far as I know, Lake County is not actually planning to annex a new community in 2023.

[2] Again, this is a real ZIP code that belongs to an actual, existing community; I am merely using it as an example here as part of the assignment.

[3] Note: while the scenario I pose in this assignment is hypothetical, the dataset I use is an actual publicly-available dataset published by Lake County.

applicable distance measure to use for setting up the clustering model; even though the distance calculation method is scale-dependent and sensitive to outliers, the uniform scale and lack of severe outliers in the LCDH dataset renders those issues moot. Indeed, when viewing the color-coded "heat map" diagram of the Euclidean distances between the LCDH data points (see Figure 2), my initial observation that the county's cancer rates reside within a static range is borne out by the relative lack of blue on the diagram ("blue" representing the largest distances between points). Since Euclidean distance lends itself particularly well to use with the *k-means clustering* algorithm, that solidifies my choice of *k-means* as the best machine learning tool to help solve this problem.

The initial requirement of *k-means*, of course, is that I know the number clusters to use in advance. To ascertain this, I can use the *elbow method*, which produces a clear elbow bend at $k = 3$ (see Figure 3). Now knowing the ideal number of clusters, I can proceed with running the *k-means* algorithm on the LCDH dataset, using each of the varying types of cancers' rates as the numerical variables. *Figure 4* in the appendix below visualizes the existing 27 ZIP codes of Lake County clustered by cancer rates, but what follows here is a description of those findings:

Cluster 1	
Title	"Low-Cancer Rate ZIP Codes"
Number of Communities	13
Cluster Center	2,070 cases per 100,000
Community	**Cancer Rate per 100,000 (All Cancers)**
60085	1,465.29
60073	1,533.54
60040	1,796.29
60064	1,830.42
60087	2,092.97
60060	2,174.09
60083	2,205.74
60031	2,217.83
60047	2,261.91
60042	2,267.41
60099	2,317.37
60046	2,340.86
60061	2,409.73

Cluster 2	
Title	"Medium-Cancer Rate ZIP Codes"
Number of Communities	11
Cluster Center	2,924 cases per 100,000
Community	**Cancer Rate per 100,000 (All Cancers**
60030	2,581.85
60084	2,596.58
60048	2,697.71
60002	2,703.15
60015	2,922.59
60089	2,991.53
60096	2,995.65
60020	3,084.13
60044	3,149.85
60041	3,200.46
60010	3,248.83

Cluster 3	
Title	"High-Cancer Rate ZIP Codes"
Number of Communities	3
Cluster Center	3,959 cases per 100,000
Community	**Cancer Rate per 100,000 (All Cancers**
60045	3,611.81
60035	3,760.43
60069	4,505.48

What these findings show is that, while cancer rates certainly vary from community to community across the county, there are three communities separated from the rest that should be clear recipients of the new state funds: the three communities that make up **Cluster 3**, "High-Cancer Rate ZIP Codes". These communities – 60045, 60035, and 60069 – have overall cancer rates ranging from 3,611 to 4,505 cases per 100,000 residents – clear steps above the center of the "Medium-Cancer Rate ZIP Codes" cluster (2,924 cases per 100,000). So while this clearly solves the LCDH's question of where to direct the state funds for cancer prevention in the *current iteration* of Lake County, there is still the outstanding question of how the 60100 community fits in. Will the new community have an even *more urgent* case for needing state funds than the three members of Cluster 3?

To tackle this final problem, I can use my existing *k-means* cluster model to predict the cluster placement of the new community. If 60100 places in Cluster 3, for example, the LCDH may have to consider allocating funds there as opposed to one of the three communities currently in Cluster 3 (or to allocate funds to all four communities, but in smaller amounts – each getting a "smaller piece of the pie", if you will). *Figure 5* in the appendix shows the cancer rates for varying types of cancer in the 60100

community. Applying my previous cluster model to this data point, I receive a result of **Cluster 2**, "Medium-Cancer Rate ZIP Codes". Indeed, even though 60100's overall cancer rate of 3,010 cases per 100,000 appears to be relatively elevated, due to exceeding the 3,000 cases per 100,000 mark and placing within the top 10 highest overall cancer rates for Lake County communities, my cluster analysis illuminates that that figure is much closer to the middle of the pack – meaning Cluster 2's center – than to the communities with the most severe cancer rates represented by Cluster 3's center of 3,959 cases per 100,000.

