Homework Assignment 3: Medicare Data Analysis

Steven Lin, Ahsan Rehman

Methodology

The variables were standardized (using Pig) to put equal weight for the k-means and have commensurate units (so that variables with greater variation are not given more importance during distance computation and cluster assignment). The numeric variables in fields 18 to 26 were chosen for clustering. The reasoning was that each of the amounts, counts and standard deviation are related to behavior of providers and will give more interesting results than just clustering on demographic variables. All of the amounts, counts and standard deviation were chosen because we might expect to have differences in each of these variables between clusters since with a behavior standpoint it makes sense for a provider not to behave the same levels across clusters.

The k-Means algorithm implemented in mapreduce (using python scripts) was run for cluster size from 3 to 6, with 3 random initial centroids and 10 iterations for each.Then the best cluster size was chosen and k-means was run again for that cluster size. Another mapreduce job was run to assign each observation to a cluster from the final centroids of the k-means. Finally, a sample dataset of 100,000 records was taken to profile the clusters on other variables. The *ReadMe.txt* contains a description of each file and steps to replicate the analysis.

Cluster Size

In order to compare the different cluster sizes, the cluster mean variables were first unstandardized. The numeric outputs can be seen in Appendix B for the different cluster sizes and initial centroids. It can be observed that general pattern of the centroid means does not vary across different initial centroids for the same cluster size. Thus, for graphical purposes, only the cluster means of the first initial centroids of each cluster size was plotted (Appendix C). From this analysis, it can be seen that cluster size of 6 is not a good choice because cluster 5 and 6 are similar in terms of their cluster means, indicating that it might have come from a split of a cluster from cluster size of 5, while other cluster sizes were distinct clusters based on the cluster means. Therefore, cluster size equal to 5 was chosen because it captures the most differentiation across its clusters until no significant differences can be seen, such as the cluster in cluster size equal to 6.

Cluster Description

Based on the mean centroids, the following clusters could be identified:

Cluster	Description
1	Specialist Providers. They have very high submitted charges and relatively low count.
2	Day-to-Day providers. They have low charges, low variance and high count
3	Emergency/Trauma providers. They have higher variance and charges, and lower count compared to cluster 2 due to different type of patient
4	New patient providers or Visiting specialist. Similar to cluster 2 but lower count
5	Specialist providers. Compared to cluster 1, they have lower submitted charges and slightly higher patient count.

Profiling

The map (Appendix D) shows a similar distribution across the US for all clusters. Cluster 1 and 5 seem to be less concentrated in the middle of the country and more focused in in the east and west coast.

Profiling by place of service and provider type (Appendix E) shows different characteristics for each cluster. For example, cluster 2 and 4 are similar, but cluster 2 has a higher number of patient visits related to internal medicine, family practice and diagnostic radiology.

Our K-Means solution produced five clusters which are distinct in terms of demographics and identify distinct type of health beneficiaries based on the HCPCS Description (Appendix F). Cluster 1 describes major patients with complex Arthroplasty and Orthopedic procedures which are usually treated by specialist doctors. This cluster has a very high submitted charge amount which largely varies from patient to patient, as each procedure demands specialized treatment. Cluster 2 describes major patients visiting for Influenza, chest x-rays or ECG tests. These patients are frequently visiting the medical provider and have specific doctors with lower submitted charge amount. Cluster 3 describes patients visiting under emergency, critical care or initial treatment, which explains these trauma patients at times could be having medium (Average $800) submitted charges which could vary at times based on their treatments. Cluster 4 describes a large chunk of patients who visit the hospital for the first time and require basic treatment, these patients might be visiting the specialist doctors for recommendations and therefore the charged amount for this cluster could be a little higher than normal patients. Cluster 5 describes patients with complex surgeries which include 'intracoronary stent' or 'Anesthesia spine cord surgery', such procedure can also be having higher submitted charge amount and can largely vary based on the level of complex treatment. Our analysis suggests that numerical and categorical attributes from Medicare data can be used to glean insights into the differences in Medicare claims across the country.

