Southern Methodist University

CSE -7331: Introduction to Data Mining

Dr. Michael Hahsler

Data mining project 2

Cluster Analysis

Dmonty, Sevil Sanjao 47568070
Panchmahalkar, Pritheesh 47524741

Table of Contents

Executive Summary

Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known.

We have hospital data from project 1 and we are going to use that data to do our cluster analysis. In this project we are going to use various clustering methods such as K-means, Hierarchical clustering and Gaussian mixture model for clustering. These algorithms take a subset of data based on primary group conditions like diabetes, hip fracture and pregnancy. We are going to use this preprocessed data and going to apply various clustering methods on them which might help us to predict future about the subset. The clusters would help the users understand some common patterns in the data and use it for medical purposes and better health care of the patients. The insurance companies also can observe some patterns in the data they might be interested in, say, patients with less stay in the hospital, or less claims. This report talks about various clustering algorithms, internal validation and choosing the optimal number of clusters and external validation based on the source of truth which is generally given by experts.

Data Preparation

In order for the hospital dataset to be used for the clustering models, it must be preprocessed. Many of the columns are not necessary in order to carry out the clustering. The attributes used for different clustering models used are as follows. For this project, only the subset of data that belonged to Year 1 (Y1) has been chosen.

Attribute	Scale	Distance Measure	Description
Claim	Numeric	Euclidean	Total number of claims filled by member within duration of year 1.
Paydelay	Numeric	Euclidean	It is number of days between the date of service and date of payment
Charlson Index	Numeric	Euclidean	A measure of the affect diseases has on overall illness, grouped by significance, that generalizes additional diagnoses.
DrugCount	Numeric	Euclidean	Count of unique prescription drugs filled by DSFS (Days since First Service). No count is provided if prescriptions were filled before DSFS zero. Values above 6, the 95% percentile after excluding counts of zero, are top-coded as “7+”.
Lab Count	Numeric	Euclidean	Count of unique laboratory and pathology tests by DSFS (Days Since First Service). Values above 9, the 95% percentile after excluding counts of zero, are top-coded as “10+”.
Length of Stay	Numeric	Euclidean	Length of stay (discharge date – admission date + 1), generalized to: days up to six days; (1-2] weeks; (2-4] weeks; (4-8] weeks; (8-12 weeks]; (12-26] weeks; more than 26 weeks (26+ weeks).
PrimaryConditionGroup	Nominal	Gower	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates, that generalize the primary diagnosis codes

Claims: - The claims attribute gives the number of claims that each member has filed. This gives us total number of claims filed by each member for year 1. The minimum value of claims is 1 and maximum is 43. By this we can assume that if a member has maximum claims could be very sick person with very high charlson index value or there is a possibility that the member was diagnosed with chronic disease such as diabetes, asthma etc. which require regular checkup and follow-up with the doctor. Eighty-eight percent of Americans over 65 years of age have at least one chronic health condition. It is possible that a patient who has just one claim might be suffering from a life-threatening disease too, if the claim is recorded at the end of Year 1. The claims attribute can be grouped along with the primary group condition and help the health care providers understand the claims per each condition, based on which the patient can be given medication. From insurance companies' point of view, the members who have less claims could be profitable most of the time when compared to the members with high claims.

PayDelay: - The attribute PayDelay is number of days between the date of service and date of payment to the vendor. It has value starts from 0 day up till 162 + days. The delay in payment could be due to various reasons like financial situation of the patients, lack of communication of the payment due date or conflicts with the insurance providers. An average PayDelay for each member is calculated using the claims data. This attribute gives an overview of the timeline of the payment process of the members. The health care providers would be happier with the patients who have less average PayDelay when compared to ones with higher average PayDelay. For the sake of this project, all the values in the PayDelay are converted to numeric values, i.e., the value "162+" is converted to 162. Based on the new attribute, the mean of the means of the pay delay of the members is around "52.26%".

Charlson Index: - According to Wikipedia, "The Charlson comorbidity index predicts the one-year mortality for a patient who may have a range of comorbid conditions, such as heart disease, AIDS, or cancer (a total of 22 conditions). Each condition is assigned a score of 1, 2, 3, or 6, depending on the risk of dying associated with each one. Scores are summed to provide a total score to predict mortality."

The patients with less Charlson Index are healthy and the one's with high Charlson index are not healthy i.e., suffering from life threatening diseases. This attribute is important to know about the health condition of a patient. The patients with high Charlson Index are likely to visit the hospitals more often when compared to the patients with low Charlson Index. Also, the hospital bills for the patients with high Charlson Index are likely to be higher. From the insurance companies point of view, it could be profitable for them to focus on patients with low Charlson index, as they are likely to pay less and visit the hospitals less often. The doctors/researchers based on Charlson index can focus their research on diseases that result in high Charlson index and design medicines that can cure such diseases.
The values in the claims data is modified as show in the table below.

Charlson Index	Value taken for charlson index
0	0
1-2	1.5
3-4	3.5
5+	5.5

Drug count: - The Attribute Drug count is total number of drugs prescribed by the doctor to the member during the year 1. In the drugs data, the values above 6 are top-coded as "7+". To keep the data numeric, this value has been changed to 7. The minimum value of Drug count is 1 and maximum value is 84. The drug count gives information about the condition of the patient. The more the drug count value of a member, the higher the probability that a patient is suffering from a disease that needs long time to be cured or can never be cured. The lower the drug count of a patient, the healthier is the patient which is the objective of the health care providers. The patients should be careful when using drugs, as the drugs could be a reason for illness too. The patients who take regular prescriptions should try to eliminate few drugs, so they can avoid the effects of the drugs.

