Assignment 1

Zijie Huang

January 26, 2020

Introduction

In the United States of American, physicians with different credentials are providing millions of services to Medicare beneficiaries. Through studying the patterns between healthcare professionals with different credentials, the Medicare system can be optimized specifically, and thence provides a better experience to beneficiaries.

This report studies the Medicare data released by the US government and finds particular patterns among different providers. Patterns of credentials are related to service counts, beneficiary counts, provider charges as well as gender, place of services and other information of the provider.

The first part introduced the methodology used in this report. The second part displays the results of clustering. The third part is discussions of the results. And the report ends with a conclusion.

1. Methodology

This report uses the *Medicare Physician and Other Supplier Data CY 2017(Hereinafter referred to as MP)*, which is released by the US Government and can be obtained on the website: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.

Due to the relatively large size of the dataset, 10,000 samples are randomly selected and analyzed from 200,000 samples every time, while the same analysis has been repeated 5 times in different data intervals. All the figures shown in this article are based on one analysis (10,000 samples), while discussions and conclusions are based on all five analyses.

1.1 Dataset

MP is composed of 26 features, 19 of which are categorical, and the other 7 features are numerical. Among the 19 categorical features, 7 features are related to this study; and among the 7 numerical features, 5 are related to this study. The next section will explain these features in detail and explain why they were chosen.

1.2 Feature Explanation and Preprocessing

1.2.1 Numerical Features

- line_srvc_cnt : Number of services provided
- bene_unique_cnt: Number of distinct Medicare beneficiaries receiving service.

- bene_day_srvc_cnt: Number of distinct Medicare beneficiary/per day services.
- average_Medicare_allowed_amt: Average of the Medicare allowed amount for the service
- average_submitted_chrg_amt: Average of the charges that the provider submitted for the service.

The histogram of these raw numerical features are extremely skewed (non-symmetric) left. Therefore, box-cox method is used to convert the original skewed distribution into a normalized distributed histogram (An example is shown in Fig 1.1).

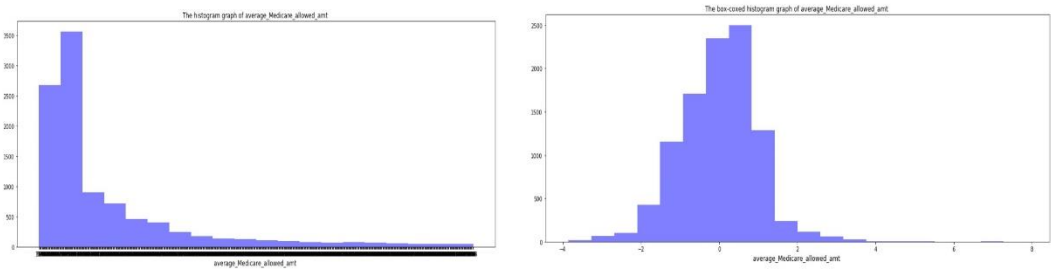

Fig 1.1 The original histogram and box-coxed histogram of feature average_Medicare_allowed_amt

Outliers which are outside of range [mean – 2 * std, mean + 2 * std] are sifted out and saved in file *Outliers.csv.* Obviously, these outliers are not erroneous since most of them contain an extremely large numbers of services provided or an extremely high value of average charges due to the specific treatments like “Injection beneath the skin” or “Anesthesia for placement or revision of blood flow shunt”. Due to the sensitivity of the clustering algorithm to outlier, these special treatment samples will not be contained in the training dataset.

The reason why feature “average_Medicare_payment_amt” and feature “average_Medicare_standardized_amt” are not chosen is that they are high linear correlated to feature “average_Medicare_allowed_amt”. The scatter figure of these three features are as followed (Fig 1.2):

