# Medicare Data Cluster Analysis

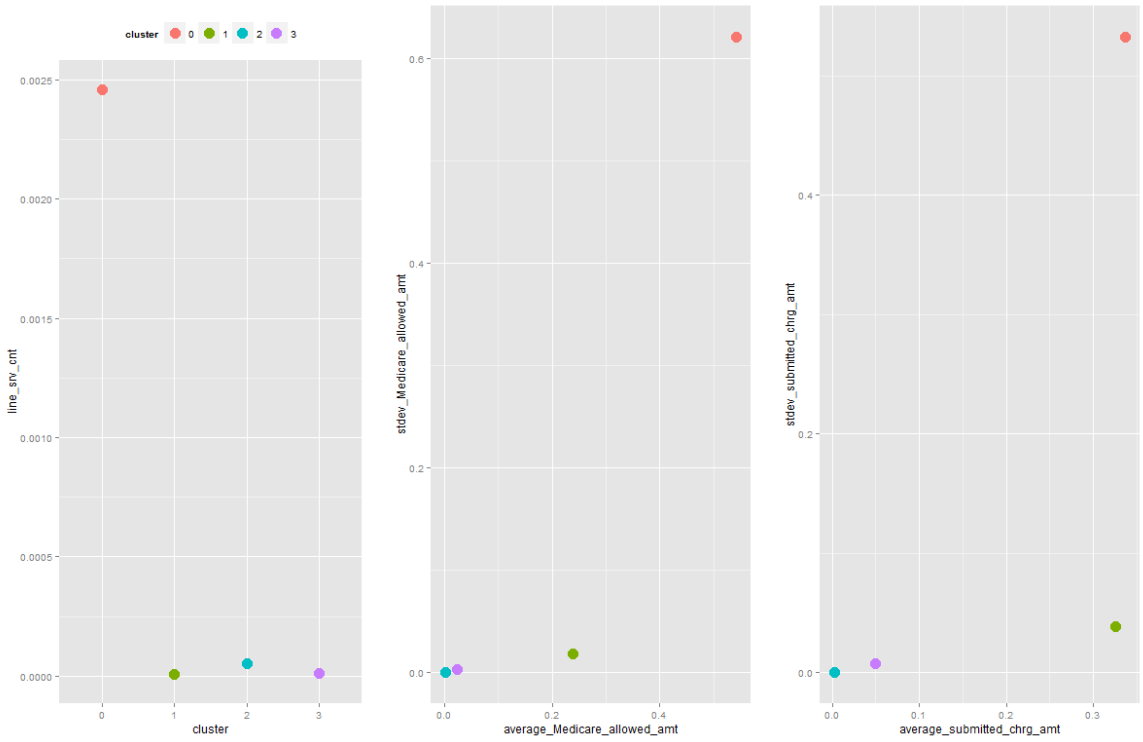
Apurvaa Subramaniam

Clustering on the variables line_srv_cnt, average_Medicare_allowed_amt, stdev_Medicare_allowed_amt, average_submitted_chrg_amt, stdev_submitted_chrg_amt, average_Medicare_payment_amt, and stdev_Medicare_payment_amt gives 4 clusters:

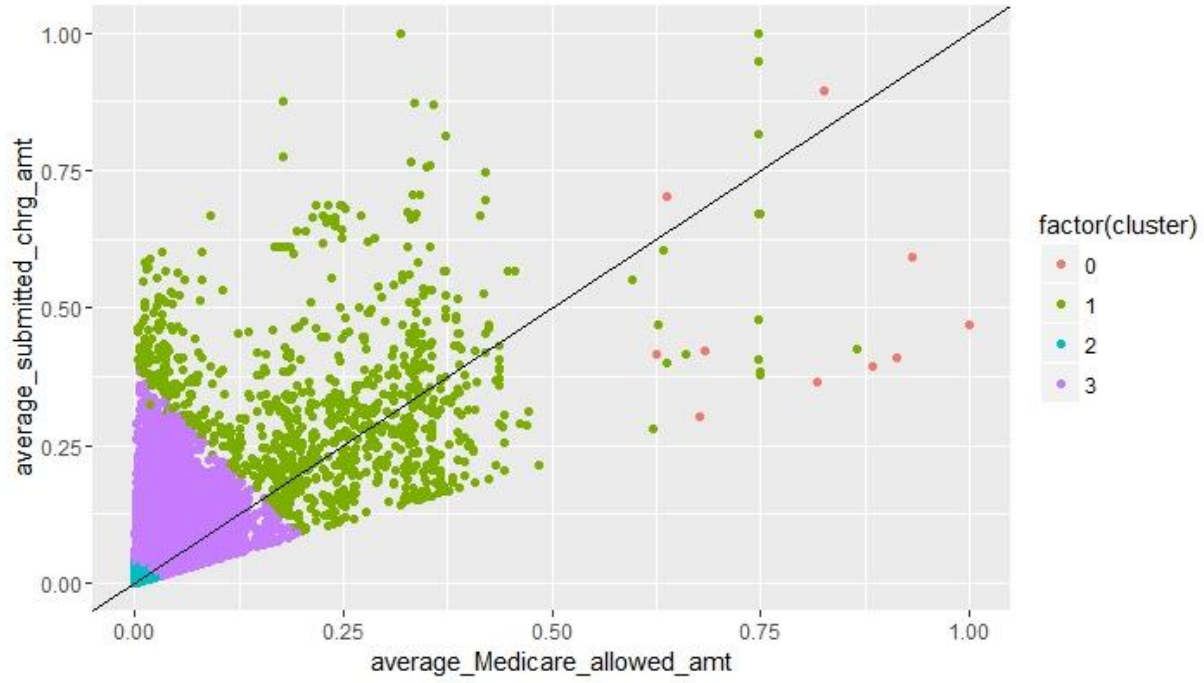
| Cluster | Count of Assigned Rows | Description (based on analysis below) |
|---|---|---|
| **0** | 15 | Outliers (Probably very large/super specialized hospitals) |
| **1** | 1024 | Underestimated Services (or trying to trick system) |
| **2** | 9,14,3478 | Average |
| **3** | 143,359 | Allowable amounts slightly higher than average (probably more specialized services) |

These variables were chosen because just clustering on the average/stdev values results in only 2 clusters i.e. not much variance in the data. Also, the average payment amount is in general the same as the allowed amount i.e. highly correlated. Adding the services count variable increases the variation in the data.

Plotting the centroids for each of these clusters for the various features shows that cluster 0 contains the outliers (with very high standard deviations as well as averages). These are probably very large hospitals. Clusters 2 and 3 are similar and represent the majority, with cluster 3 having slightly higher average values. Cluster 1 has those who submitted amounts significantly higher than their allowable limit i.e. perhaps the service cost is underestimated (or alternatively they could be trying to trick the system). Further analysis can be done by looking at the other categorical data but that is outside the scope of this analysis.




The graph below shows an overall view of utilization i.e. value submitted/value allowed, with the line representing utilization of 1 or 100%. This shows that for cluster 3 utilization is >100% for the majority i.e. they submit more than their allowed amount, which indicates that these services are likely to be for terminal/difficult to treat illnesses (such as cancer) or recurring treatments such as physiotherapy, which are relatively more expensive than common ailments. It is interesting that most of the outliers have utilization much lower than 100%. It would be worthwhile to take a further look.




Zooming in on the outliers shows that apart from super specialized services (represented by high charges and low standard deviations below, and low number of services but very high charges), some services have very high standard deviations but low charges. These are likely to be emergency services which are highly variable in nature.