Lab count: - The Attribute Lab count is the number of Laboratory tests prescribed by the doctor to the member in the year 1 data. In the labs data, the values above 9 are top-coded as "10+". To keep the data numeric, this value has been changed to 10. The minimum value of lab counts is 1 and maximum is 80. Lab tests like blood tests, urine tests, MMR, X-rays etc. are necessary to determine the health condition of the patient. We can assume that the member having maximum number of Lab count are the least healthy people when compared others. Also, there is possibility that members with higher Lab count could be suffering from chronic disease for which that member must visit doctor on regular basis and get the tests done.

Length of stay: The Attribute length of stay is the total number of days a member spent in the hospital of health care facility. There are missing values in the length of stay attributes which we have changed to 0 days, we have considered that a member whose length of day is missing might not admitted to hospital but instead that member was admitted in morning and later discharged in evening, or some members got treated in ambulance and got discharged on same day. we have changed values for length of stay attribute from string to numeric is as follows,

Length of Stay	Value taken for Length of Stay
Missing data	0
1 day	1
2 days	2
3 days	3
4 days	4
5 days	5
6 days	6

1-2 Weeks	7
2-4 weeks	14
4-8 weeks	28
8-12 weeks	56
12-26 weeks	84
26+ weeks	182

The sum of length of stay of each member is calculated based on their claims in the claims data for the clustering methods. The longer the length of stay, the higher is the probability that the patient is suffering from a disease that requires more medical care. It could also be possible that the missing values in the data could be due to the reason of privacy, but for the sake of simplicity, we have chosen a value 0 instead of NA. The highest length of stay was observed as 562 days, which results due to the way the data was changed. It is possible that the patient has been admitted for 2-4 weeks throughout the year or may be the claims were recorded multiple times each kind of Primary Condition group, place of service, specialty.

PrimaryConditionGroup: - As per American Medical Association there are various types of primary conditions, we have total 45 types of primary conditions listed in our dataset. Every member who has filled claim must have one of the primary condition listed. It is important attribute for clustering as it gives us primary condition under which the member was receiving treatment. This attribute can help us to get the value of clarlson index by which one can understand the level of illness the member has. Primary Condition group has been used to select a subset of data from the claims.

Modeling

K-Means Clustering

Method	Cluster Data set	K value
K-means (Lloyd)	Pay Delay	3
	Lab Count	
	Claims	

The K-Means clustering is done on the subset of the data related to the patients suffering from diabetes. The features average PayDelay, LabCount and claims have been considered for the patients who had claims with PrimaryConditionGroup value as **METAB1** which stands for "Diabetic ketoacidosis and related Metabolic" based on the claims data. There are 987 patients with claims related to diabetes based on the claims data on Year1.

The diabetic patients are likely to visit the doctors often in order get the tests like blood test to get a report of their health condition. Hence, it would be interesting to analyze these patients based on the lab count and number of claims. As these patients would visit the doctors often, they probably have more claims and hence the insurance providers would also be more interested in the clustering.

Cluster#	Size	DrugCount	LabCount	Claims
1	463	-0.4297883	-0.62463502	-0.7355004
2	201	1.5598283	0.08926798	0.4592263
3	323	-0.3545928	0.83982400	0.7685208

Table: Results of K-Means algorithm

The above table shows the results of the k-means algorithm. Cluster#1 is the largest cluster with size more than 450. Cluster#1 is the cluster of patients who have all the attributes less when compared to the other clusters. It is clear that the cluster#1 is plotted using the red color based on the result of K-Means algorithm. Cluster#2 has patients that have higher DrugCount. The Cluster#2 has been plotted using the green color and it is the smallest of the three clusters. Cluster#3 has patients with low DrugCount, higher LabCount and Claims. Cluster#1 has relatively more healthier patients of the three clusters. Based on cluster#2, it is possible that these patients visit the doctors regularly for the prescription of drugs to maintain their health in control. Cluster#3 contains patients with high LabCount but low DrugCount. It could be possible that the lab tests may have resulted in normal condition, so they didn't need drugs and these patients had the tests done regularly as they are health conscious. The insurance companies would be more interested in the patients in the cluster#1 as they have low claims, labs and drugs counts.

The health care providers and medical scientists would focus their research on designing medicines or procedures for the patients in cluster#2 and cluster#3, so that these patients get healthier.