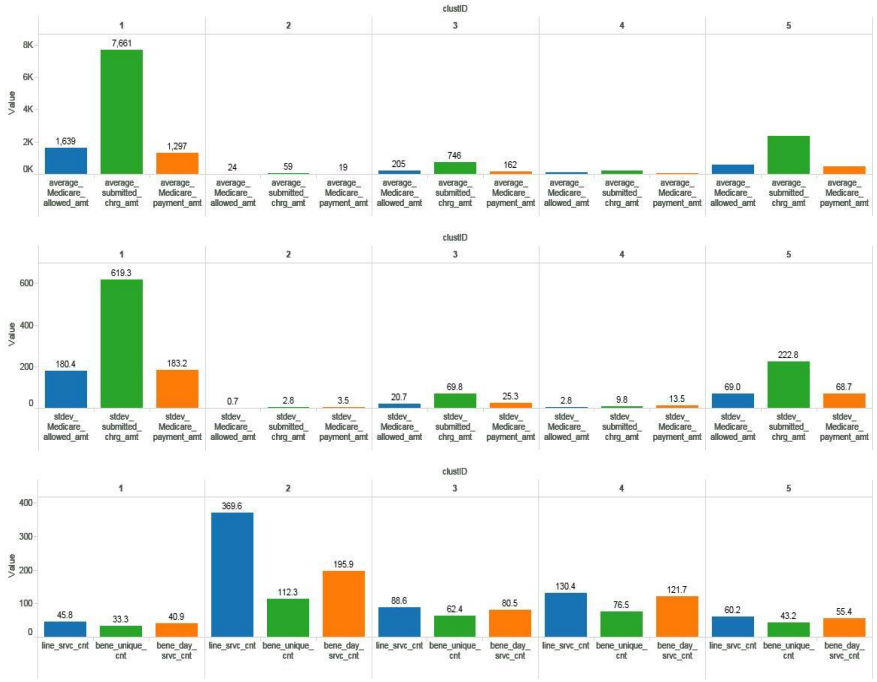
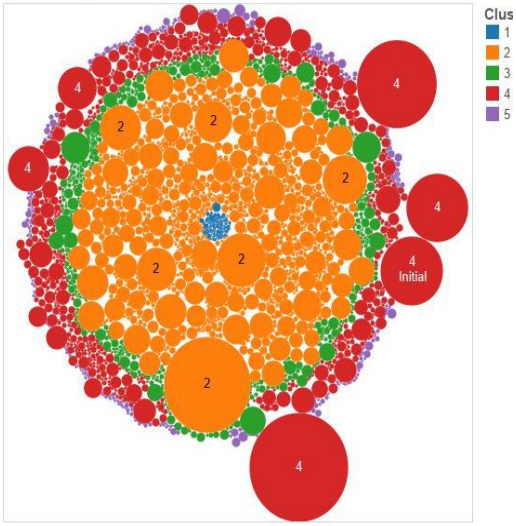


Figure 1 Cluster Analysis



Cluster ID and Hcpcs Description. Color shows details about Cluster ID. Size shows count of Cluster ID. The marks are labeled by Cluster ID and Hcpcs Description.