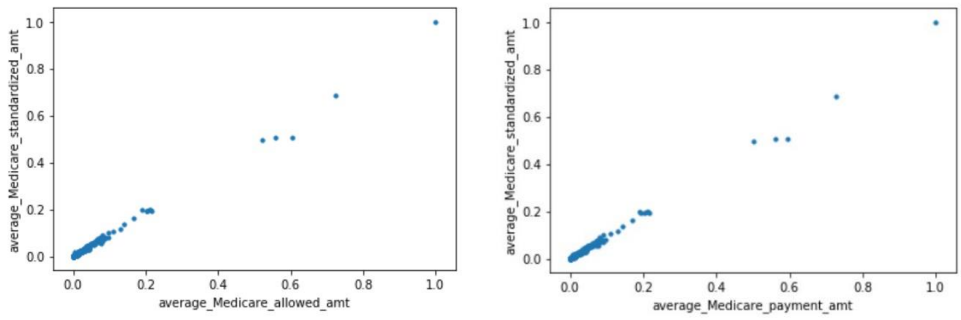

Fig 1.2 Scatter figure of three linear correlated features

After applying standardization, ten figures formed by pairing the five features are shown below. (Fig 1.3)

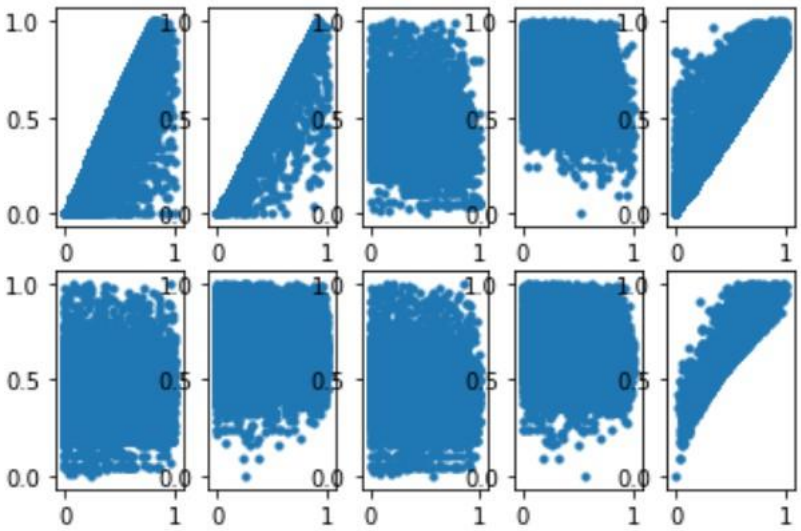

Fig 1.3 Standardized Scatter figures by pairing five numerical features

1.2.2 Categorical Features

- nppes_credentials : These are originally the credentials for an individual and blank for an organization. For the sake of easy identification, blank value is relabeled as “Organization”, and any class less than 0.5% of the whole dataset is classified as “Other”. In addition, credentials like “M.D.”, “M.D”, “MD” that are similar with each other are merged.
- nppes_provider_gender: Gender of provider. Blank values are relabeled as “O” for organization.
- nppes_entity_code: Type of entity reported in NPPES. “I” for an individual and “O” for an organization.
- nppes_provider_state: The state where the provider is located. Any class less than 0.5% of the whole dataset is classified as “O” for “Other”
- provider_type: Name of the treatment that is provided.
- place_of_service: “F” for facility and “O” for non-facility.
- hcpcs_drug_indicator: “Y” for on the Drug Average Sales Price (ASP) File and “N” for not.

While preprocessing categorical features, any class less than 0.5% of the size of the training data is relabeled as “O” for “Others”. It is noticeable that feature “country” and “zip code” are not included in the training data. For feature country, more than 99.9% of the data in this feature is labeled as “US”, which is almost meaningless for clustering. For feature zip code, there are too many classes in this feature, and if only the first three digits are considered, it has the same meaning as feature “state”. Therefore, these two features are not selected into the training data. For other categorical features such as name, ID, etc., obviously they are not very useful for clustering algorithm.

In addition, all the selected categorical data are binary encoded and then standardized before training.

1.3 Clustering Method

K-Means clustering and agglomerative clustering are applied in processing data. Before training the model, all data was standardized and applied t-SNE method for better visualization.

1.4 Validation Method

Scree plot (Elbow method) and Silhouette values are applied to evaluate the result of clustering in this report.

2. Result

The results of K-Means method and agglomerative method are displayed according to the number of clusters. The features used in training data are listed as followed: line_srvc_cnt, bene_unique_cnt , bene_day_srvc_cnt , average_submitted_chrg_amt average_Medicare_allowed_amt , std_la_p_service ,std_la_gender,std_la_drug.