So what will I report back to the Lake County Department of Health? My summarized findings are:

- Even though the new 60100 community that will be added to Lake County in 2023 does have a relatively high cancer rate - placing within the top third of communities county-wide for overall cancer cases per 100,000 residents – it is **not** one of the best recipients for the new state funds for cancer prevention.
- The communities which would benefit the most from the additional state funds are **60045, 60035, and 60069**, represented by *Cluster 3.*

Appendix

Figure 1 – Snapshot of the Lake County Department of Health's publicly-available dataset

	ZIP <chr>	Colorectal <dbl>	Lung_Bronc <dbl>	Breast_Can <dbl>	Prostate_C <dbl>	Urinary_Sy <dbl>	All_Cancer <dbl>
1	60002	218.0621	419.6667	399.0948	259.2059	259.2059	2703.148
2	60010	258.9157	335.4647	504.3228	499.8199	227.3955	3248.829
3	60015	153.4359	230.1538	478.5738	442.0414	222.8473	2922.588
4	60020	292.7972	507.5151	214.7179	302.5571	370.8764	3084.130
5	60030	221.5354	284.4406	404.7808	322.7306	210.5954	2581.845
6	60031	163.5021	221.5190	414.0295	303.2700	160.8650	2217.827

Figure 2 – Euclidean distances between the cancer rates of Lake County, IL communities