Figure 2 HCPCS Description Analysis for Clusters

Appendix

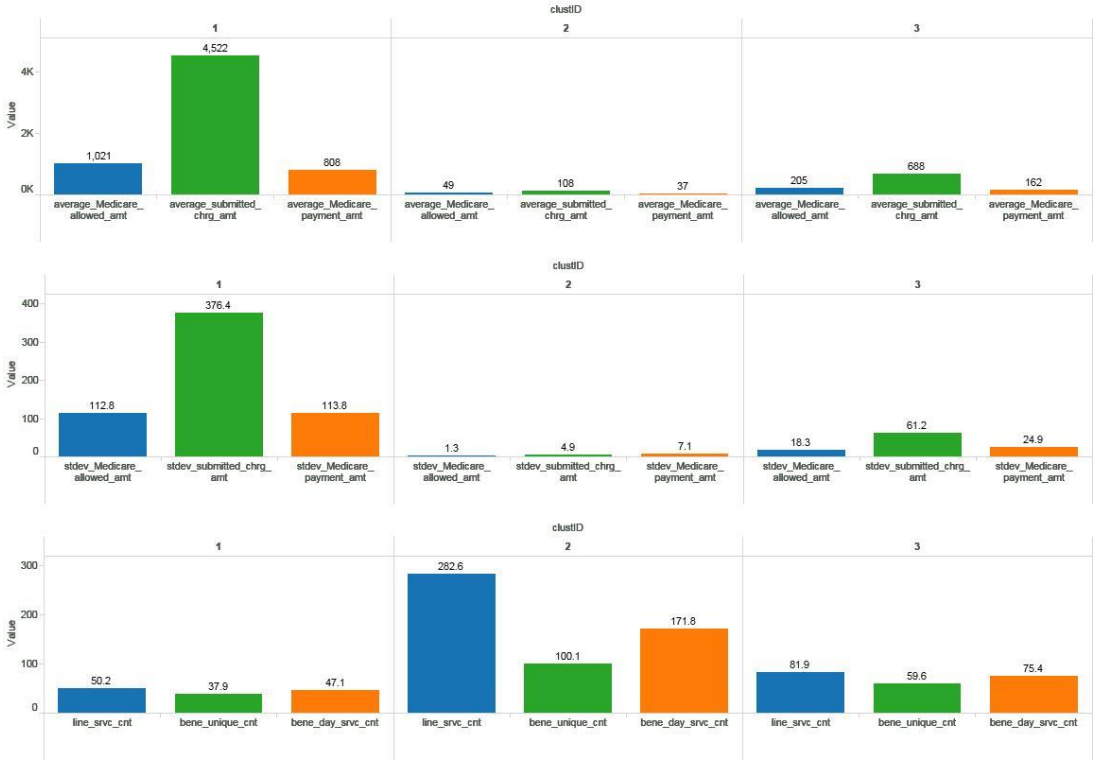
A. Dataset Overview

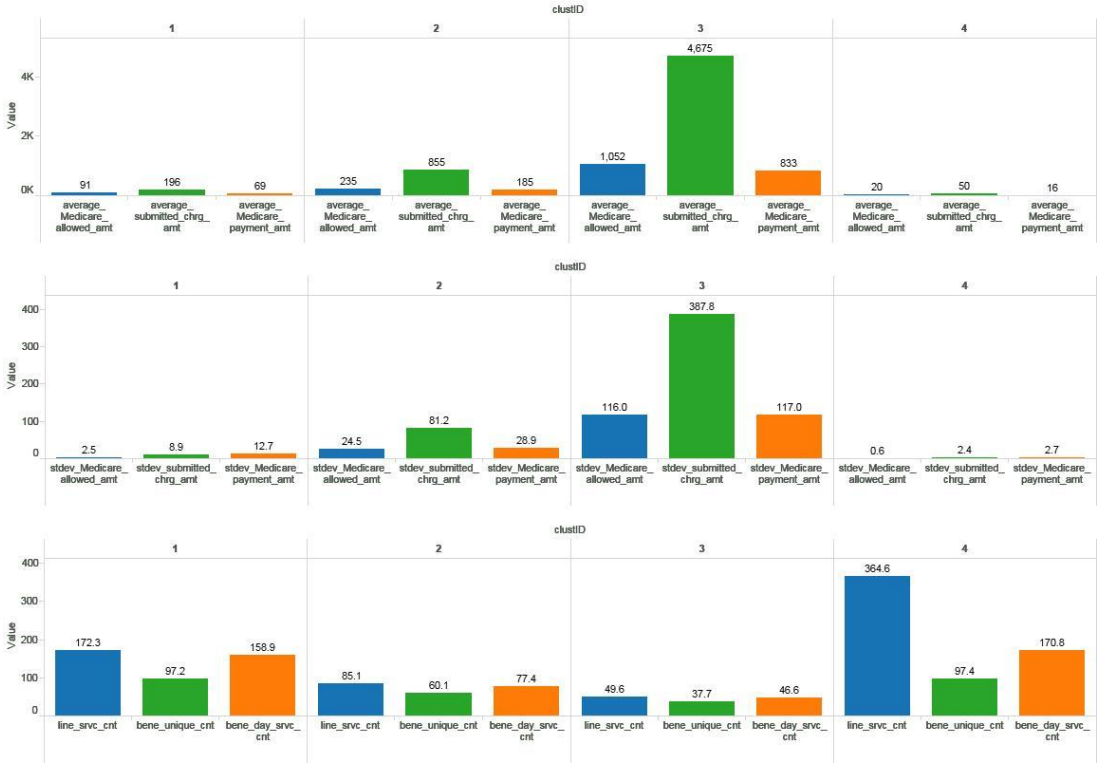
This public dataset 'Physician and Other Supplier PUF' has been provided by the Centers for Medicare & Medicaid Services (CMS). This dataset contains information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. The dataset covers calendar year 2012 and has around 10 million records. The goal of this assignment was to run K-means to achieve meaningful clusters in the dataset and derive insights on any particular patterns.