Due to the limited space, only figures of 2, 4 and 9 clusters have been posted as representative ones. (Fig 2.1-2.4) In each figure, the left displays silhouette coefficient values of each clusters, while the right displays visualized clustering result after applied TSNE method.

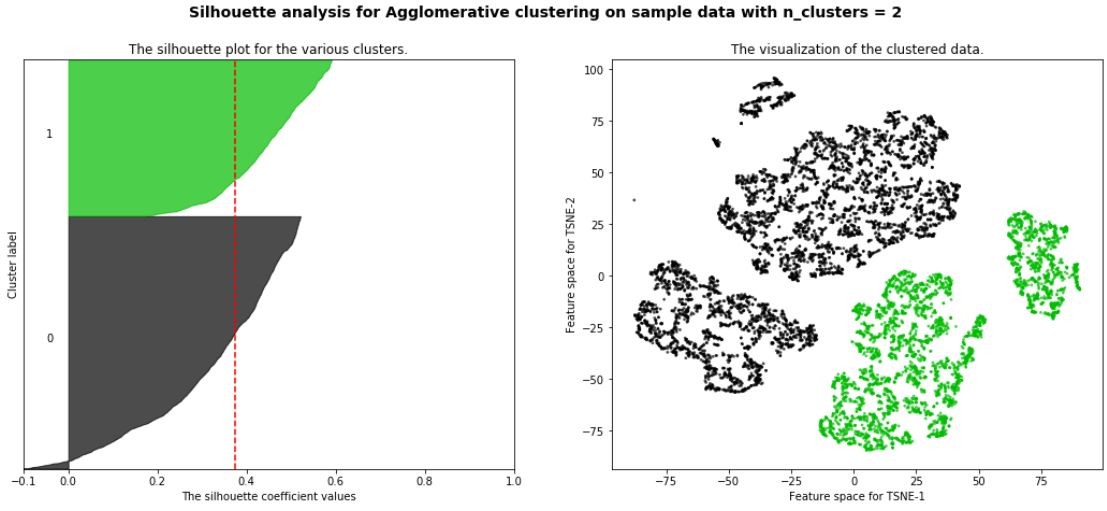

Fig 2.1 Agglomerative method with n_clusters = 2