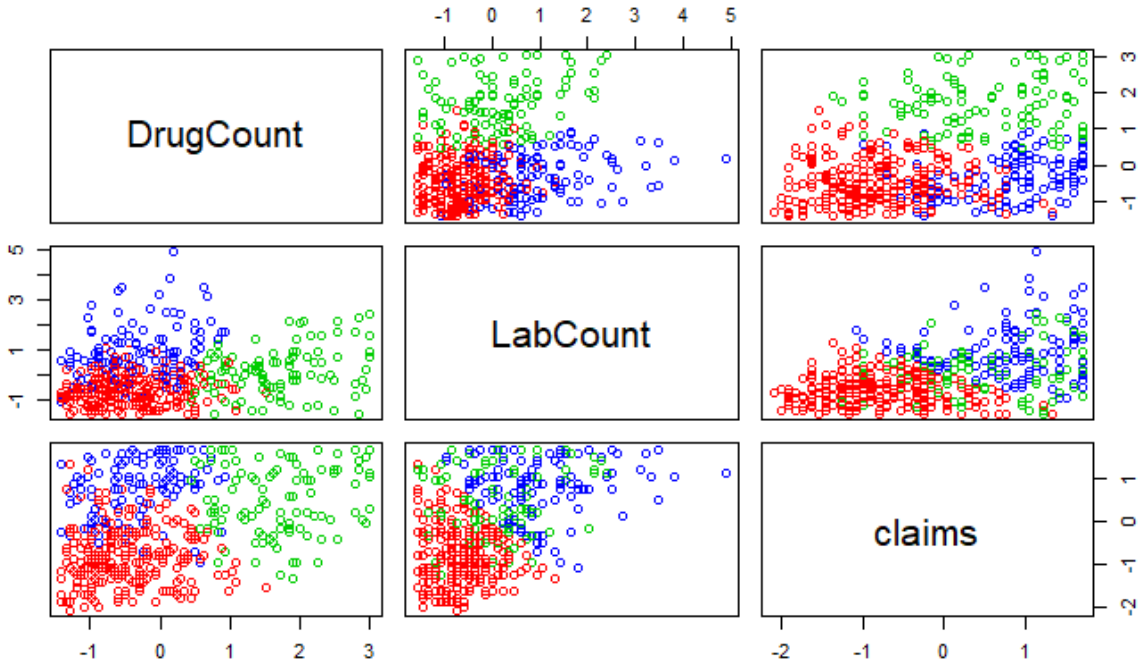

Figure 1: K-Means plot of DrugCount, LabCount, claims on the patients suffering from Diabetes

Determining suitable number of clusters and Internal Validation of clusters

K-Means is an algorithm which takes the number of required clusters as one of its parameters. The optimal number of clusters for the K-Means algorithm has been decided using two metrics, Within Sum Squares (Elbow method) and Average Silhouette width. Figure#2 and Figure#3 are the plots plotted using the metrics obtained from K-Means algorithm for different number of clusters from 2 through 10. Based on the plots, the optimal number of clusters is chosen as 3.

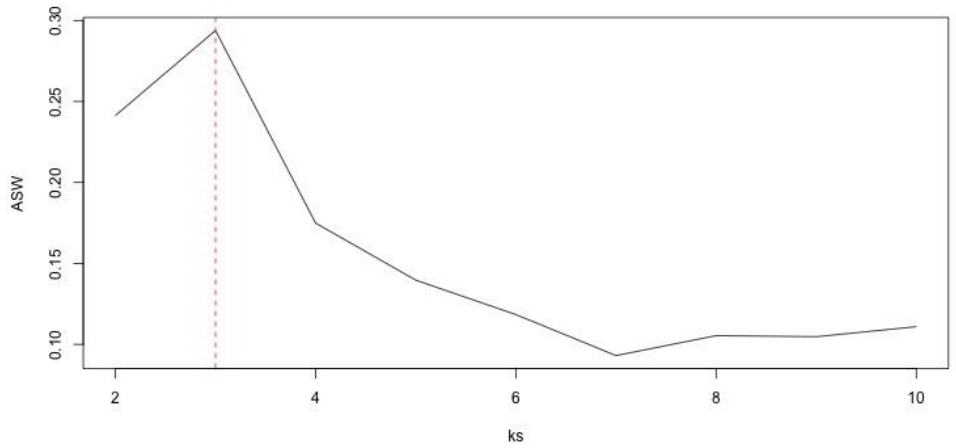

Figure2: Plot showing the result of average silhouette width for number of clusters ranging from 2 through 10 using the k-means algorithm

The internal validation of the clusters is done using the cluster.stats methods from the package fpc in R. The cluster.stats method is used with the K-Means result and some important metrics calculated are tabulated below.

Within cluster sum of squares	1449.128
Average silhouette width	0.3364
Dunn Index	0.0172
Average distance between clusters	2.6468
Average distance within clusters	1.511

The within cluster sum of squares which is the objective function of k-means is obtained based on the Euclidean distance matrix. It is desirable to have value for within sum of squares for good quality of clusters.

Silhouette width is used to measure the degree of similarity of an object in one cluster compared to other clusters. High value of Silhouette width indicates that the object is similar to the objects in the cluster and dissimilar to objects in other clusters.

Dunn Index is calculated as minimum separation / maximum diameter. The higher the Dunn Index, the better are the clusters.