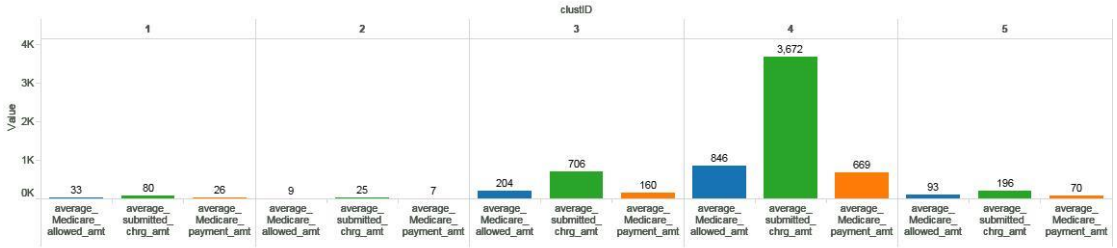
B.Output K-means

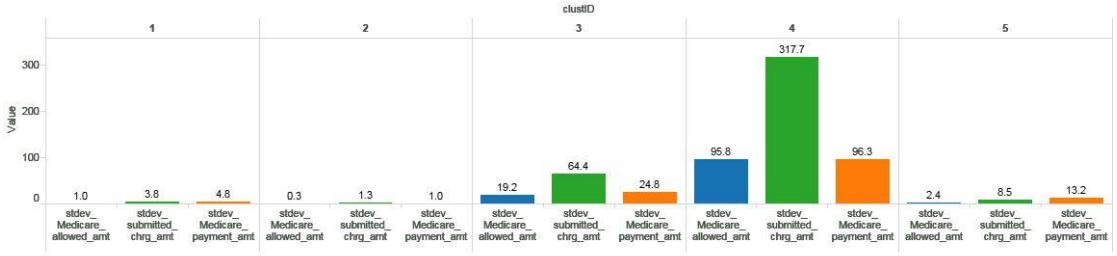
clustl	line_srv	bene_unique_cnt	bene_day_srvc_cnt	average_Medicare_allowed_amt	stdev_Medicare_allowed_amt	average_submitted_chrg_amt	stdev_submitted_chrg_amt	average_Medicare_payment_amt	stdev_Medicare_payment_amt
1	50.1	37.92	47.11	1020.97	112.83	4521.72	376.40	808.21	113.82
2	282.	100.14	171.79	48.88	1.32	108.21	4.85	37.29	7.08
3	81.9	59.60	75.44	205.41	18.30	688.31	61.21	161.51	24.89
1	353.	110.02	193.10	27.85	0.79	64.06	2.97	21.67	4.08
2	60.9	44.29	55.63	624.46	74.36	2666.27	243.80	493.93	74.52
3	115.	72.55	108.09	124.04	6.31	330.15	22.03	95.56	16.03
1	45.6	34.12	41.59	1488.00	163.51	6834.88	548.56	1178.07	165.80
2	251.	94.06	157.09	66.70	2.66	165.36	9.56	51.41	8.87
3	72.8	52.40	65.98	401.26	50.08	1631.96	162.38	317.28	50.45
1	172.	97.17	158.86	90.93	2.53	196.24	8.93	68.92	12.75
2	85.1	60.08	77.43	234.87	24.48	855.29	81.20	185.07	28.89
3	49.6	37.72	46.55	1051.91	115.96	4675.34	387.78	832.74	117.02
4	364.	97.35	170.84	19.68	0.57	49.94	2.44	15.80	2.67
1	57.5	41.35	52.68	734.89	85.49	3164.52	282.09	581.45	85.66
2	102.	68.45	96.77	141.59	8.48	398.92	29.21	109.66	18.11
3	231.	80.65	141.25	33.32	0.94	75.60	3.41	25.64	4.89
4	3654	9854.51	17351.68	23.34	1.77	65.30	9.21	20.30	2.67
1	472.	129.61	229.17	14.81	0.43	38.23	1.94	12.31	1.86
2	151.	80.90	136.93	66.78	1.78	143.02	6.13	50.06	10.17
3	84.9	61.51	79.05	167.82	12.04	510.79	41.05	131.14	20.54
4	56.7	40.87	52.11	782.07	89.90	3379.39	297.50	618.83	90.20
1	121.	58.55	100.68	32.98	0.98	79.96	3.82	25.55	4.83
2	737.	189.62	330.34	8.52	0.30	24.89	1.34	7.39	1.01
3	85.2	61.05	77.97	203.83	19.16	706.34	64.41	160.35	24.76
