Andre Schweizer
Prof. Diego Klabjan
IEMS 308 – Data Science & Analytics
January 23rd, 2018

HOMEWORK ASSIGNMENT I
Clustering

I | EXECUTIVE SUMMARY
Medicare is the federal health insurance program for people over 65 years old, younger people with certain disabilities, and people with end-stage renal disease. To better serve its stakeholders (i.e. beneficiaries and service providers), as well as make the processing of charges more efficient, it should understand what is it that beneficiaries are using the insurance for.

Clustering is a method that groups a set of data samples according to the similarities they have in different features. In Medicare's case, we can use this technique in the information it supplies about the usage frequency of the services it covered in 2016 and their average submitted charge. This will allow us to identify the price and volume of services and characterize, to a first extent, the kind of procedures Medicare beneficiaries are trying to use the insurance for.

After clustering the data into five groups using the K-Means algorithm, I found that the vast majority of charges submitted are for low-cost procedures, and that, despite a high aggregate volume, these are being offered by many small providers rather than few bigger ones. It was also possible to see that the opposite is true for more expensive services: they are low in volume and have a consolidated supply. By incorporating these insights into its marketing campaigns and operations, Medicare could be make itself more appealing to a higher number of beneficiaries, as well as increase the efficiency in its processing of charge submissions.

II | PROBLEM STATEMENT
The Medicare Physician and Other Supplier Data CY 2016 contains several pieces of information on the services provided in 2016 to beneficiaries under Medicare Part B, which covers certain doctor services, outpatient care, medical supplies, and other preventive services (i.e. excludes hospital bills and drug expenses)[1]. Having that in mind, I would like to investigate what beneficiaries in Illinois are trying to use Medicare for, and what are the subsequent implications. More specifically, I want to see whereas beneficiaries are using Medicare to pay for more expensive, low volume procedures or cheaper, higher volume ones. In other words, are beneficiaries trying to use the insurance for more serious health problems (e.g. a heart surgery) or more mundane issues (e.g. visits to allergists)?

[1] https://www.medicare.gov/what-medicare-covers/your-medicare-coverage-choices/whats-medicare

III | ASSUMPTIONS

- *Providers submit charges for all procedures performed on Medicare beneficiaries*: it is possible that some patients pay for some of the procedures out-of-pocket. Because there is no data on these, we assume that the submitted charges are an accurate reflection of the actual cost of the services.
- *List of services in data is exhaustive:* all services used by Medicare beneficiaries are listed in the Medicare Physician and Other Supplier Data document – it is a truthful depiction of what people are trying to use the insurance for.
- *No errors in data entry:* I assume there are no typos in the data, as I wouldn't be able to predict what these were. Data entry errors would skew the results by a bit, as numbers would be different than they should've been. But taking into consideration the theoretical rarity of these errors in comparison to the size of the data, it's safe to assume their impact would be negligible.
- *Normally distributed feature values*: for outlier detection, I assumed the values for each of the features to be normally distributed, especially since AVERAGE_SUBMITTED_CHRG_AMT is a mean value.

IV | METHODOLOGY

1. *Clean data*

   To make the code more efficient, I removed the columns and rows I wasn't interested in. Out of the 26 columns present in the initial data file, the information needed for my investigation was contained in only 3 of them: BENE_DAY_SRVC_CNT, NPPES_PROVIDER_STATE, and AVERAGE_SUBMITTED_CHRG_AMT. Hence, I first removed all rows that had a NaN value for either BENE_DAY_SRVC_CNT or AVERAGE_SUBMITTED_CHRG_AMT. I then removed all rows whose services weren't provided in Illinois. Finally, I only kept the BENE_DAY_SRVC_CNT and AVERAGE_SUBMITTED_CHRG_AMT columns.

