Matthew Paras
IEMS 308
1/23/2019
Clustering Recommendations

Executive Summary

An issue that is prevalent in many fields is the pay gap between men and women with the same level of experience. However, it is also possible that some services are dominated by one gender over the other, and thus this pay gap can exist based on specialization rather than any discriminatory reasons. There exists the potential for regional differences in these pay gaps or disparities in specialization. For example, one large city in the US may have many specialists of one kind, while another city may have many specialists of a different type.

Clustering is an analytical tool that groups observations with similar attributes together, and can be used to identify high level characteristics (groups) or in outlier detection. The US Government provides Medicare provider information each year that contains the demographic information for each provider (including name, gender, address, place of service), as well as the service provided, how often the service was provided, and the payment information for each service. By clustering on the standardized amount paid, the number of patients treated (number of times the service was provided), the location type, and the gender of the provider, we can attempt to identify and differences or inequalities between these groups and if there are any areas of specialization that are dominated by one gender.

For this analysis, we subset the data on the city of Chicago in order to have a more manageable data set size, and in order to examine any issues that may occur within a community. After performing systematic exploration of the data, including histograms and high level descriptives of the data, the dataset was clustered. The results show that for generic and common medical procedures, men and women are nearly identical in their payment information. However, certain areas of specialization emerged which are male dominated, in particular, knee and joint treatment and chemotherapy. This opens the avenue for women to enter these areas to take advantage of the areas of specialization. These results would be particularly insightful for new providers who are looking to pick a specialization in Chicago and could pick a field of expertise that is underutilized and pays well.

Problem Statement

Within the city of Chicago, we wish to identify the differences in services provided in locations defined as “facilities” and “non-facilities”, as well as the differences in gender and payment that exist within these groups. Ideally from this analysis, we will identify where certain services are provided and if there are any differences between the gender of the providers as well. Some services may be dominated by one specific gender and may reveal the demographics of certain services within the city of Chicago. Understanding what services are

provided, by what genders, and in what facilities may provide insight into where specialists may want to set up business.

Assumptions

- The Medicare data provided matches the documentation for the data provided, such that the data represents exactly what it should
- The data itself is accurate and that there were no fabrications made or transcription errors made in the process of creating the data.
- Demographic information provided is correct, such that medicare providers can be classified as either male or female, and that their places of work can be classified as either a facility or a non-facility.
- The description for whether a facility is a facility or a non-facility is correct

Methodology

Because the data set is too large to perform clustering on, I decided to subset the data by a geographic region, to provide some sort of insight into a specific area. For this analysis we chose the city of Chicago. To subset on providers located in Chicago, I selected the observations that have provider zip codes located in Chicago after finding a list of zip codes for the city of Chicago. We are also concerned with individuals rather than organizations, so we removed observations that were from organizations rather than individuals (since individuals cannot be classified as male or female). This also removes the issue where some medical labs that processed an extremely large number of procedures (such as blood samples), were inflating the service counts.

Then I selected the features that we wanted to cluster on, which included “line service count”, “average medicare payment”, “gender”, and “service location”. By clustering on these fields, we can attempt to segment the services by the cost and frequency of service, as well as where the service is provided and the gender of the provider. After the one-hot variable encoding is done on both “gender” and “service location”, we performed beta normalization (mean of 0, standard deviation of 1) on each column to prepare the data for clustering.

In order to identify the appropriate number of clusters, I analyzed the scree plot and also performed silhouette analysis. Silhouette Analysis was too large to run on the entire data set, so we sampled 10,000 rows from the data set and performed the silhouette analysis on that data. It yielded similar results from the scree plot, so 9 clusters were chosen. See **Figures 1 and 2, Table 1** for more information regarding the analysis run.

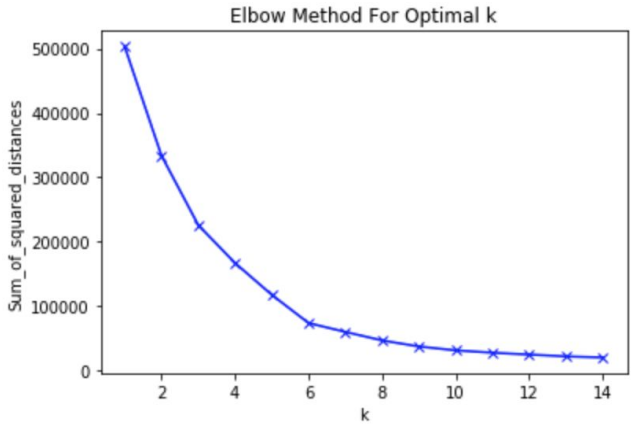

Figure 1: Scree Plot

Silhouette Stuff:

Number of Clusters	Average Silhouette Score
2	0.5014393627600267
3	0.6513966861060004
4	0.6561384036761084
5	0.7854284943056873
6	0.8032503301965597
7	0.8106385377831983
8	0.8154185508981553
9	0.8161898284585932
10	0.7415005587344172
11	0.675070328386609

Table 1: Silhouette Scores

Some example plots from the analysis:

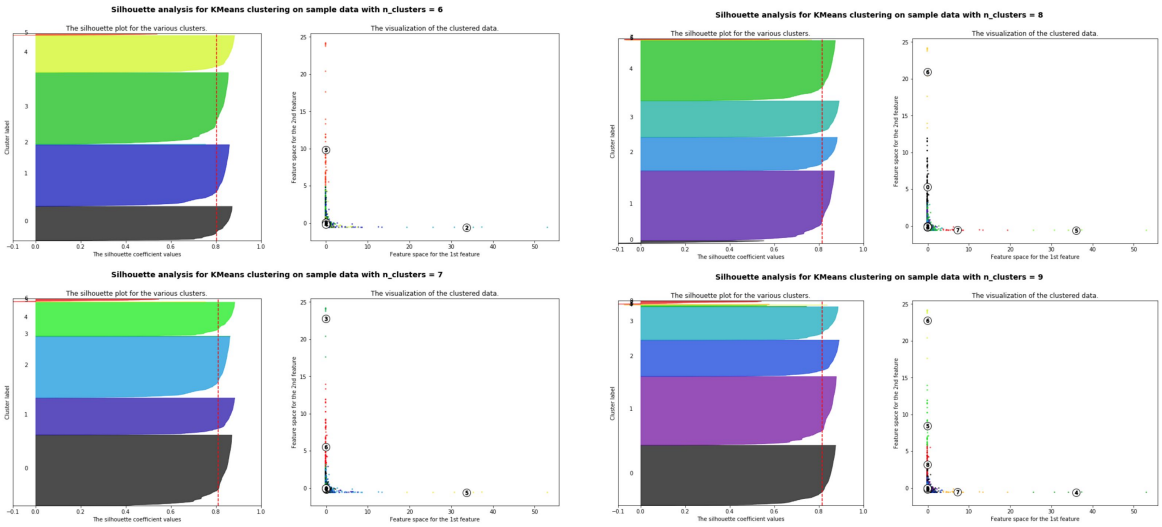

Figure 2: Silhouette Plots for n = 6, 7, 8, 9

In addition, before analysis could be done, I needed to understand what HCPCS codes were. I used this website:

https://coder.aapc.com/cpt-codes/27477

To understand what each code did and how the codes were defined.

Also, there were varying definitions for what defines a “facility” versus a “non-facility”. In order to explore this further and understand what the differences were and how different locations were defined, I read through this link:

https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/downloads/medicare-physician-and-other-supplier-puf-methodology.pdf

And found these definitions:

APPENDIX C – Place of Service Descriptions

Table C-1. Non-Facility Based Place of Service (place_of_Service ="O")

Place of Service Code	Non- Facility Place of Service Description
01	Pharmacy
03	School
04	Homeless Shelter
05	Indian Health Service Free-standing Facility
06	Indian Health Service Provider-based Facility
07	Tribal 638 Free-standing Facility
08	Tribal 638 Provider-based Facility
09	Prison/ Correctional Facility
11	Office
12	Home
13	Assisted Living Facility
14	Group Home
15	Mobile Unit
16	Temporary Lodging
17	Walk-in Retail Health Clinic
20	Urgent Care Facility
25	Birthing Center
32	Nursing Facility
33	Custodial Care Facility
49	Independent Clinic
50	Federally Qualified Health Center
54	Intermediate Care Facility/Mentally Retarded
55	Residential Substance Abuse Treatment Facility
60	Mass Immunization Center
57	Non-residential Substance Abuse Treatment Facility
62	Comprehensive Outpatient Rehabilitation Facility
65	End-Stage Renal Disease Treatment Facility
71	Public Health Clinic
72	Rural Health Clinic
81	Independent Laboratory
99	Other Place of Service

Table C-2. Facility Based Place of Service (place_of_Service ="F")

Place of Service Code	Facility Place of Service Description
21	Inpatient Hospital
22	Outpatient Hospital
23	Emergency Room – Hospital
24	Ambulatory Surgical Center
26	Military Treatment Facility
31	Skilled Nursing Facility
34	Hospice
41	Ambulance - Land
42	Ambulance – Air or Water
51	Inpatient Psychiatric Facility
52	Psychiatric Facility-Partial Hospitalization
53	Community Mental Health Center
56	Psychiatric Residential Treatment Center
61	Comprehensive Inpatient Rehabilitation Facility

FIGURE XXXX

Essentially, facilities are defined as traditional hospital locations, whereas non facility locations are defined as either offices or any non traditional locations of medical service.