4	54.5	39.71	50.14	845.68	95.81	3671.60	317.66	669.26	96.33
5	141.	79.98	131.27	93.12	2.44	196.09	8.54	70.48	13.17
1	7339	1720.76	3166.01	16.52	0.83	41.68	3.90	13.61	2.59
2	52.8	39.10	49.06	911.79	102.00	3982.07	339.11	721.67	102.74
3	161.	61.19	102.86	18.20	0.50	46.13	2.21	14.74	2.41
4	139.	78.08	128.01	80.10	2.13	170.90	7.42	60.25	11.80
5	81.4	59.71	75.35	197.28	17.23	655.69	57.87	155.08	23.89
1	153.	81.58	138.44	66.80	1.86	146.40	6.60	50.02	10.24
2	54.1	39.61	49.90	892.94	99.11	3839.91	318.75	706.71	100.09
3	455.	126.38	223.21	15.71	0.45	40.23	2.02	12.95	2.02
4	79.3	57.64	71.82	232.69	31.86	1014.69	117.29	183.62	33.10
5	89.0	63.70	83.76	147.84	5.04	332.04	14.13	114.81	16.51
1	367.	165.70	293.26	96.18	1.89	138.64	5.44	72.35	14.52
2	52.7	39.05	48.94	914.73	102.33	3998.48	340.46	724.00	103.06
3	85.0	60.82	77.78	216.72	20.99	751.29	69.51	170.64	26.33
4	433.	99.24	166.85	10.65	0.32	28.95	1.48	9.10	1.29
5	85.6	57.79	78.86	87.53	3.23	274.91	13.68	66.86	11.36
6	137.	62.94	112.99	35.84	1.12	88.48	4.13	27.55	5.24
1	363.	92.08	150.80	10.40	0.31	28.58	1.43	8.90	1.26
2	140.	79.47	130.42	92.61	2.43	196.38	8.61	70.06	13.13
3	53.1	39.09	49.07	898.35	100.64	3916.86	334.59	711.01	101.36
4	5003	14344.37	25279.33	23.21	1.81	67.36	9.87	20.42	2.45
5	158.	69.84	128.93	36.58	1.11	87.82	4.22	28.10	5.45
6	81.6	59.38	75.16	208.37	19.74	721.68	66.00	163.96	25.33
1	177.	88.22	159.85	65.60	1.59	113.31	4.63	48.76	10.46
2	96.3	61.44	81.74	59.53	3.50	269.50	17.79	45.74	8.13
3	85.5	59.95	77.35	245.16	28.42	946.48	94.41	193.41	31.42
4	394.	101.94	179.57	16.99	0.48	43.96	2.12	13.85	2.22
5	51.0	38.39	47.92	978.45	108.44	4306.58	361.00	774.49	109.37
6	179.	110.04	169.32	134.04	3.51	260.47	10.69	103.00	16.58