2. *Pre-process*

   The first step was to normalize the data to ensure that the difference in scale between features doesn't make some of them "overpower" others in the clustering process. Hence, from each sample-feature data point, I subtracted the feature's mean and divided by its standard deviation. Thus, each column now had a mean close to 0 and a standard deviation close to 1. The second step was to remove outliers, as some of the data points were up to 5 standard deviations away from the mean and, thus, greatly impacted the clustering results. And so, I removed all values that were more than 2.33 standard deviations away from the mean for both features – around 99% of the data was still being used in the analysis.

3. *Cluster*

   K-means was the clustering method selected for this analysis. First, the selection of the optimal number of clusters was two-fold: it included the elbow method using a Scree Plot and visualizing a Silhouette Plot. For the elbow method, I created and fit a K-Means object to the cleaned data 10 times, each time using a different number of clusters from 1 to 9. I then calculated the distortions, the within-group sum of squares, for each of the number of clusters. The kink on the graph indicated that 5 was optimal. For the silhouette plot, I implemented an

algorithm found on the scikit-learn website[2] to make silhouette and clustering graphs for each of the number of clusters, from 2 to 9. The second plot showed that 4, 5, or 6 would be reasonable choices. Taking into consideration the results of the Scree Plot, I determined the optimal number of clusters to be 5, and selected the corresponding cluster plot for the analysis below. *See Appendices A and B for plots.*

V | ANALYSIS

To determine what Illinois Medicare beneficiaries are trying to use the insurance for, the main metric used was the relationship between the average submitted charge by the provider of each service and how many times that service was offered in 2016. This allows us to look at the location of the services in terms of those two variables and determine whether beneficiaries are using the insurance more for cheaper, lower volume procedures or more expensive, higher volume ones. Figure 5.1 gives us an initial overview of this:

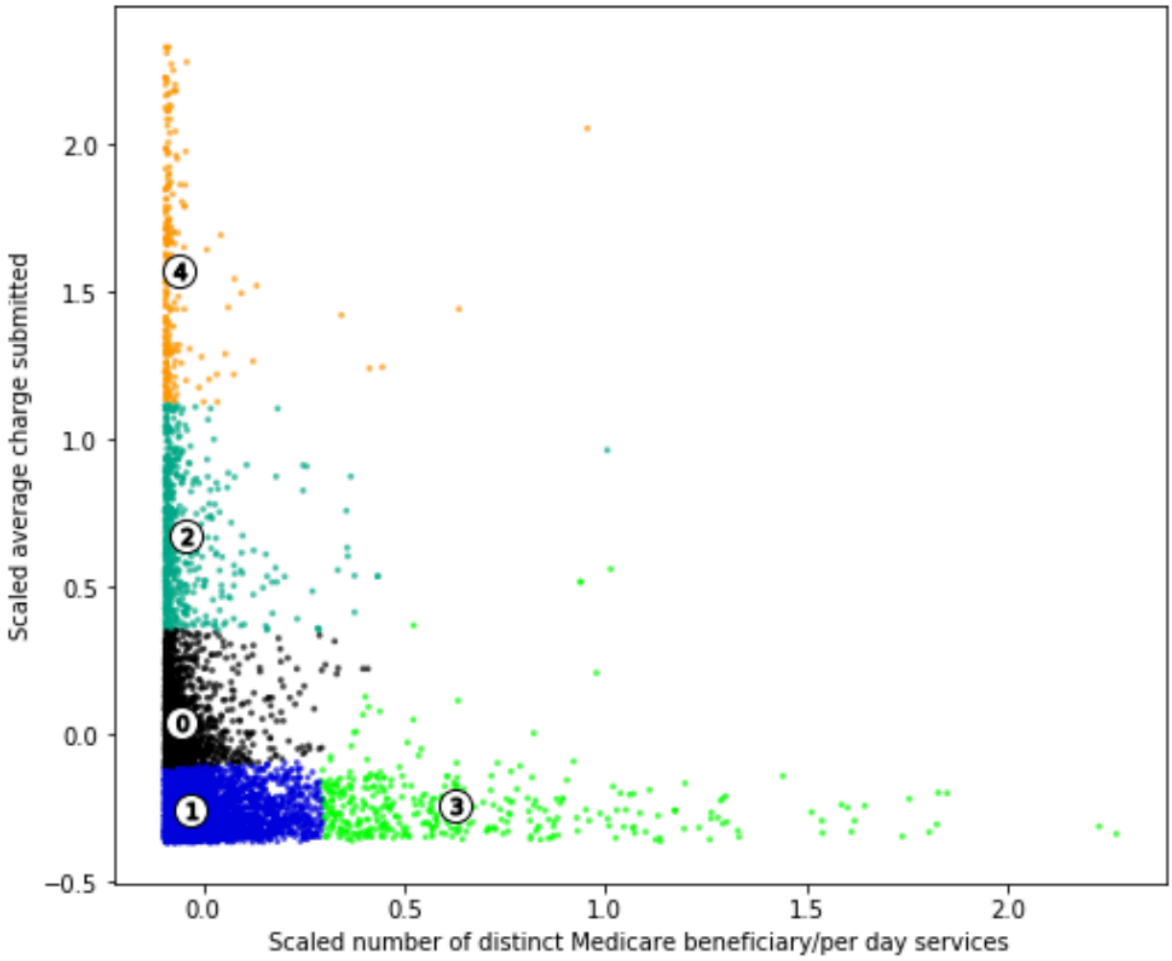

Figure 5.1: Visualization of clusters

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

It's interesting to note a few things:

- There are practically no services that were very high in both the average charge submitted and the frequency it was provided. This makes sense, as it would be hard to picture a provider that offered expensive services and still obtained high volumes – this would probably be a monopoly of some kind and showed up in the outliers that were previously removed (e.g. the University of Michigan hospital).
- The vast majority of services (314,357, or 81.5%) have average submitted charges less than or equal to the mean, indicating that relatively cheap services are by far the most used ones.
- The vast majority of services (305,054, or 79.1%) have average number of times it was provided less than or equal to the mean, indicating that relatively low volume services are by far the most used ones.
- Samples are more distributed for average charge submitted than for total number of services provided (i.e. there are more samples further from the mean in the former than the latter).

Moreover, some statistics on each of the clusters also reveals interesting insights:

Table 5.1: Statistics on clusters and overall data (scaled values)

Cluster	Mean of total # services provided	Mean of average charge submitted	# of data points	Classification
0	- 0.055	0.035	78,966 (20.5%)	Average cost, low volume, medium-consolidated supply
1	- 0.037	- 0.255	258,179 (67.0%)	Low cost, low volume, highly consolidated supply
2	- 0.054	0.666	23,369 (6.1%)	Medium cost, low volume, medium-consolidated supply
3	0.598	- 0.247	16,744 (4.3%)	High volume, low cost, low-consolidated supply
4	- 0.065	1.627	8,364 (2.2%)	High cost, low volume, low-consolidated supply

- The breakdown of number of samples in each cluster allows us to look into the distribution of services a little deeper to see that in fact, low cost and low volume procedures (grouped in cluster 1) make up the majority of data points. This is could be due to three reasons: (1) most services used by Medicare beneficiaries are "single-time ones" (i.e. a physical therapist doesn't charge for multiple sessions when a patient visits them only once), (2) most services aren't performed as frequently in individual patients, but are needed by a lot of the beneficiaries (e.g. flu shots), or (3) and most likely, cheap services are reported more times, by a greater number of providers. In other words, cheap services in cluster 1 are high in aggregate volume but they get broken down into lower volume data points because they are offered by a wider range of providers.
- Cluster 4 represents expensive, low volume procedures like surgeries and prostheses. The high average charge submitted and small number of data points indicates that these are specialized services that are only being offered by few providers and being used by few beneficiaries.

- Samples in cluster 3 are similar to those in cluster 1: they are high in volume and low in price. The difference, however, is that these are more specialized services and, thus, are being reported by fewer providers – there are less data points and their individual volumes are higher due to the aggregation in supply.
- Clusters 0 and 2 are intermediaries between 1 and 2.