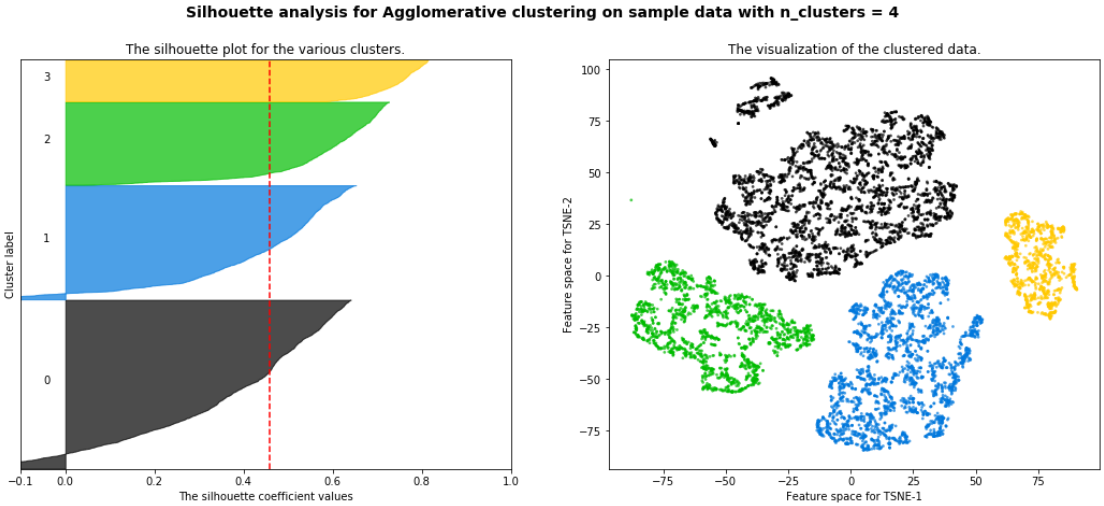

Fig 2.2 Agglomerative method with n_clusters = 4

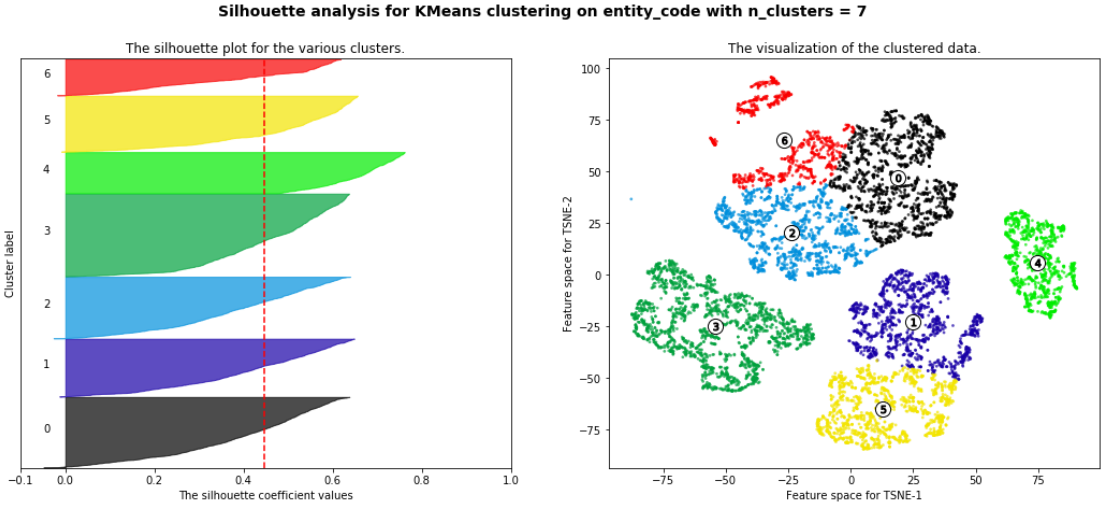

Fig 2.3 Agglomerative method with n_clusters = 7

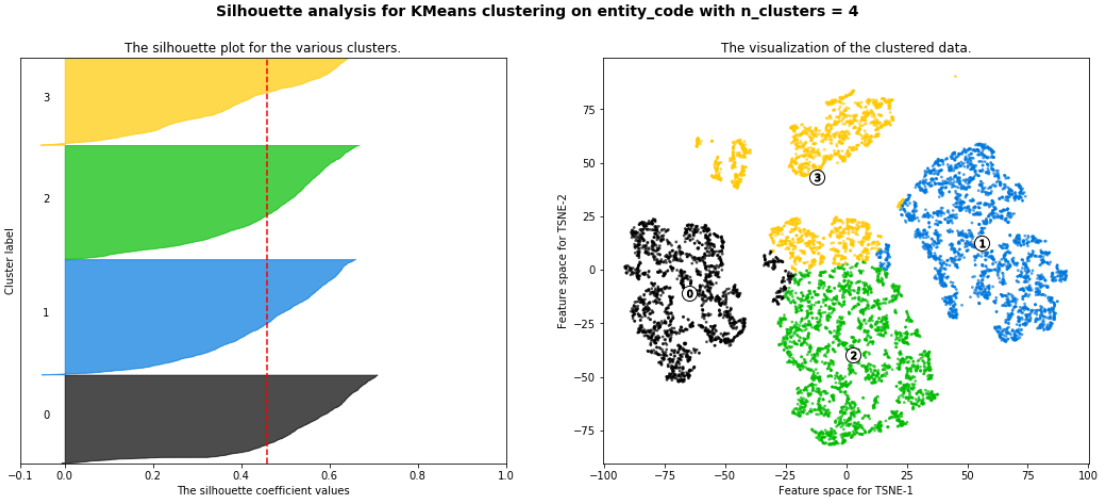

Fig 2.4 K-Means method with n_clusters = 4

The evaluation result of using K-Means method and agglomerative method are shown below:

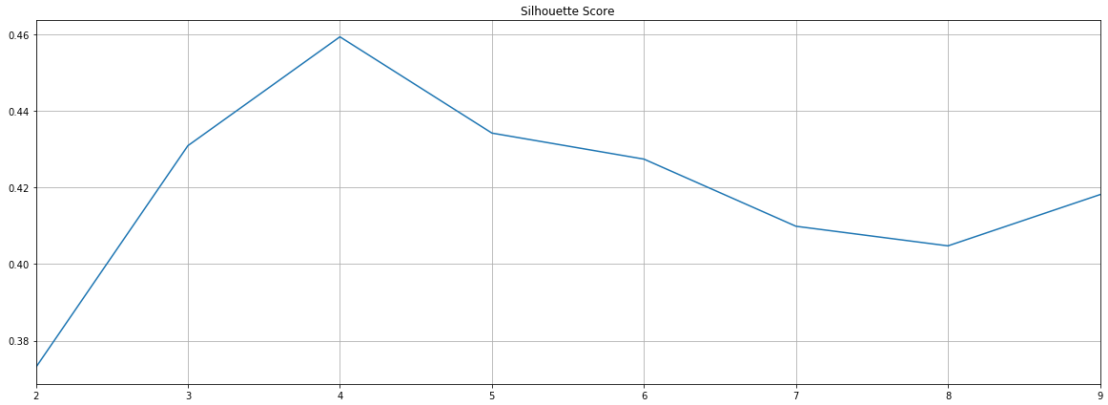

Fig 2.5 Silhouette score using agglomerative method

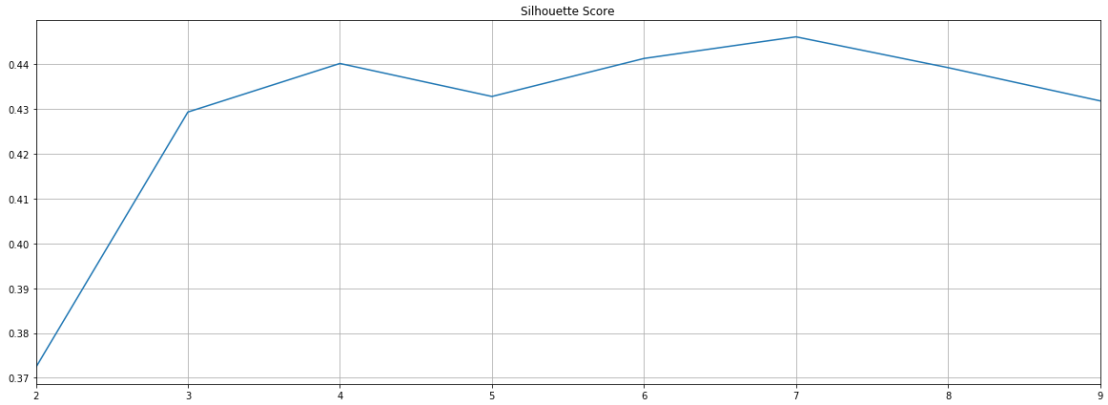

Fig 2.6 Silhouette score using K-Means method

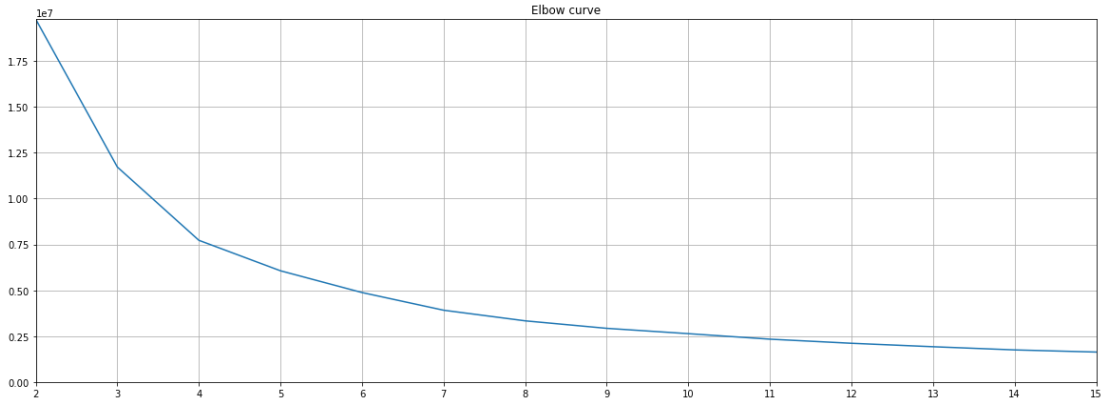

Fig 2.7 Elbow curve using K-Means method

The result of 4-clusters Agglomerative clustering is stored in file *Result_for_4clusters.csv*. The following is an example of this file.