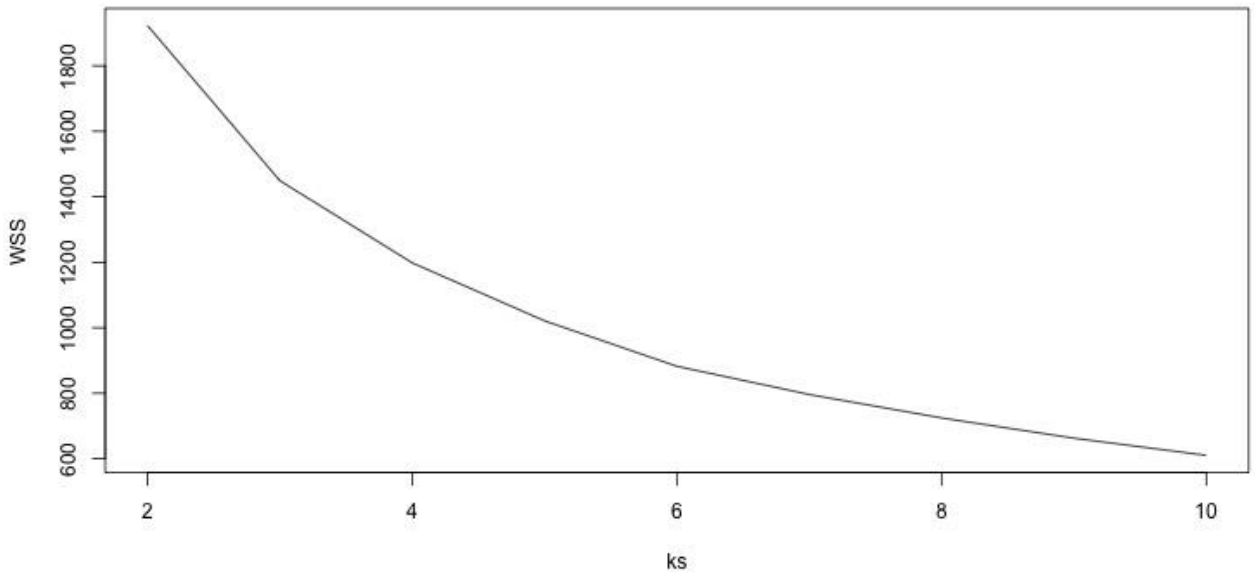

Figure 3: Plot showing the result of Within sum squares for number of clusters ranging from 2 through 10 using the k-means algorithm

The Figure#3 doesn't really give much information about the number of clusters that could be chosen for clustering. But, a value of 3 was chosen considering both the silhouette plot and the within sum squares plot

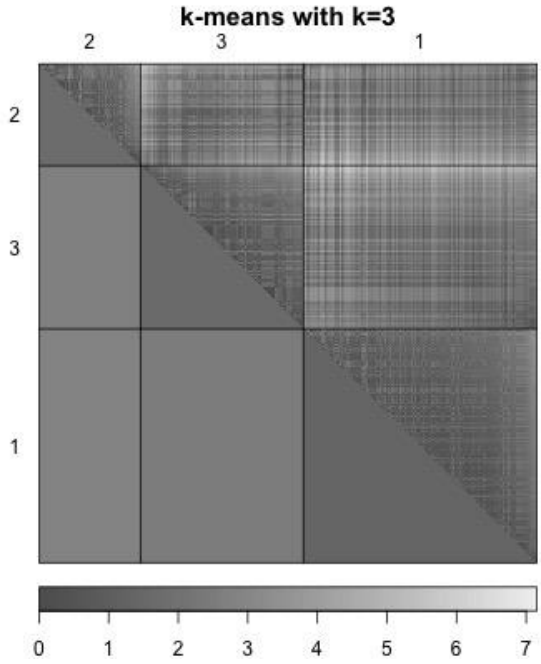

Figure 4: Dissplot for kmeans algorithm on patients with diabetes for 3 clusters

From the figure#4, the number of clusters chosen as 3 is good. The cluster#1 is the large cluster and is at the bottom right of the plot, whereas, the smallest cluster, cluster#2 is at the top left. The regions in these clusters are dense when compared to the other parts of the plot.

External Validation

External validation is done on the clusters based on source of truth as Charlson Index. The actual Charlson Index of the members were grouped as “low”, “medium” and “high”. Now, based on the values of Claims, DrugCount, LabCount for the members, a value for CharlsonIndex has been assigned to check for external validation. For Cluster1 as the patients are healthy, CharlsonIndex of “low” was assigned. CharlsonIndex of “low” was assigned for the patients in the cluster#2 as well and CharlsonIndex of “medium” was assigned to patients in the cluster#3. Around 67% of the patients assigned to clusters were found to be equal.

Hierarchical Clustering

Method	Cluster Data	Number of centers
Hierarchical (complete)	Drug Count	6
	Length of Stay	
	Claims	

For the Hierarchical Clustering method, the data related to patients suffering from Hip Fracture was taken. Table above shows data set we have taken for Hierarchical clustering, we found total 964 members who has Hip Fracture as Primary condition in year 1 data. Figure #5 shows a dendrogram of Hierarchical clustering based on the data.

The data of the patients with hip fracture was observed using the hierarchical clustering using 6 clusters. The older age patients are more likely to suffer from hip fracture as the hip bones are not as healthy for them when compared to the young people. Hence, these patients are likely to visit the doctors, get X-rays more often. Two large clusters with more than 75% of the patients were formed along with 4 other clusters with the rest of the patients. Cluster#1 contains people that have relatively less DrugCount, LabCount and claims when compared to the other claims. The insurance providers would be more interested in the patients in this cluster. The other clusters vary with DrugCount, LabCount and claims. The patients in cluster#3 use relatively more drugs and have higher claims when compared to the other patients. Cluster#5 is small with 13 patients but the labcount and claims for these patients is high which could be possible because they may be suffering from other conditions as well which require lab tests.