VI | CONCLUSIONS

The analysis in (V) reveals insights related to the purpose proposed for this investigation:

- *Medicare beneficiaries are trying to use Medicare mostly for cheaper, higher volume services*
  A possible implication of this for Medicare is related to quality control. Because most of these services are commonplace, it is hard to ensure that they are being delivered with care and are aiming at the beneficiary's best interest – Medicare would have to pay more attention to those in comparison to others. Another implication has to do with fraud. It is much easier for a provider to make up a fake cheap procedure than it is for an expensive one. Having in mind that cheaper services make up the bulk of submitted charges, Medicare's anti-fraud procedures should be especially geared towards those. Lastly, when encouraging people to use Medicare insurance, the government should create campaigns that depict these simpler, higher volume services, as those are clearly in demand by more Medicare beneficiaries.

- *Commonplace, high volume services have diversified supply*
  Medicare's information systems should be geared towards receiving forms from multiple different places rather. It should simplify the charge submission form (since most of them are only submitting a couple with only a few things on them), and streamline the receiving end to ensure there's no backlog due the high volume of small requests. Moreover, it should ensure that beneficiaries are aware that there are so many providers who work with Medicare, as this would help someone find a provider close to them.

- *Expensive services are low in demand and have limited supply*
  This makes it easy for Medicare to perform quality and fraud control on the high-priced, critical charges it receives. Since only few providers are submitting those, they could make more detailed questions and maybe even visits to ensure beneficiaries are receiving the best treatment, and that no one is making fortunes out of health insurance fraud.

VII | NEXT STEPS

This investigation fomented further questions that would also yield insights relevant to the question originally posed.

- It would be interesting to use text analytics to explore the HCPCS_DESCRIPTION column of the data, which contains the descriptions of the services being provided. These could be mapped and compared to the plots in (V) to see what kinds of services fit in each of the clusters.
- Similarly, we could look at the PROVIDER_TYPE column and see what providers fit in which clusters to analyze the range of services offered by each provider (i.e. a chiropractor is likely

to only offer medium volume, medium cost services whereas a large hospital would probably offer procedures in all different points of that range).

- It would also be insightful to look at the frequency that each of the services is needed and compare that to how often they are being covered by Medicare. In this analysis, we looked at the frequency people use Medicare for each kind of service, but it is obvious that higher demand services will be more seeked. A better metric for the effective usage of the insurance would be the proportion between how many times beneficiaries used Medicare to pay for a service out of all the times they actually needed it – is Medicare covering users' needs?
- This clustering method yielded reasonable clustering results (with silhouette average values around 0.6 and clear clusters formed). The nature of the data, however, makes clustering hard, as the graph of the data points is a steep curve. Hence, it would be interesting to try different transformations to the data or other clustering algorithms to assess the quality of those results in comparison to the ones obtained in this investigation.

