Analyses of Medicare Data

By: Swap Chhabra, Alex Milut, Kyle Hundman

We employed K-means clustering in Hadoop for the following dimensions from the file, Medicare-Physician-and-Other-Supplier-PUF-CY2012.txt*:

1. bene_day_srvc_cnt
2. average_Medicare_allowed_amt
3. stdev_Medicare_allowed_amt
4. average_submitted_chrg_amt
5. stdev_submitted_chrg_amt
6. average_Medicare_payment_amt
7. stdev_Medicare_payment_amt

*Each variable was log normalized and scaled such that the final values ranged from 0~1.

Variable Selection:

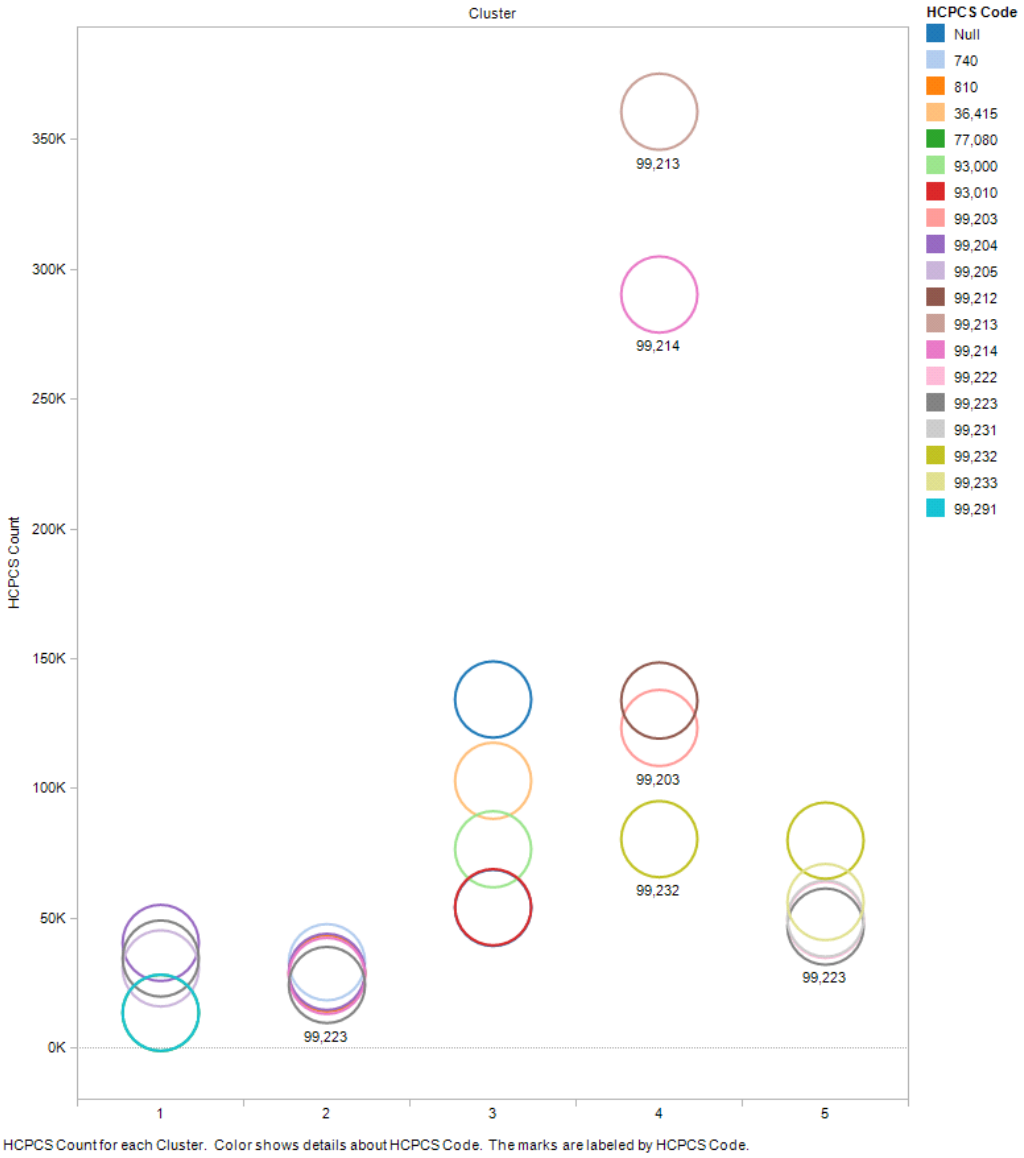
These variables were selected because we wanted to see how average charge, allowance, and payment converged and diverged across clusters. One count variable was also included that would provide insight into the magnitude of services that were being represented by each cluster. Additional counts were left out to avoid inflating the importance of related counts in the k-means algorithm.

Cluster Analysis:

5 iterations of the clustering algorithm were generated to arrive at a **5 cluster solution** detailed below. The centroid means (re-transformed back from their scaled values) for the resulting clusters are shown below with extreme values highlighted in green or orange. In particular note the stark differences between Cluster 1 and Cluster 3. Cluster 3 appears to the lowest values for all the numeric variables considered in this study by a large margin.

	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5
bene_day_srvc_cnt	36	43	60	85	30
average_Medicare_allowed_amt	177.33	133.24	9.83	62.98	83.43
stdev_Medicare_allowed_amt	5.65	7.81	0.15	0.30	0.07
average_submitted_chrg_amt	498.45	431.55	27.74	126.90	201.07
stdev_submitted_chrg_amt	0.10	72.91	0.40	0.80	0.73
average_Medicare_payment_amt	135.94	102.72	8.48	45.85	68.42
stdev_Medicare_payment_amt	24.61	18.18	0.78	10.79	0.13

Next we examined state-wise distribution as well as prominent HCPCS codes for each of the clusters. For all of the clusters prominent states in terms of overall counts included: CA, FL, TX, IL and PA. The HCPCS code distribution however was unique for each cluster. The plot below shows the top five HCPCS codes that were observed for each of the clusters.



HCPCS Count for each Cluster. Color shows details about HCPCS Code. The marks are labeled by HCPCS Code.

As seen in the table above Cluster 3 appeared to stand out from the rest. The HCFCS codes unique to this cluster included those associated with the Influenza Virus (G0008 and Q2038 - shown as blue circles and annotated as NULL in the plot) and electrocardiogram (routine ECG Codes 93000 and 93010). Further analyses are required to understand the practical significance of the generated clusters. Apart from the inherent limitations of K-means method, the current study was limited to five iterations of the K-means algorithm. Hence these observations should be taken with a grain of salt. However the current analyses demonstrates that numerical attributes from medicare data may be used to glean insights into the differences in medicare claims across the country.