Cluster#	Size	DrugCount	LabCount	Claims
1	363	-0.3986094	-0.3025196	-0.9804352
2	381	-0.2143262	-0.2537183	0.7229669
3	85	2.16089197	0.08047061	0.92860203
4	83	-0.3382762	1.4621376	0.4143745
5	13	-0.0568253	6.2071386	0.9011432
6	39	1.83316217	-0.06176762	-1.14337242

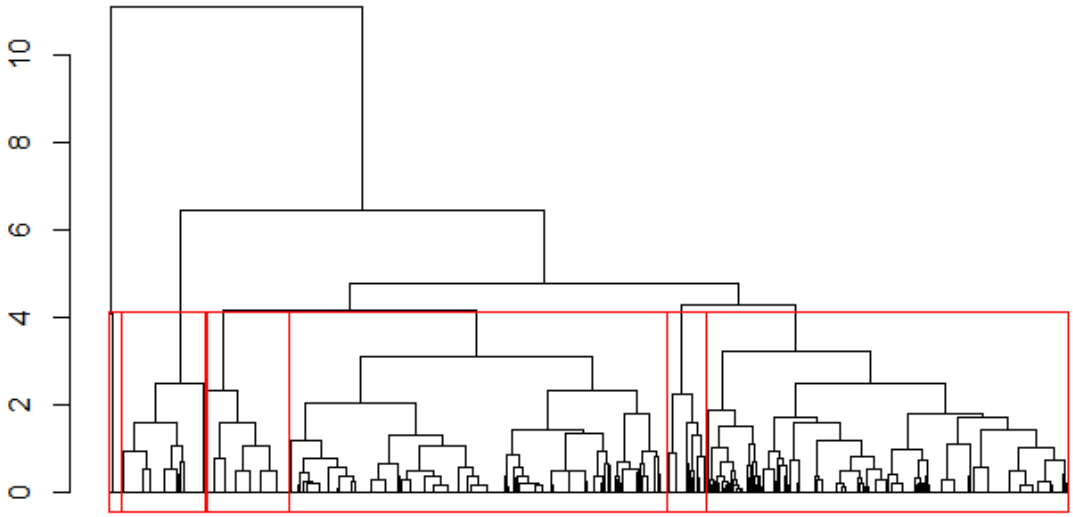

Figure 5- Hierarchical Plot Drug count, Length of stay and Claims with the number of members having Primary Condition of Hip Fracture.

Determining suitable number of clusters and Internal validation

Hierarchical clustering taken the number of required clusters as one of the required parameter. The optimal number of clusters has been decided by using average Silhouette width. Figures #6, #7, #8 are the plots plotted using the metrics obtained from Hierarchical clustering algorithm for different number of clusters ranging from 1 to 10. Based on the plot, the value for optimal number of clusters we choose is 6.

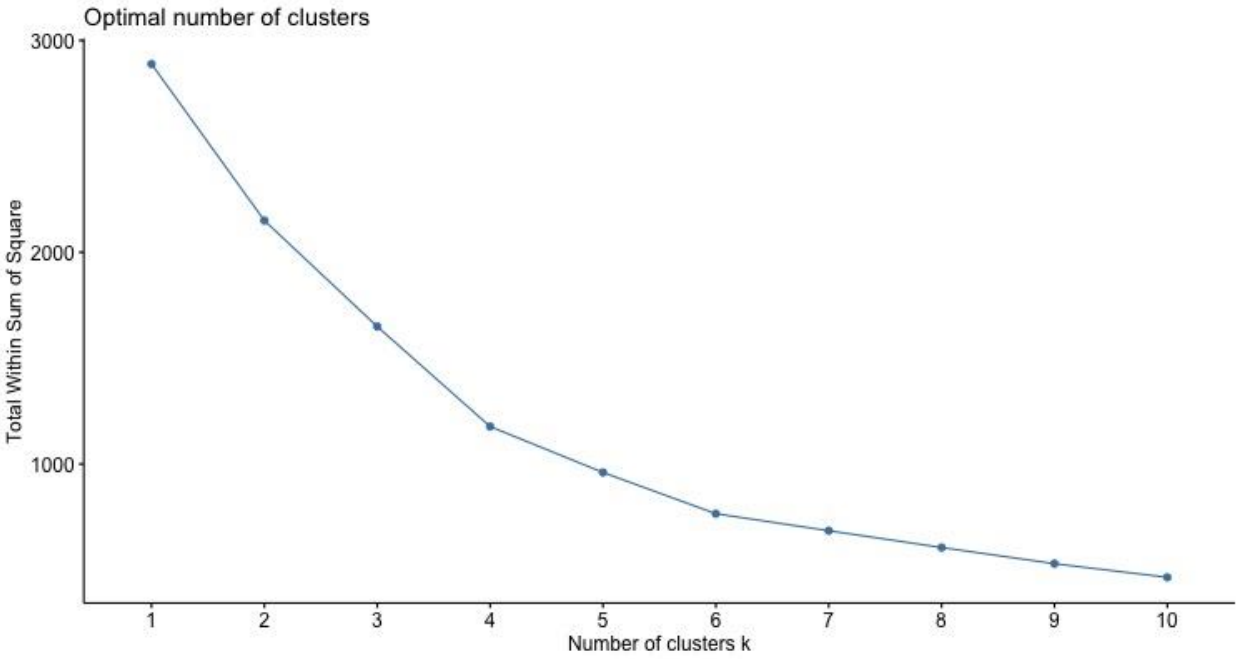

Figure 6: Plot showing the result of Within sum squares for number of clusters ranging from 1 through 10 using the Hierarchical algorithm