A | SCREE PLOT

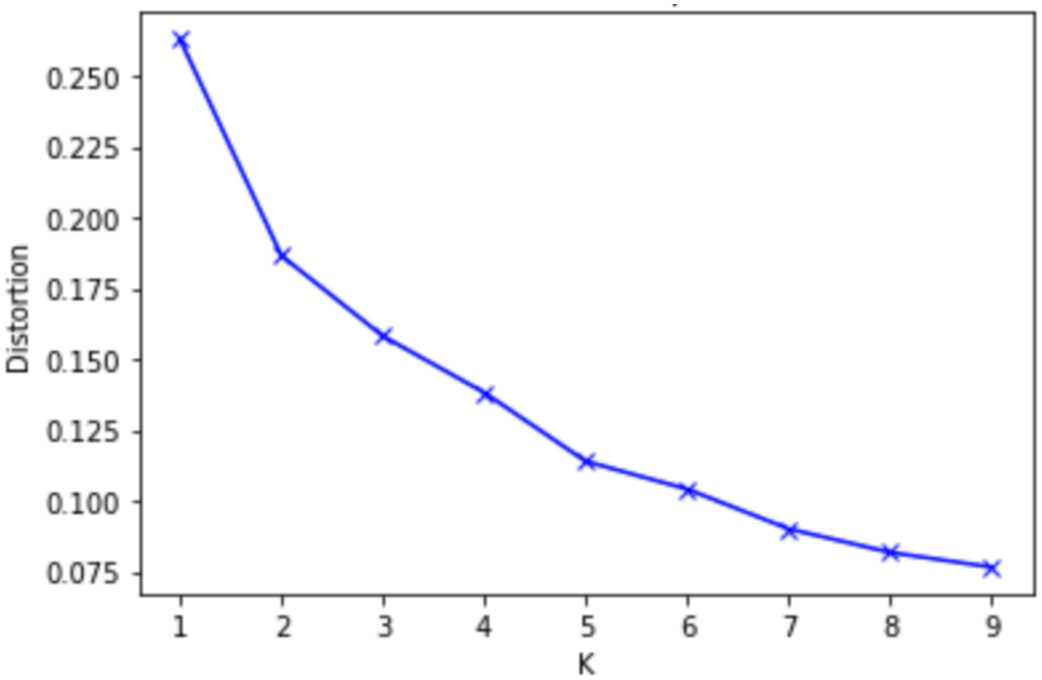

Figure A.1: Scree Plot used to find optimal number of clusters

B | SILHOUETTE PLOTS

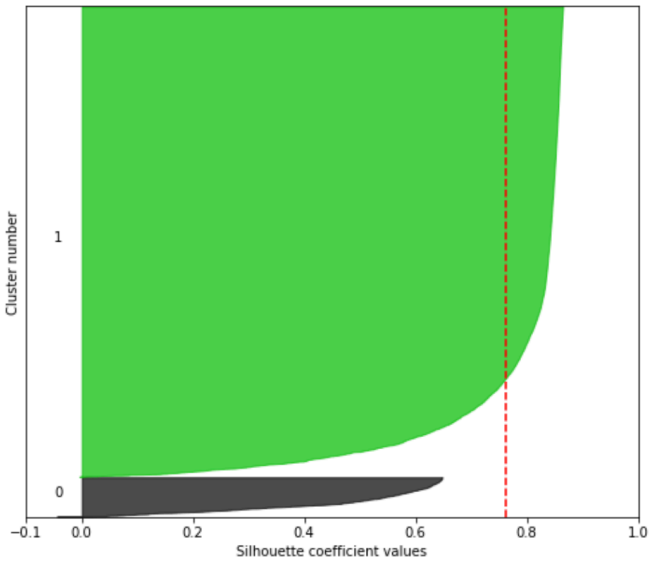

Figure B.2: 2 clusters

