Medicare Data Analysis Summary Report

**Executive Summary**

This analysis seeks to highlight underlying reasons for which certain institutions charge the Medicare system greater sums of money than others, and if so, can these features be acted on in a legislative sense so as to reduce the amount of money that is charged to the Medicare system. Factors considered were a geographically standardized Medicare charge amount, provider gender, provider entity type (individual vs organization), office vs facility (such as hospital) and whether the institution/individual was a participant in the Medicare pricing system. Cluster analysis was performed to group similar institutions into groups for cross comparison.

The analysis indicates that all of the features aside from whether the institution is a participant in the Medicare system seem to have little to no effect on the price Medicare is charged. Interestingly, those that do not participate in the Medicare system charge substantially less than those that do- the mean value is two orders of magnitude lower than both the subset’s overall average the average of each of the other cluster.

Based on this analysis, I recommend a qualitative inquiry into the charging practices of facilities that do and do not participate in the Medicare system. Clearly there is a discrepancy present here, and if rectifying the difference in pricing practices could bring the average charge from member institutions down to the level of non-member institutions, then a substantial amount of money could be saved by the Medicare system.

**Problem Statement**

This analysis seeks to determine whether or not healthcare providers with certain similar characteristics routinely charge higher payments to the Medicare system. This information could be potentially beneficial to legislatures within the United States working to reduce costs incurred by the Medicare system.

**Methodology and Assumptions**

All of the data used for this analysis was obtained via the U.S. governments yearly public disclosure of Medicare information. A clustering analysis was used in determining if certain similarly characterized healthcare providers are responsible for incurring larger charges to the Medicare system; the data set included many factors, however for the purposes of this analysis, we chose to include the following:

* ***nppes\_provider\_gender*** – Categorical. The gender of the provider, if the provider was an individual.
* ***nppes\_entity\_code*** – Categorical. Whether the provider was an individual or an organization.
* ***medicare\_participation\_indicator*** – Categorical. Whether the provider participates in the Medicare program or not.
* ***place\_of\_service*** – Categorical. Whether the place of service in this case is a facility or a non-facility (which is typically an office).
* ***average\_medicare\_standardized\_amount*** – Numeric. The average amount charged to Medicare for the service, standardized for geographical factors. This was used over the other non-standardized factors to eliminate the need to consider the effect of different price levels across states and cities.

Additionally, only data points from the United States were considered in this analysis, as the root question to be answered here would likely be useful in a legislative context. As such, it makes sense to only consider the cases which the United States would be able to legislate over.

**Analysis and Results**

Before performing the clustering analysis, a random subset of the data was taken so as to make processing faster and possible within the limitations of my own computer’s hardware. To subset the data, first a large portion of the entire data set was loaded into R (in this case, 100,000 rows) then a random sample of 10,000 rows was taken. Following the subset, the data was then filtered, and the categorical variables re-represented using one-hot encoding so that they could be included in the cluster analysis. Before cluster analysis occurred, all variables were standardized with a mean of 0 and a standard deviation of 1. For this analysis, k-means was the clustering method employed.

For the problem at hand, there is no intuitive number of clusters to choose, and so a skree plot was employed to determining the number of clusters to use:

