Clustering Report

Enzo Profli

1 Executive Summary

One of the most pressing health problems in the United States is that of the opioid epidemic. Thousands of American die each year due to opioid overdoses, often prescribed by legitimate doctors across the country. As explained during the text, literature indicates that these doctors receive many patients in so called "clinics" and often overcharge for prescribing certain drugs. However, in many cases, the number of patients these doctors receive is limited by law.

In this scenario, a K-means clustering analysis has been performed to assess if doctors in areas with more drug overdose deaths and who receive a larger number of patients charge more for their services. The analysis indicates that this is not case: areas with relatively fewer overdose deaths and few-patient doctors charge more for medical visits.

There are a couple of possible explanations to this phenomenon. The first is that doctors in more affluent areas have naturally fewer patients, and the second is that these number of patients restrictions are binding enough that doctors overprescribing opioids serve fewer patients.

2 Problem Statement

There is abundant investigation about medical practices in the United States that overprescribe opioids, contributing to the current opioid crisis in the United States. For example, Sontag (2013) discusses that addicted patients are subject to overcharges and medical practice closures, but also discusses that, in many cases, the number of patients for doctors prescribing opioids are limited by law. Hoffmann (2008) analyzes the example of a California doctor who was the largest provider of opioids in the states, and overcharged visits by $150. Given this discussion, are medical visit costs related to pain management larger in counties with large presence of opioids? And are these costs larger for doctors who receive more patients? To answer this question, we will utilize the Provider and Other Supplier Public Use Data from the Centers for Medicare & Medicaid Services (CMS) and Drug Poisoning Mortality in the United States by County from the National Center for Health Statistics (NCHS).

3 Assumptions

- This analysis assumes that doctors operate in individual cities, and that patients do not travel far to visit doctors.
- Obviously, the analysis assumes that the provided data is accurate, and that no errors about doctor visit types and about each doctor's area of expertise are made.

4 Methodology

The first step in this process is cleaning and merging the relevant data, by county. Then, we subset the data for medical visits to pain management physicians, leaving us with 7700 observations (7200 after merges and cleaning). Once the data is adequately organized, we generate our variable of interest: Medicare surcharges.

As described in the CMS data dictionary, Medicare and Medicaid allow a certain amount of funds for each type of medical service (allowed amount), and then doctors submit service charges (submitted charge), usually above the allowed amount. So, for each physician, we have the average surcharge (relative to Medicare-allowed amounts), as defined by Equation (1) below:

$$Surcharge\ Ratio = \frac{Allowed\ amount}{Average\ submitted\ charge} \tag{1}$$

The next step is clustering the data on the number of patients and the county drug poisoning mortality rate (a proxy for opioid overdoses on a county level). We will be using the K-means clustering algorithm, since we are dealing with Euclidean distances. In order to cluster the samples, we normalize these variables and take their logarithm, so that they resemble normal distributions, as shown in Figure 3 in the Appendix. However, note that we are averaging over different types of services (say, 15-minute doctor visits and 45-minute doctor visits), which have plausibly very different number of patients per physician. So, when normalizing the number of patients, we do so in groups. For example, if a doctor received 50 15-minute visits, and the maximum number of 15-minute visits a physician has had was 100, the normalization yields 0.5, while a doctor who received 50 30-minute visits (and that was the maximum) will be normalized to 1 - even though the number of patients is equivalent. Finally, we want to cluster by doctor, and doctors provide different services, and so we take the normalized number of patients of that doctor to be the mean of the normalized number of patients in each service. We perform a similar operation to yield the doctor's average surcharge. After grouping by doctor, we possess 2005 observations.

5 Analysis

Figure 4 in the Appendix displays the total sum of squares to centroids using up to 20 clusters. We see that the total sum of squares does not decrease significantly below 6 clusters, so we will be using 6 clusters in our analysis. Figure 1 below displays our clustering on the number of patients and the county's opioid overdose death rate. Cluster 1 represents "central" values while clusters 2-6 represents clusters outside the center (for example, cluster 6 groups observations with large number of patients, but variable number of county overdose deaths).

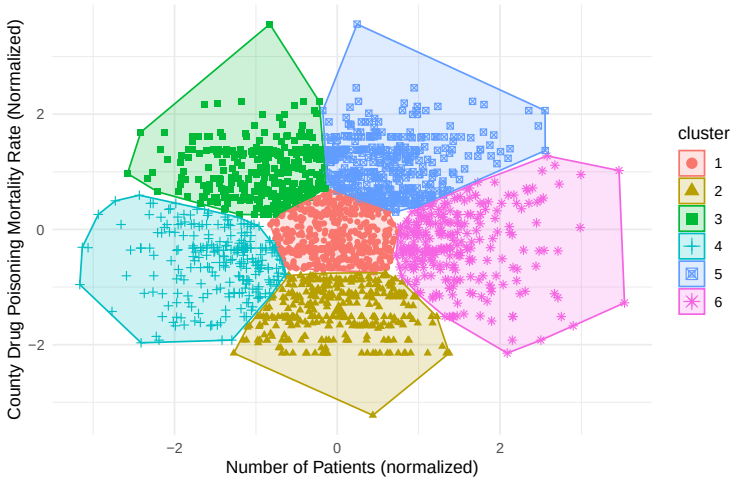

Figure 1: Cluster Plot on normalized variables

Now that we have the clusters, we can see the surcharge distribution for each cluster, as displayed in Figure 2. We see some interesting variability in the distributions. For example, cluster 1's distribution is very

concentrated on lower levels, while cluster 4 has a much smaller peak over a larger range of surcharge ratios. We can also see some interesting density increases in some clusters. For example, group 4 has a small peak on a level slightly below 10, while groups 1 and 2 have small peaks in levels above 10. To gain a more analytical look at these results, we will inspect some summary variables in the next paragraph.