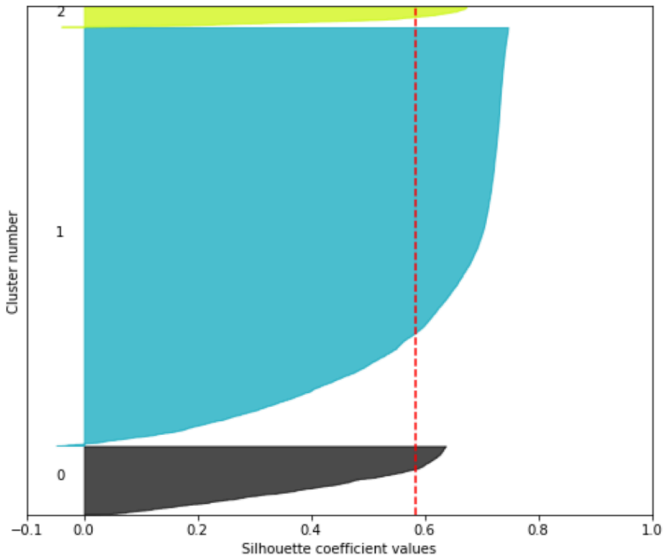

Figure B.2: 3 clusters

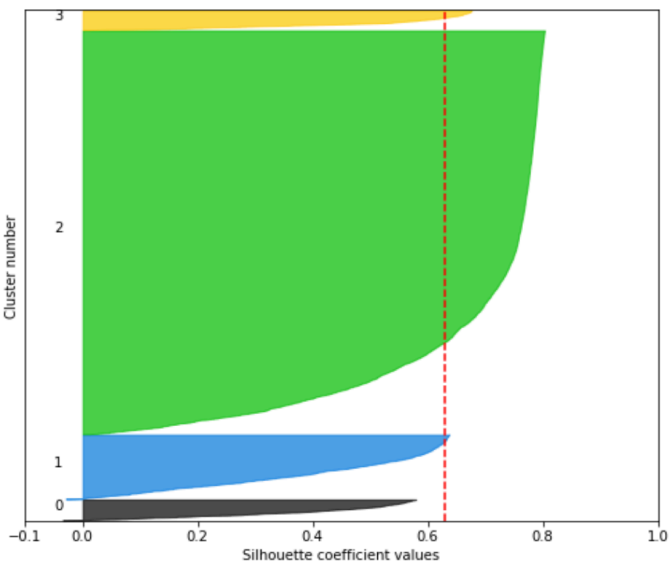

Figure B.3: 4 clusters

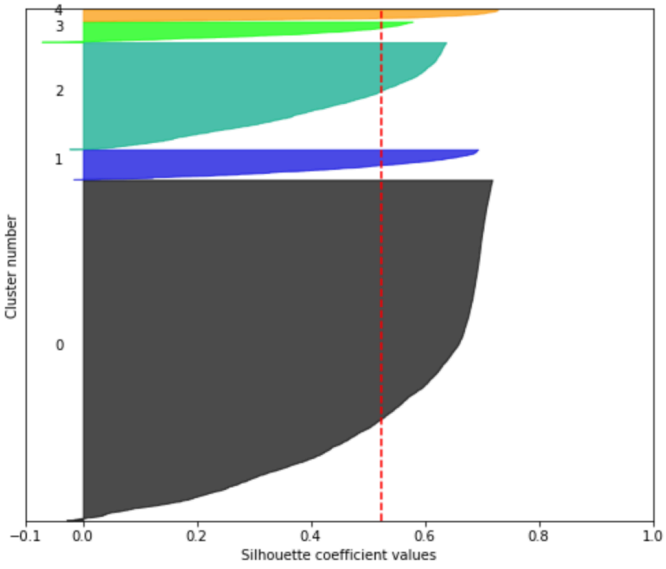

Figure B.4: 5 clusters

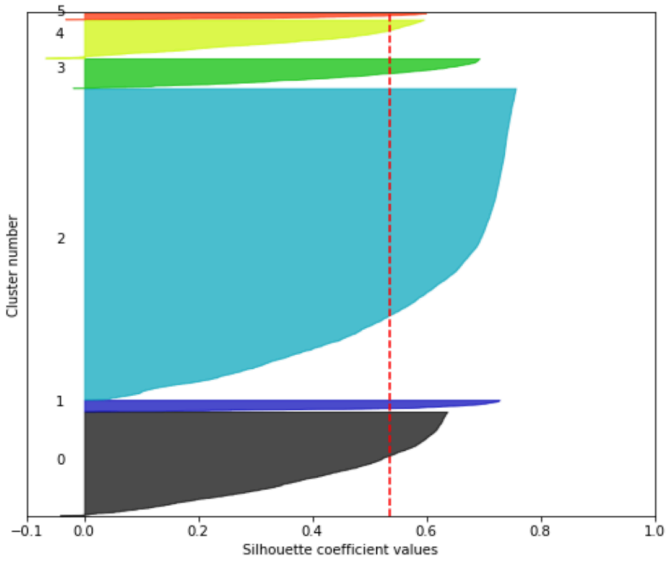