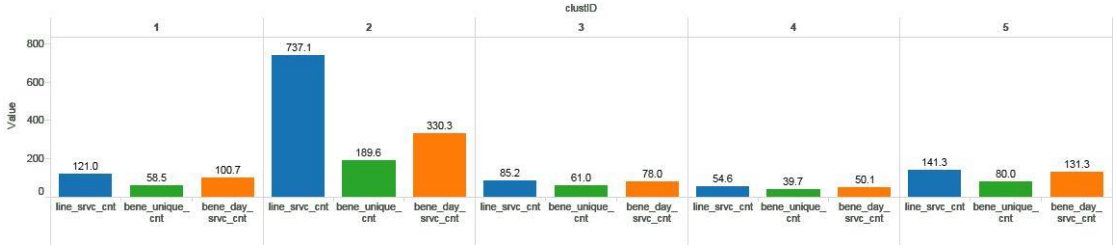
C. Graphical Output k-means

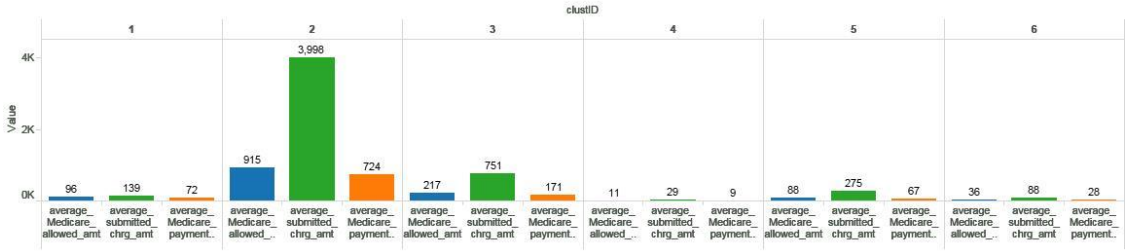


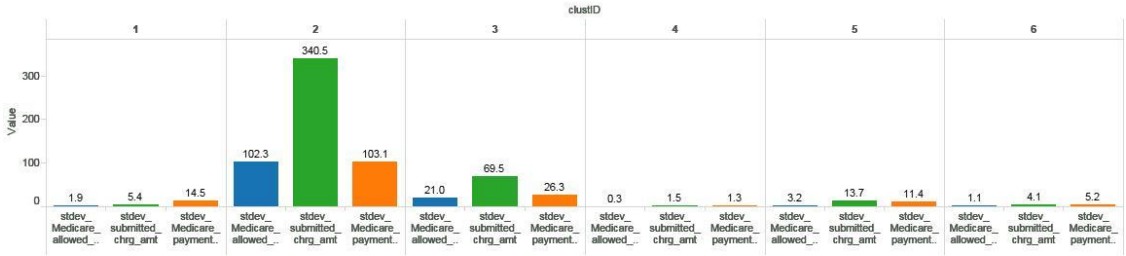


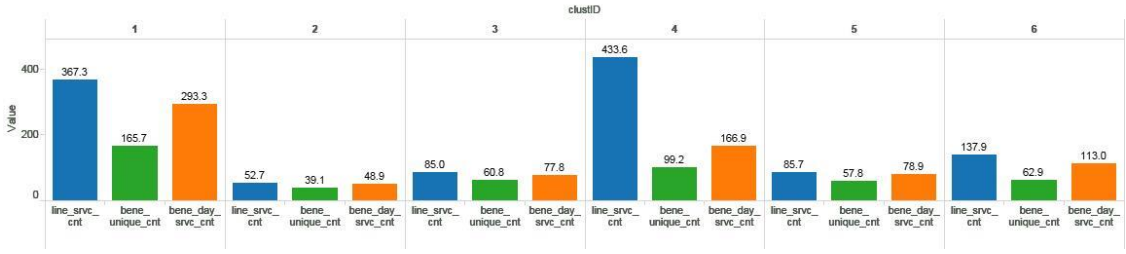

clustID
1
2
3
4
5
Value
4K
3K
2K
1K
0K
33
80
26
9
25
7
204
706
160
846
3,672
669
93
196
70
average_ Medicare_ allowed_amt
average_ submitted_ chrg_amt
average_ Medicare_ payment_amt


clustID
1
2
3
4
5
Value
300
200
100
0
1.0
3.8
4.8
0.3
1.3
1.0
19.2
64.4
24.8
95.8
317.7
96.3
2.4
8.5
13.2
stdev_ Medicare_ allowed_amt
stdev_ submitted_ chrg_amt
stdev_ Medicare_ payment_amt