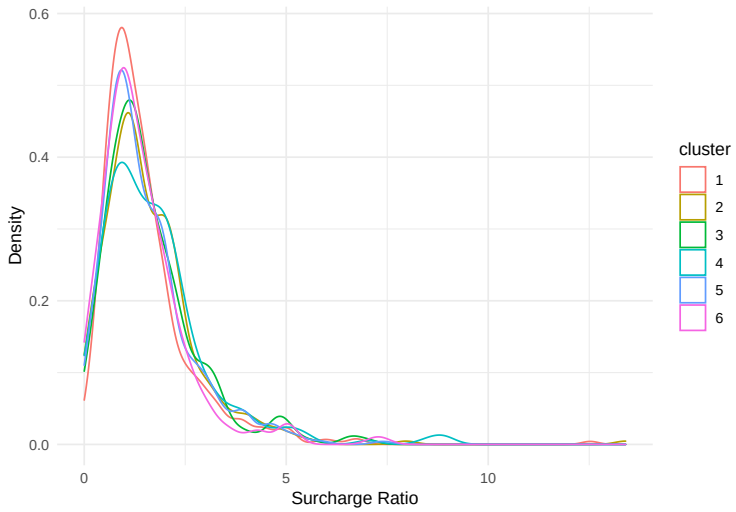

Figure 2: Surcharge ratio density plot for each cluster

Table 1 below displays some summary variables for each cluster. We see that cluster 4, with low patient number and low death rates, contains the highest median surcharge ratio, while the lowest is on group 6 (high patient amount, average death rates). Cluster 5, with high amount of patients and high overdose death rates, present a substantially lower median surcharge ratio. I suspect that this is due to socioeconomic factors, as it is plausible that less affluent areas contain both a larger amount of overdose deaths (deaths of despair) and a higher proportion of patients per doctor.

Table 1: Summary Statistics per cluster

cluster	Mean Surcharge	Median Surcharge	Min Surcharge	Max Surcharge	Proportion above 2	Cluster Description
1	1.52	1.22	0.02	12.53	0.09	Average patients, average mortality
2	1.63	1.39	0.00	13.40	0.11	Average patients, low mortality
3	1.63	1.34	0.00	7.03	0.11	Few patients, high mortality
4	1.74	1.42	0.00	8.90	0.12	Few patients, low mortality
5	1.54	1.25	0.00	7.39	0.10	Many patients, high mortality
6	1.43	1.19	0.00	7.40	0.06	Many patients, average mortality

6 References

Hoffman, D. (2008). Treating Pain v. Reducing Drug Diversion and Abuse: Recalibrating the Balance in out Drug Control Laws and Policies. Saint Louis University Journal of Health Law & Policy. Vol. 1: pp. 231-210.

Sontag, D. (2013). Addiction Treatment with a Dark Side. New York Times. Retrieved from: https://www.nytimes.com/2013/11/17/health/in-demand-in-clinics-and-on-the-street-bupe-can-be-savior-or-menace.html

7 Appendix

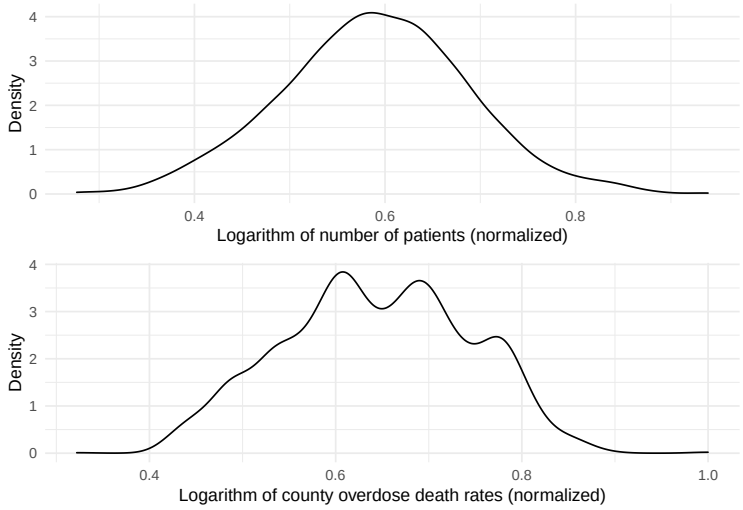

Figure 3: Distribution of clustering features

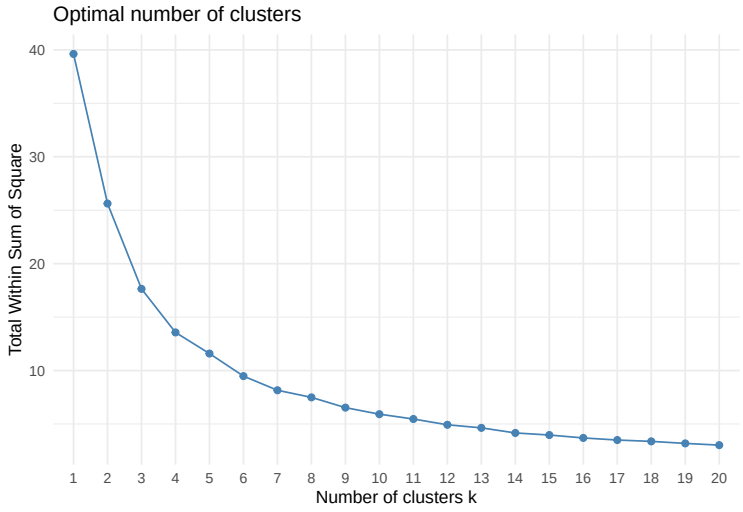

Figure 4: "Elbow Method" to determine number of clusters