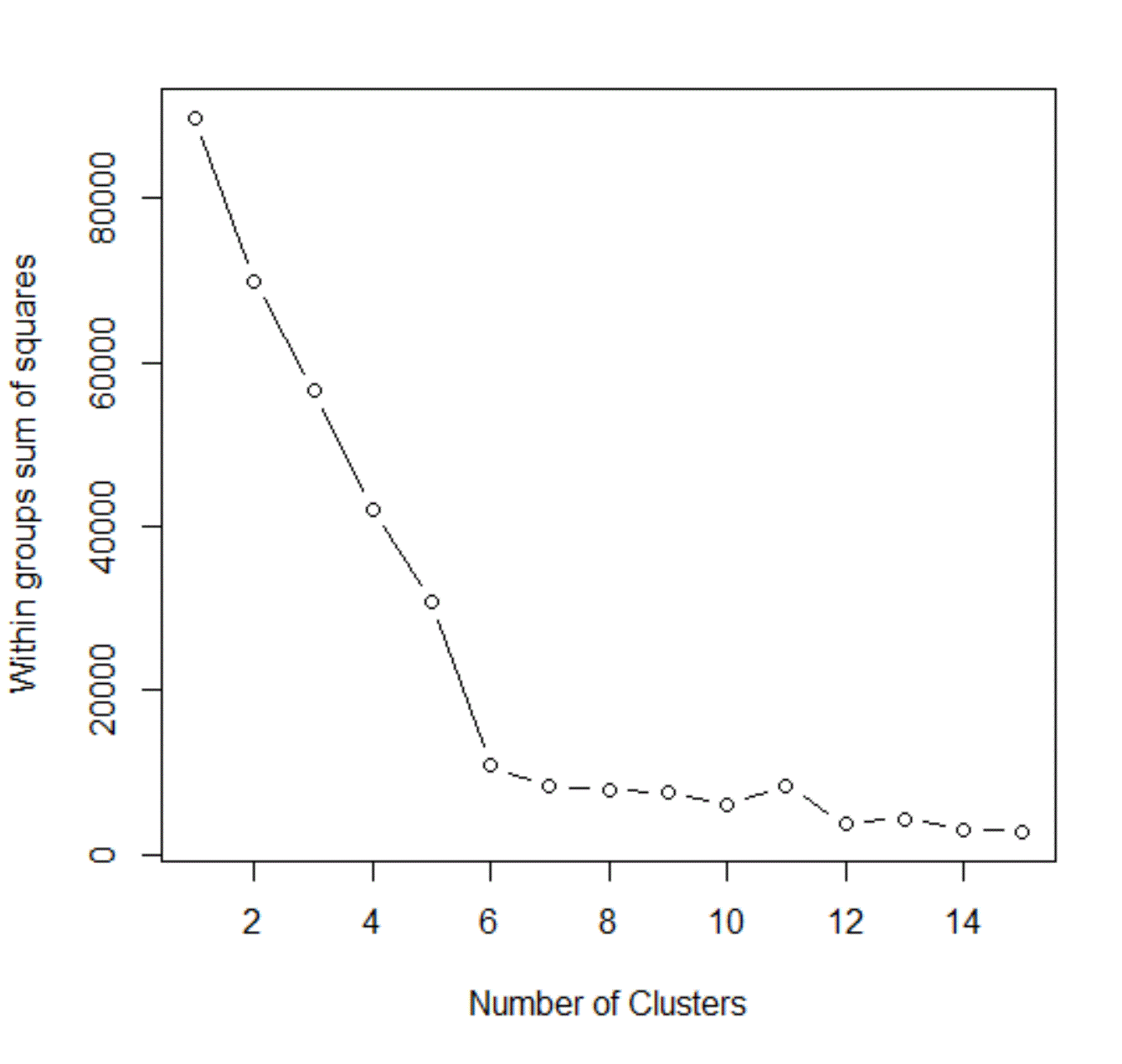


Figure 1: A scree plot showing how within group sum of square decreases as clusters increase. Note the kink at 6.

From the skree plot, it was determined that the number of clusters to use for the analysis would be 6. After that, R’s built in *kmeans* function was used to group the data points into clusters. Projected onto a two-dimensional space, the graph of the clusters formed is not particularly informative, however it does show that the data has broken down into very stark clusters:

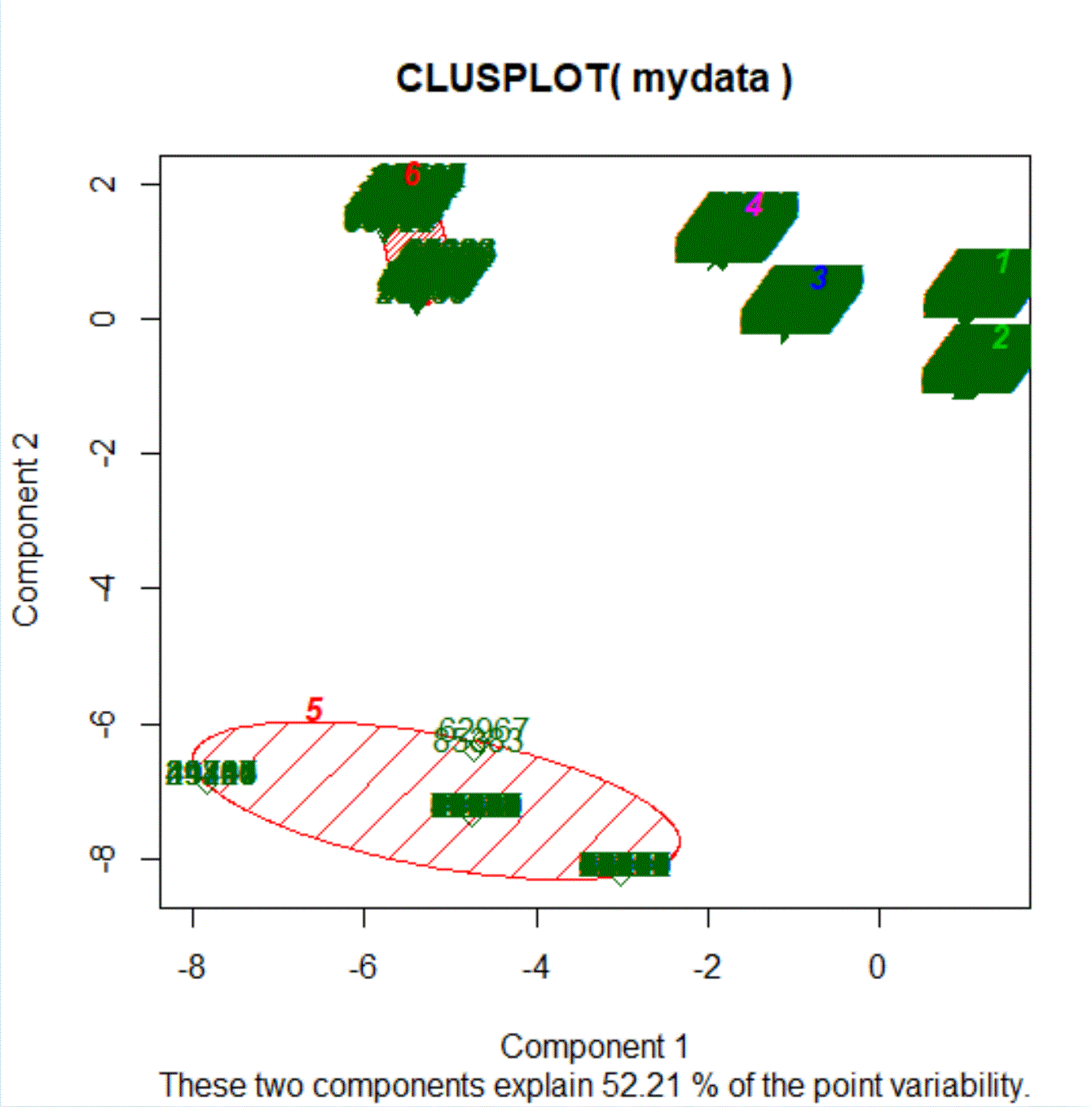


Figure 2: A two-dimensional projection of the clustering solution

Each of the six clusters formed can be broken down as having the following attributes:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Size** | **Mean Average Medicare Standardized Amount** | **Standard Dev. Average Medicare Standardized Amount** | **Percentage Male (vs female)** | **Percentage Office (vs facility)** | **Percentage Medicare Participant** | **Percentage Individual (vs organization)** |
| **Entire Subset** | 10000 | $34,520.86 | $21,419.56 | 65% | 58% | 97% | 96% |
| **1** | 3548 | $33,271.89 | $20,737.28 | 100% | 100% | 100% | 100% |
| **2** | 2761 | $39,012.00 | $21,170.88 | 100% | 0% | 100% | 100% |
| **3** | 1027 | $39,666.31 | $21080.77 | 0% | 0% | 100% | 100% |
| **4** | 2041 | $33,274.51 | $20,564.22 | 0% | 100% | 100% | 100% |
| **5** | 256 | $246.13 | $2,849.58 | 65% | <1% | 0% | 94% |
| **6** | 342 | $28,863.12 | $16217.01 | 0% | 74% | 100% | 0% |

Figure 3: A table describing the features of the different clusters, with the average amounts across the entire subset shown for comparison

Coming from a perspective of cost, the cluster that stands out the most here is cluster 5. Interestingly, it by far has the lowest average cost, however it also has the unique facet of being the cluster which includes entirely all of the non-Medicare participating facilities. This huge discrepancy in price coinciding with lack of Medicare participation perfectly points to this possibly being a causal factor in the price that the Medicare system is charged.

**Conclusion and Next Steps**

With my limited knowledge of the healthcare system, it is my advice that this information be used to look into Medicare participating individuals and institutions knowingly overcharging the Medicare system. Reasons for this could include superior bargaining power by private healthcare insurance companies to pay lower rates, or a sort of “guarantee” that the Medicare charge will be paid due to its backing by the government, however it should be noted that this is purely conjecture. From the analysis, it certainly appears that institutions that are not participating in the Medicare system are charging the system far less when they do use it, and I would recommend a more qualitative investigations take place into the charging practices of participating and non-participating institutions.