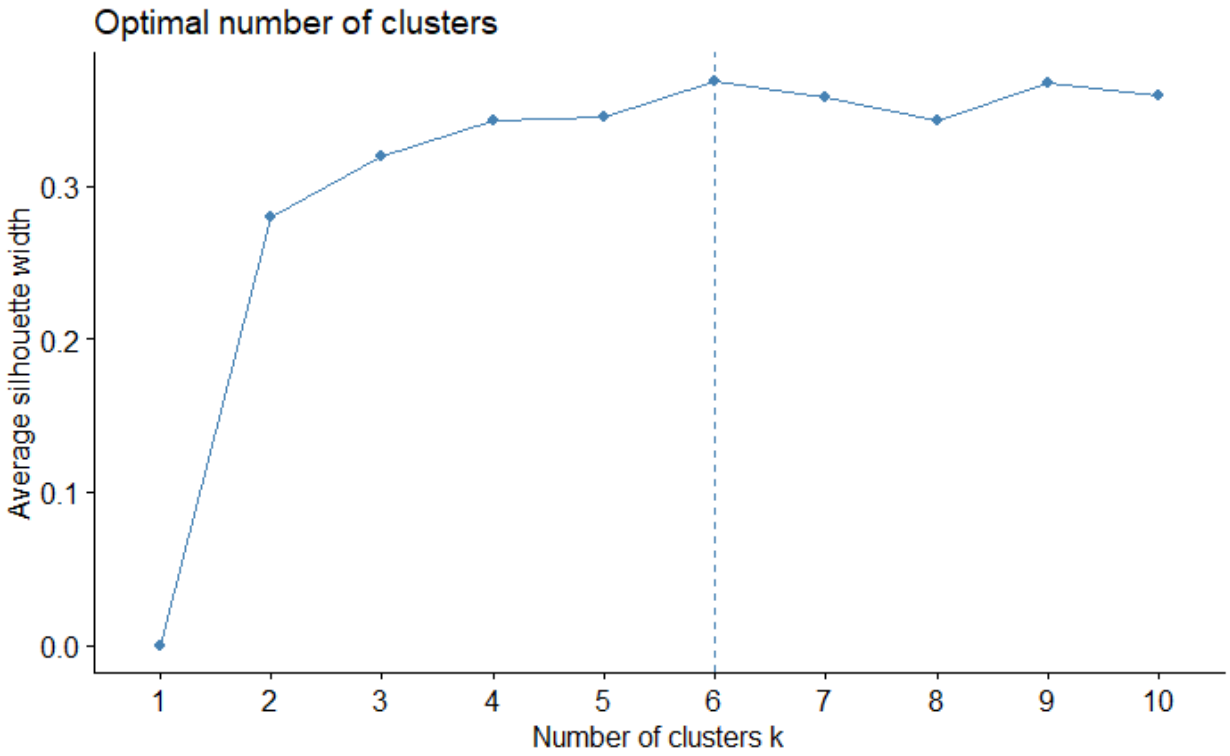

Figure 7: Plot showing the result of Within sum squares for number of clusters ranging from 1 through 10 using the Hierarchical algorithm

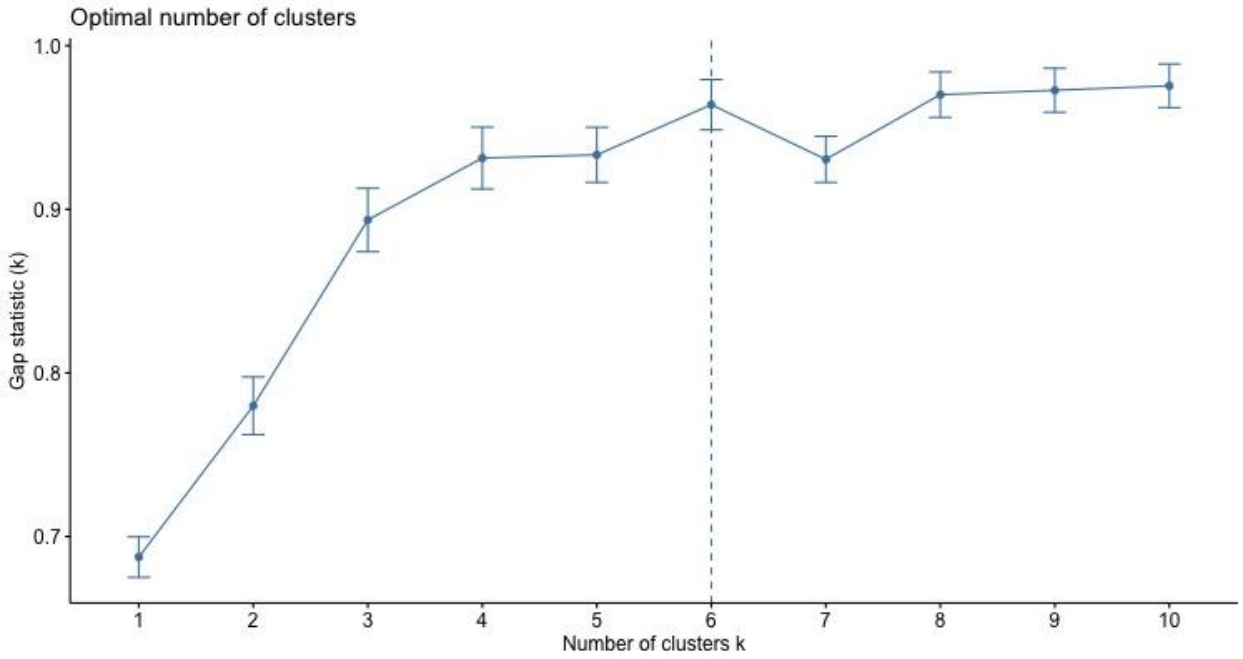

Figure 8: Plot showing the result of average silhouette width for number of clusters ranging from 1 through 10 using the Hierarchical algorithm

Based on the results of the clustering, the cluster stats of the clusters are calculated and tabulated as shown below.