In addition, to augment the methodology and results, see the Jupyter notebook entitled "Clustering (source code)" and the python script "clean_data_source_code.py".

Analysis of the Results

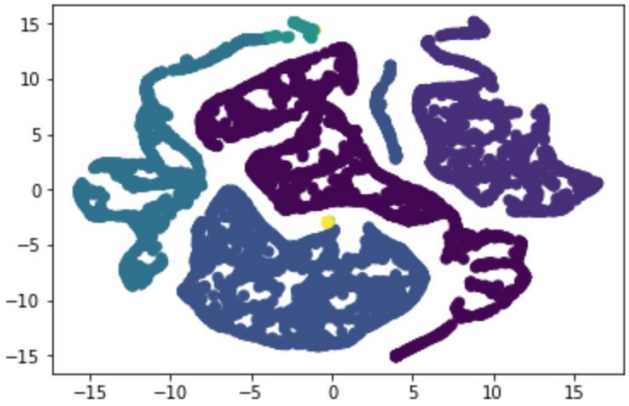

Figure 3: Visualization of the clustering using t-sne

Cluster	Cluster Size	Line Service Count Mean	Avg Medicare Standardized Amount Mean	#Unique Men	# Unique Women	Facility Count	Non-Facility Count	# Unique Services
1	28754	104.4421263	70.60975819	3588	0	3588	0	1137
2	15555	148.2343105	53.41271809	0	2917	0	2917	818
3	24492	165.6070431	60.01463595	3314	0	0	3314	1106
4	13749	95.02113608	70.72064437	0	2503	2503	0	778
5	1090	47.52293578	741.4743092	538	102	512	128	215
6	3	141664	0.510047257	1	1	0	2	3
7	20	50782.65	2.001396338	12	4	0	16	10
8	68	34.22058824	2662.465872	49	14	20	43	16
9	168	11447.53571	9.339226462	72	27	1	98	40

Table 2: Statistics about Clusters

In **TABLE 2** we can see some statistics regarding the different clusters. Since the clustering was performed on the service count, the average medicare standardized amount mean, the gender and the facility, we can see distinct groups emerge. In addition, the sizes of the groups vary wildly. Clusters 1 and 4 are associated with men and women who both work in normal facilities, and clusters 2 and 3 are associated with men and women who both work in non facilities. These means and counts are relatively similar, and after analyzing the distributions of services provided, these are fairly standard medical situations. We also see that the payment amounts are relatively similar and that the services provided are also relatively similar, so we can say that for general medical usage there appears to be no different between male and female providers. However, the interesting groups appear in clusters 5 through 9, which I outline here, after analyzing the distribution of services provided:

Clusters:

- Cluster 1:
    - Male, Facilities, standard procedures offered
- Cluster 2:
    - Female, Non facility, standard procedures offered
- Cluster 3:
    - Male, Non facility, standard procedures offered
- Cluster 4:
    - Female, Facilities, standard procedures offered
- Cluster 5:
    - Cataract surgery, coronary therapeutic services, knee and joint repair
    - Extremely expensive, male dominated, performed in facilities

- Cluster 6:
    - Injection, low iron
    - Routine injections
- Cluster 7:
    - Pre surgery routines, bone marrow injections
    - Many services provided
- Cluster 8:
    - Chemotherapy
    - Extremely expensive, male dominated, performed in non facilities
- Cluster 9
    - Pre surgery procedures, bone cancer routines, chemotherapy

The interesting clusters are seen in clusters 5 through 9. For instance, cluster 5 and cluster 8 are both primarily men, contain services that are very expensive and most definitely in certain specialties. For instance, cluster 5 contains cataract services, heart therapy services, and knee and joint repair. These are areas of specialty in which women in the future may want to get into in the area of Chicago, since there are a lack of women in this field.

Conclusions

- There appear to be very little differences between men and women with regards to general medical services within the city of Chicago, regardless of the location of service (be it a facility or a non facility)
- There are certain services that are male dominated (more so than the general male domination in the general population of doctors)
- These services are expensive and most likely highly lucrative
- There are certain fields that may be inviting for women to join, since they are predominantly male skewed
- These provide opportunities for entrance into certain fields

Next Steps

The next steps would be to analyze other metropolitan areas of the United States to see if these results are standard. Performing the same analysis on cities such as New York City, Los Angeles, San Francisco, Atlanta, etc. would yield interesting results. We could revisit the features that we selected to cluster on, and maybe clustering on others would yield more interesting results. We may also want to seek supplementary data sets, as maybe the Medicare provider data does not have the most pertinent information. The Medicare data does not include anything about private insurance or the services that these providers may be providing exclusively for private insurance beneficiaries. We could also do analysis in more rural areas to see if there are pay discrepancies in those regions as well.