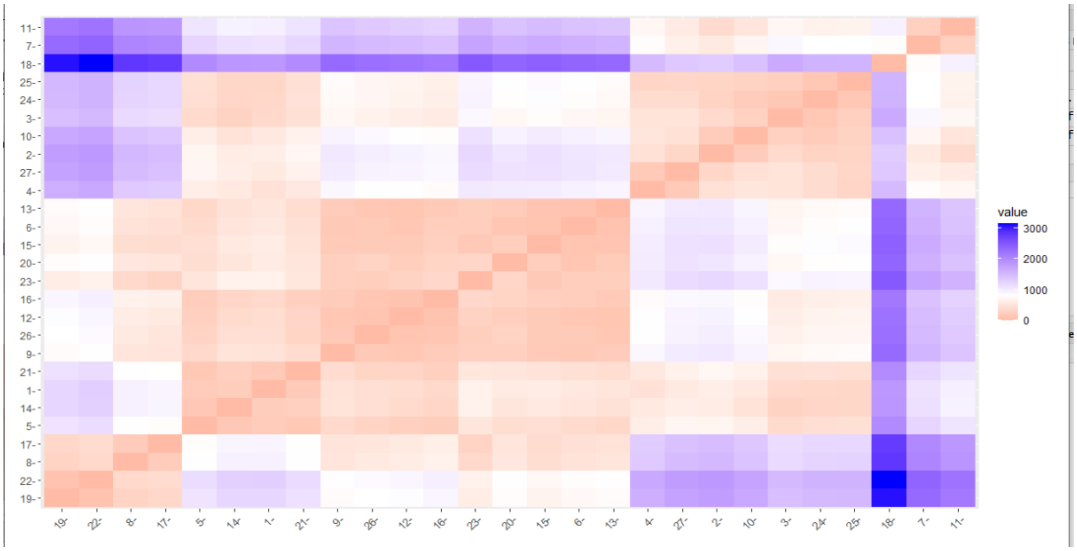

Figure 3 – Elbow bend curve of the Lake County, IL communities' cancer rates

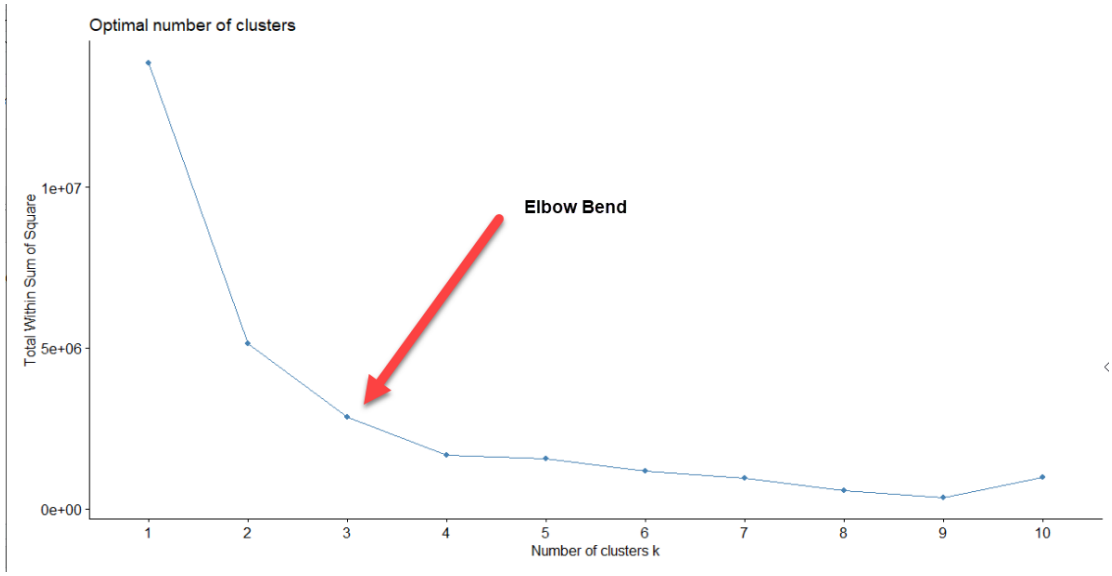

Figure 4 – K-means cluster distribution of Lake County, IL cancer rates

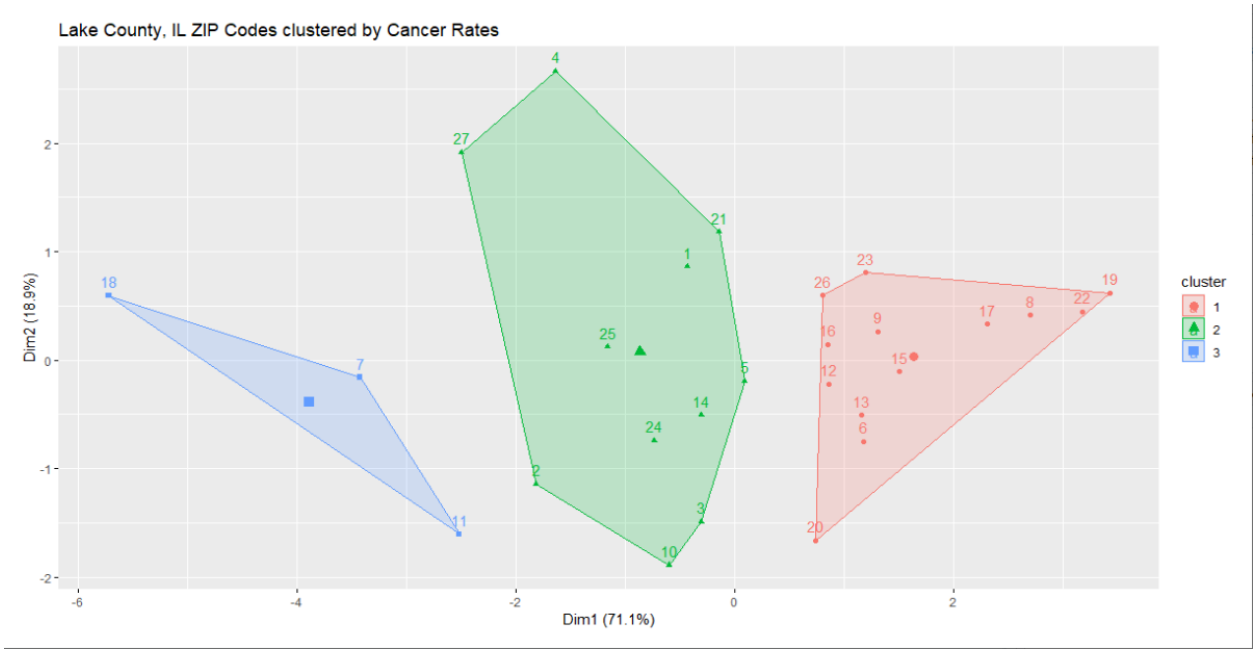

Figure 5 – Cancer Rates of 60100 (Lake County, IL's 2023 annexation candidate)

ZIP	Colorectal	Lung_Bronc	Breast_Can	Prostate_C	Urinary_Sy	All_Cancer
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
60100	270	150	250	350	275	3010

References

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for Business Analytics: Concepts, techniques and applications in Python*. John Wiley & Sons, Inc.