Within cluster sum of squares	842.3952
Average silhouette width	0.3544387
Dunn Index	0.06095805
Average distance between clusters	2.54771
Average distance within clusters	1.214715

"The idea behind their approach was to find a way to standardize the comparison of log W_k with a null reference distribution of the data, i.e. a distribution with no obvious clustering. Their estimate for the optimal number of clusters K is the value for which log W_k falls the farthest below this reference curve. This information is contained in the following formula for the gap statistic:

$$Gapn(k)= E*n\{logWk\}–logWk$$ "

The number of clusters is chosen as the using the value of cluster for which the gap statistic is maximum which is at k = 6.

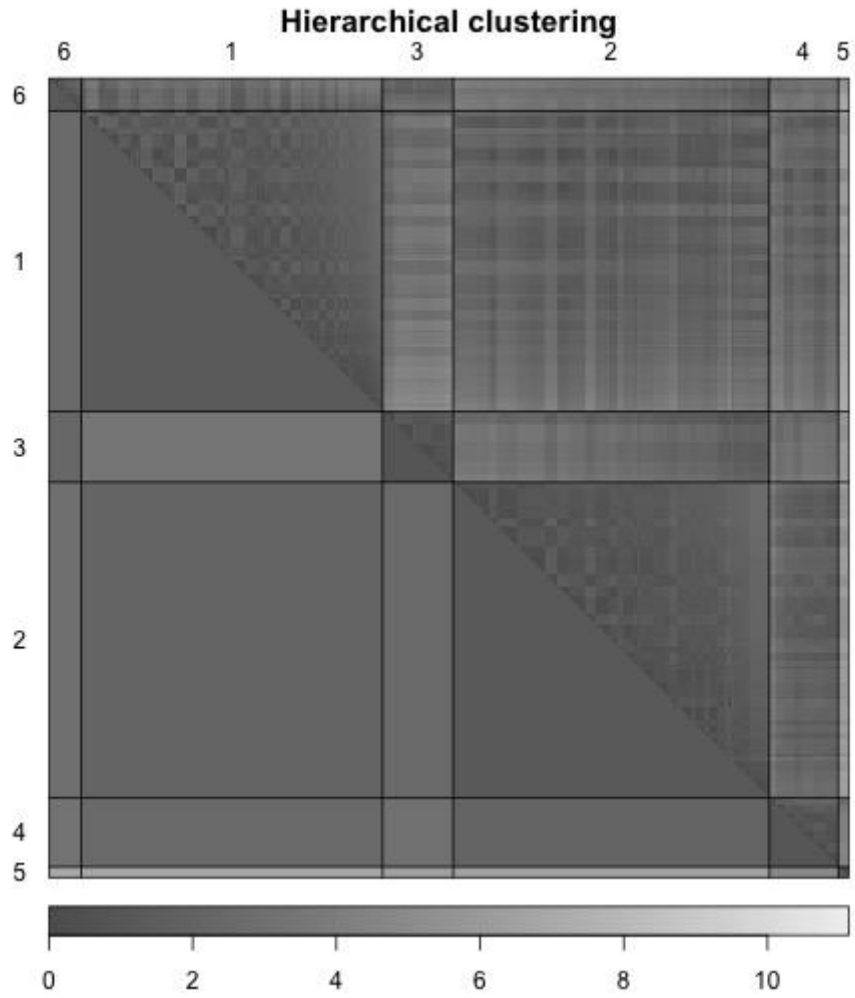

Figure 9: Dissplot for Hierchical clustering algorithm on patients with hip fracture for 6 clusters

The above plot is plotted using the results of the hierarchical algorithm and it shows that 6 clusters is a good choice for the data subset as the diagonal has the color in the range close to 0 when compared to the other regions in the plot.

External Validation

The external validation is done on the clusters based on the CharlsonIndex as source of truth for the clusters. The Charlson Index for the patients has been grouped as “low”, “medium” and “high”. Based on the values of DrugCount, LengthOfStay and claims a value for the Charlson Index is predicted and checked with the actual value. All the clusters except the fifth have been assigned a low Charlson index and the fifth has been assigned a Charlsonindex of “medium”. The total accuracy was found to be 97.5%.

Gaussian Mixture Model based Clustering

Method	Cluster Data	Centers
Gaussian Mixture Model	Drug Count	6
	Lab Count	
	Length of Stay	

For the Gaussian Mixture Model based clustering we have taken all of the members from year 1 data who have filed claims under their primary condition as PRGNCY which stands for Pregnancy, along with their Drug Count, Lab count and Length of Stay in Hospital. Figure #10 shows the plotting of members with their Drug count, Lab Count and Length of stay. we have taken random sample 2500 number of members with pregnancy claims, table below shows cluster wise distribution of data.

Cluster#1,2 and 3 have less Length of stay, Drug count and Lab Count then cluster#4,5 and 6. By this we can assume that first 3 cluster members are relatively healthier than other cluster members. We can see that there is 50-50 distribution between first 3 clusters and last 3 clusters or we can say that there is 50% distribution between healthy and unhealthy patients. We also observed that cluster 4 and 6 members has longer length of stay and lab counts which might because of members might gone through surgery and for which they need more care and observation. Cluster#5 is also very interesting it has less members and less length of stay but drug count and lab count is high comparing with other clusters. This might be because of the member has some complication during her pregnancy and doctor might have prescribed more drugs to them.