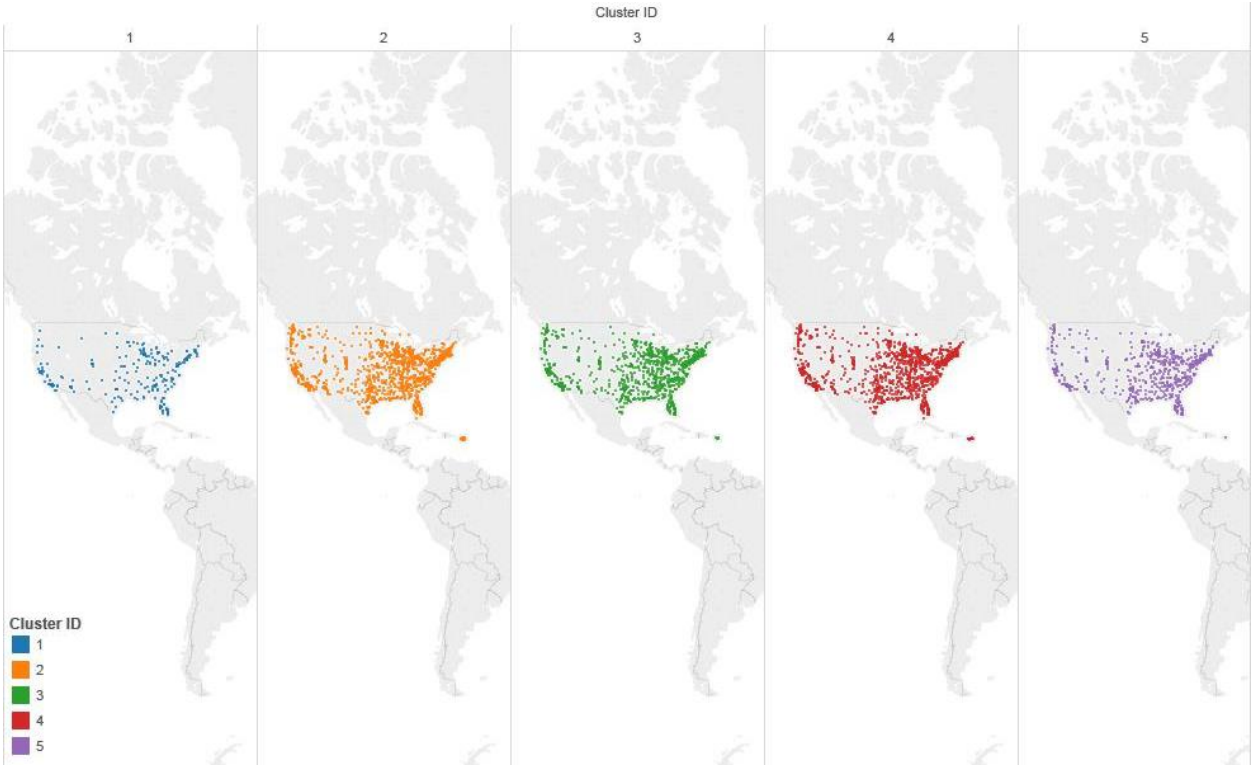

clustID
1
2
3
4
5
Value
800
600
400
200
0
121.0
58.5
100.7
737.1
189.6
330.3
85.2
61.0
78.0
54.6
39.7
50.1
141.3
80.0
131.3
line_srvc_cnt
bene_unique_ cnt
bene_day_ srvc_cnt


clustID
1
2
3
4
5
6
Value
4K
2K
0K
96
139
72
915
3,998
724
217
751
171
11
29
9
88
275
67
36
88
28
average_ Medicare_ allowed_amt
average_ submitted_ chrg_amt
average_ Medicare_ payment..