	labels	credentials	p_type	no_serv	no_bene	no_dis_bene	ave_Me_pay	ave_Sub	place of service	gender	drug
0	1	MD	Internal Medicine	59	57	59	137.31	625	F	M	N
1	1	MD	Internal Medicine	114	110	114	206.06157895	923.99122807	F	M	N
2	1	MD	Pathology	438	289	294	29.240251142	88	F	M	N
3	1	MD	Pathology	24	17	18	27.02	179	F	M	N
4	1	MD	Anesthesiology	22	14	22	113.55909091	202.27272727	F	M	N
...	...	...	...	...	...	...	...	...	...	...	...
9246	1	MD	Nephrology	22	22	22	105.04	307.31818182	F	M	N
9247	0	DPM	Podiatry	59	28	59	23.18	125.19559322	O	M	N
9248	0	DPM	Podiatry	17	15	17	78.6	210.52	O	M	N
9249	2	Other	Other	185	25	185	44.87	60	O	F	N
9250	0	MD	Cardiology	378	155	378	125.98	696.76261905	O	M	N

Fig 2.8 An example of the result csv

3. Discussion

3.1 Clustering Validation Analysis

According to the trend of Silhouette value(Fig 2.5, 2.6), the value of Agglomerative Silhouette reaches its peak when the number of clusters is 4, and the value of K-Means Silhouette reaches its peak when the number of clusters is 4 and 7. According to the Elbow Curve(Fig 2.7), the “elbow” appears roughly at 4, 7 and 10 clusters. Therefore, 4 and 7 clusters both seem to be reasonable choices. However, through observing different visualized distribution of clustering, 4-clusters is a more reasonable choice.

Due to the limitation of K-Means which tends to produce similar size clusters, the result of 4-clusters K-Means (Fig 2.4) is not as good as Agglomerative (Fig 2.2). It can be told directly from the figure that the boarders of different clusters are blurred when applying K-Means.

Therefore, the result of 4-clusters Agglomerative clustering is chosen as the final result.

3.2 Clustering Results Analysis

To explore the connection between credentials and the clustering result, count of values of different credentials are calculated according to the labels produced by the result. The count of values is saved as file: *Res_value_counts.xlsx*, where numerical data are recorded as an average.

Evidently, there are significant differences between different clusters. Providers from Cluster0 are male who provide services in non-facility places, sometimes provide

services that are on the list of ASP file, and have higher value in numbers of services and lower value in amounts of submitted payment.

Providers from Cluster1 are male who provide services in non-facility places, almost never provide services that are on the list of ASP file, and have lower value in numbers of services and higher value in amounts of submitted payment.

Providers from Cluster2 are female who provide services in non-facility places, seldomly provide services that are on the list of ASP file, and have higher value in numbers of services and lower value in amounts of submitted payment.

Providers from Cluster3 are female who provide services in facility places, almost never provide services that are on the list of ASP file, and have lower value in numbers of services and higher value in amounts of submitted payment.

It is also noticeable that providers who are not individuals but organizations, are all classified as Cluster0.

According to the above classifications, most credentials have a significant higher proportion in Cluster0 and Cluster2, which means they provide services at non-facility places and have a relatively high number of services with low amount of charges. The ratio of Cluster0 + Cluster2 and Cluster1 + Cluster3 indicates the degree of this unevenly distribution. Among all credentials, PT has an extremely high value of ratio which is 50 : 1, while OD, DPM and Organizations have a ratio over 4 : 1, and PT as well as DO have a ratio around 2:1.

Oppositely, CRNA has a ratio of 1: 20, which indicates that CRNA providers are mainly providing relatively low numbers of services and having a high amount of charges. In addition, the biggest group MD has no such unevenly distribution, with a ratio close to 1: 1.

Likewise, most credentials have a higher proportion in Cluster0 and Cluster1, which indicates a higher proportion as male. As an exception, NP is mainly composed of female, while PT and PA have a higher ratio of female as well.

Moreover, more than 75% of the providers who provide services that are on the list of ASP file are MD and Organizations. And their number of services provided and amount of charges are all below the average of the same cluster.