Figure B.5: 6 clusters

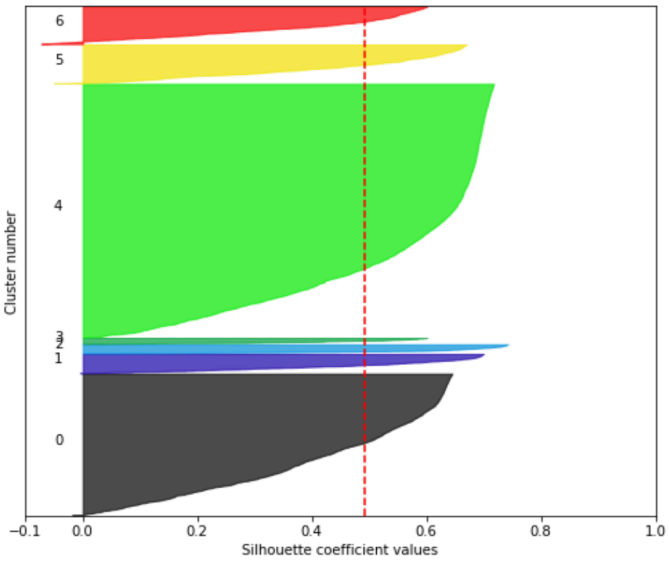

Figure B.6: 7 clusters

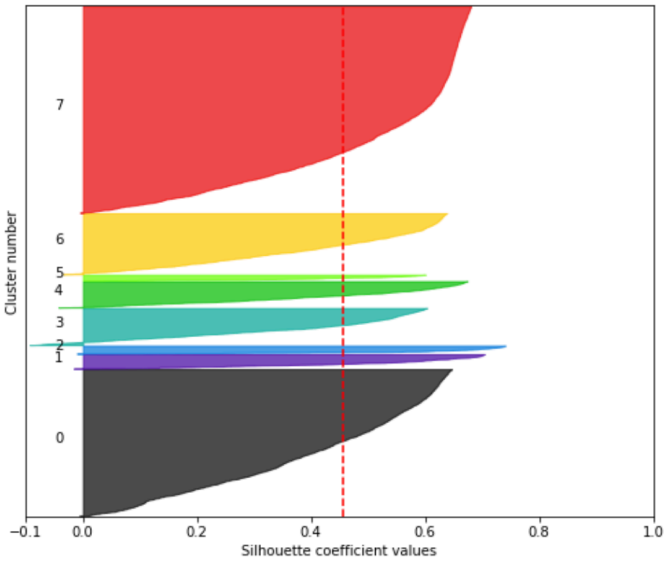

Figure B.7: 8 clusters

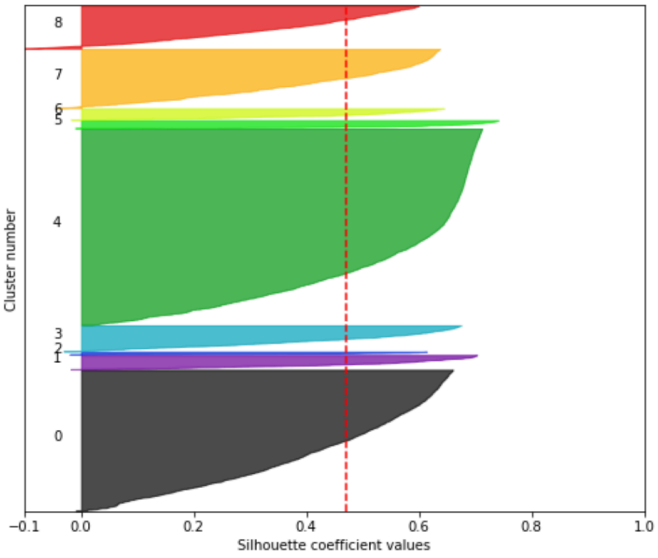

Figure B.8: 9 clusters