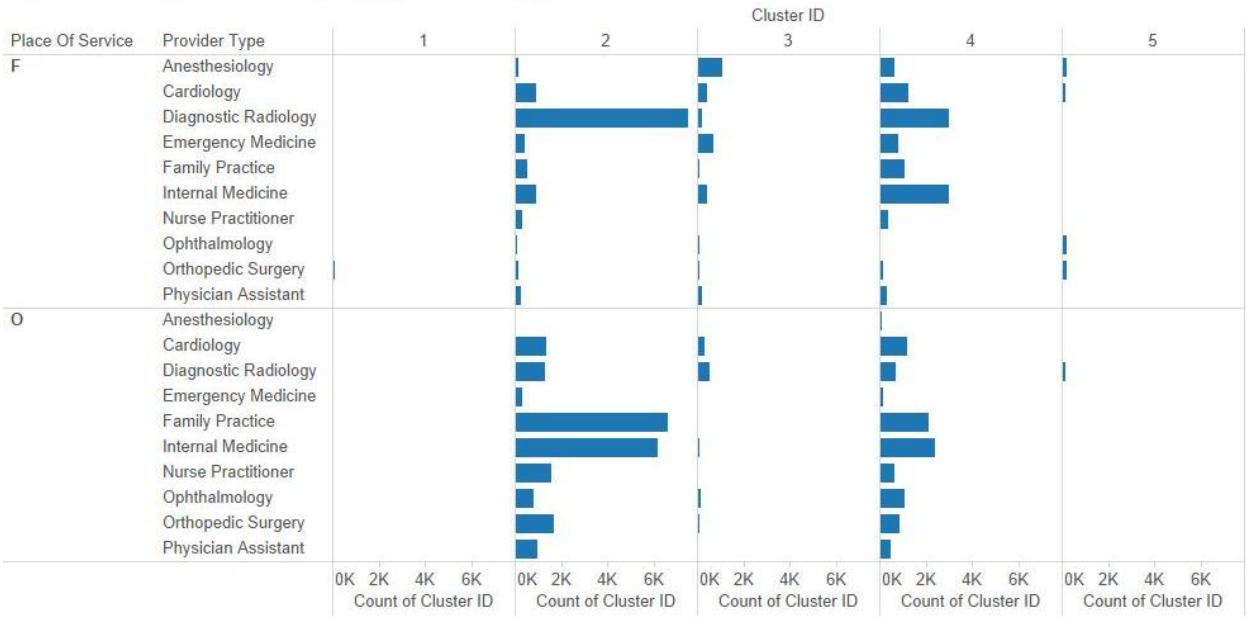

clustID
1
2
3
4
5
6
Value
300
200
100
0
1.9
5.4
14.5
102.3
340.5
103.1
21.0
69.5
26.3
0.3
1.5
1.3
3.2
13.7
11.4
1.1
4.1
5.2
stdev_ Medicare_ allowed_..
stdev_ submitted_ chrg_amt
stdev_ Medicare_ payment..


clustID
1
2
3
4
5
6
Value
400
200
0
367.3
165.7
293.3
52.7
39.1
48.9
85.0
60.8
77.8
433.6
99.2
166.9
85.7
57.8
78.9
137.9
62.9
113.0
line_srvc_ cnt
bene_ unique_cnt
bene_day_ srvc_cnt

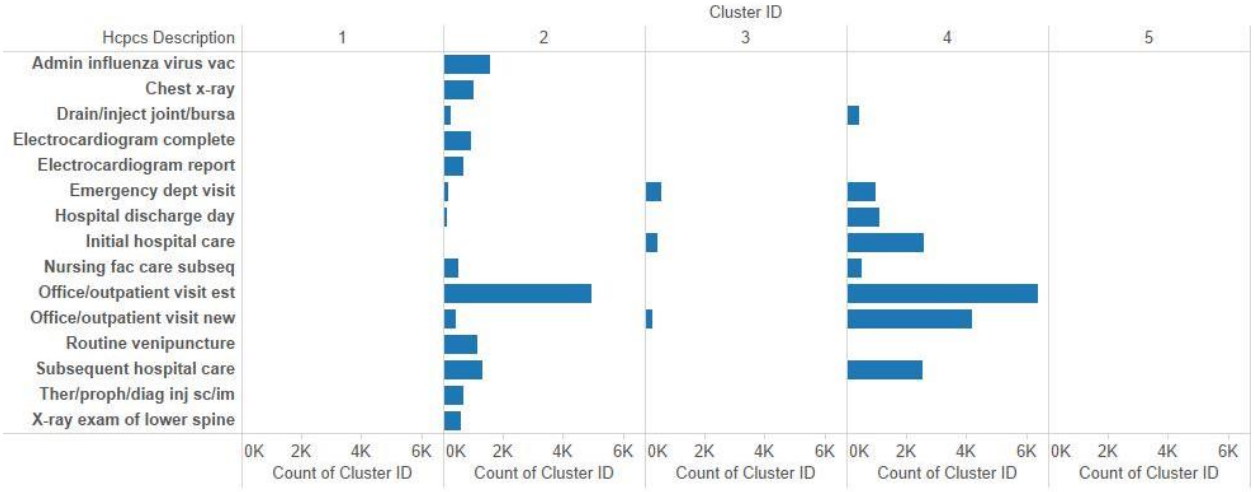
D. Map by city



E. Place of Service & Provider Type



F. HCPCS


Cluster ID
Hcpcs Description
1
2
3
4
5
Admin influenza virus vac
Chest x-ray
Drain/inject joint/bursa
Electrocardiogram complete
Electrocardiogram report
Emergency dept visit
Hospital discharge day
Initial hospital care
Nursing fac care subseq
Office/outpatient visit est
Office/outpatient visit new
Routine venipuncture
Subsequent hospital care
Ther/proph/diag inj sc/im
X-ray exam of lower spine
0K 2K 4K 6K
Count of Cluster ID