3.3 Other Analysis

Single categorical feature has no significant influence on distribution of numerical data. Take feature “gender” as an example, the result obtained by clustering numerical data has no significant connection with the labels of feature “gender”. (Fig 3.1)

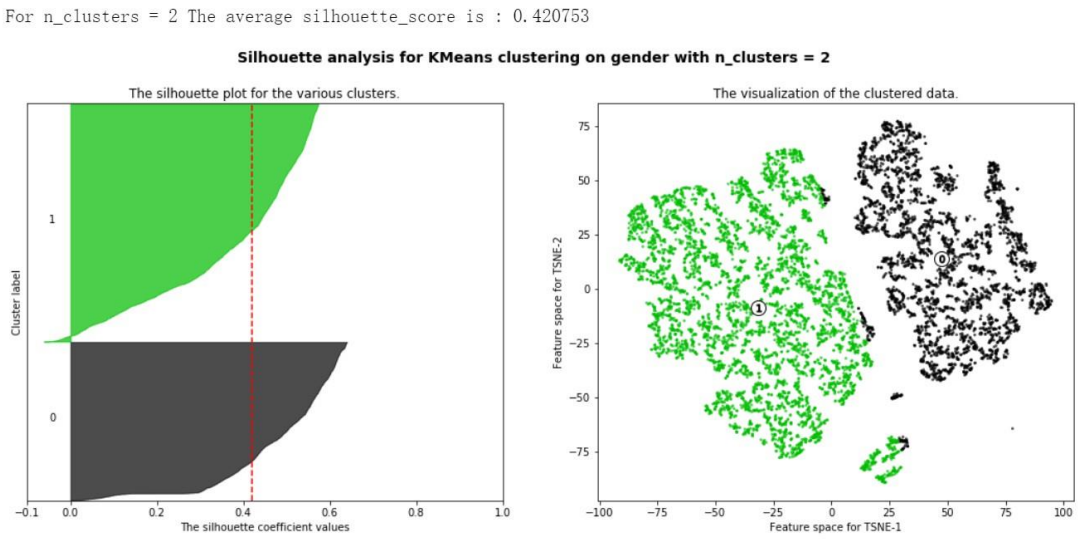

Fig 3.1 Clustering Distribution for feature gender

Although data is separated into two parts according to the clustering distribution, each part contains a considerable number of different labels of feature “gender”. As a matter of fact, the precision of this clustering result is only around 58%, which indicates that feature “gender” does not affect the distribution of numerical data.

Similar results have been found by applying same methods to different features like entity code, place of service, etc. In other words, single categorical feature including gender, entity code and place of service has no significant influence on the number services provided, number of beneficiaries and the amount of Medicare payment.

4. Conclusion

Based on the result of clustering and discussion, following conclusions can be drawn:

1. Providers with credentials of PT, PA, OD, DPM and Organizations have similar patterns: they provide services at non-facility places and have a relatively high number of services with low amount of charges. Among them, PT and PA has a proportion of female slightly higher than male, while the rest have a proportion of male significantly higher than females.
2. Providers with credentials of MD have following pattern: they are evenly distributed in most features without significant difference. However, the proportion of males is around 3 times higher than that of females.
3. Providers with credentials of NP have following pattern: they are evenly distributed in most features without significant difference. However, the proportion of females is more than 15 times higher than that of males.
4. Providers with credentials of CRNA have following pattern: they provide services at facility places and have a relatively low number of services with high amount of charges, and there is no significant difference in gender.
5. More than 75% of the providers who provide services that are on the list of ASP file are MD and Organizations.

6. A single feature, such as gender, place of service, etc., has no obvious relationship with payments and submitted charges.
7. Providers who are not individuals but organizations have the following pattern: they provide services at non-facility places and have a relatively high number of services with low amount of charges.

Reference:

1. Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview: The Centers for Medicare and Medicaid Services, Office of Enterprise Data and Analytics: 04.07.2014
2. http://www.cms.gov/Outreach-and-Education/Outreach/FFSProvPartProg/Downloads/2013-03-08-standalone.pdf

Appendix:

Credentials Abbreviation correspondence table

MD: Doctor of Medicine
NP: Nurse Practitioner
CRNA: Nurse Anesthetists
PA: Physician Assistant
OD: Doctor of Optometry
DPM: Doctor of Podiatric Medicine
DO: Doctor of Medicine
Org: Organizations