Clusters#	Size	Length of Stay	Drug Count	Lab Count
1	111	-0.2962620	-0.1494608	-1.5137589
2	332	-1.0129094	-0.4030962	-0.7973055
3	803	-0.1877747	-0.6248324	-0.1026199
4	765	0.41027022	-0.02658375	0.32871744
5	80	-0.1527193	3.6996392	0.1601204
6	409	0.5337774	0.9205963	0.6133446

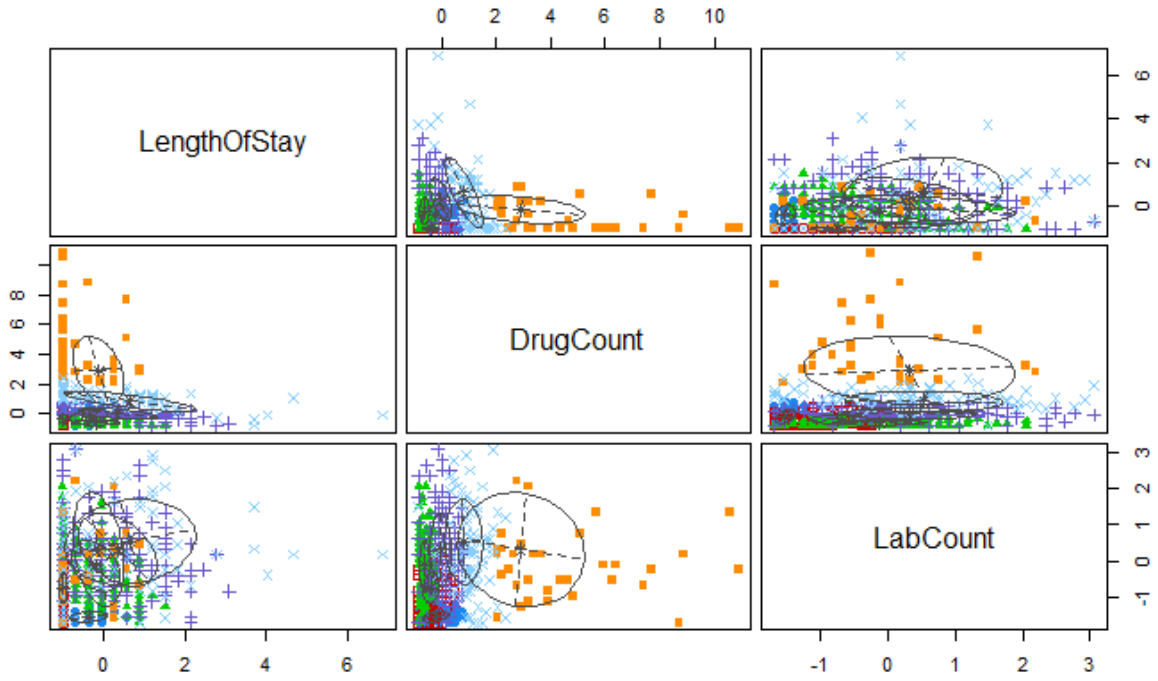

Figure 10 - Plot showing clusters created using the Gaussian Mixture Model clustering.

Determine the optimal number of clusters and Internal Validation

The optimal number of clusters has been decided by using average silhouette width plot, Figure #11 is the plot plotted using the metrics obtained from Gaussian Mixture Model clustering algorithm from different number of clusters ranging from 1 to 10. Based on Figure #11, #12 the value we chose the optimal number of clusters is 6.

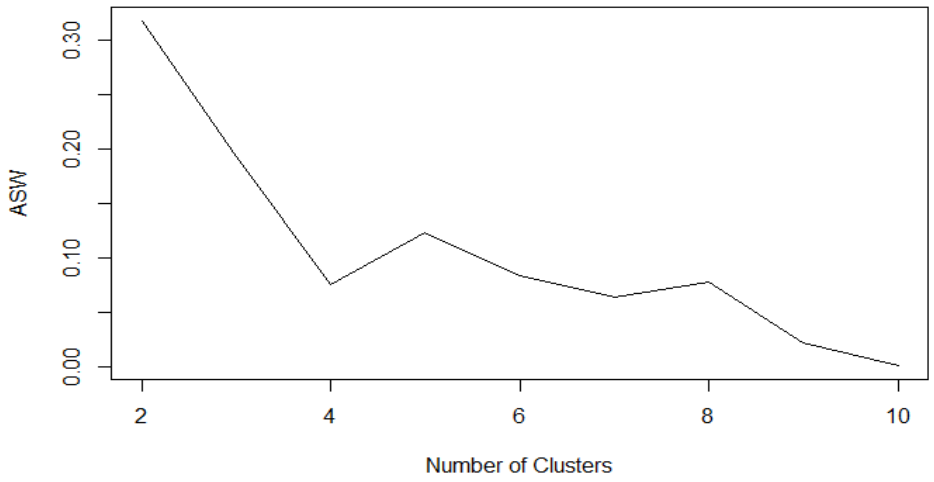

Figure 11- Plot showing the result of average silhouette width for number of clusters ranging from 2 through 10 using the Gaussian Mixture Model based algorithm.

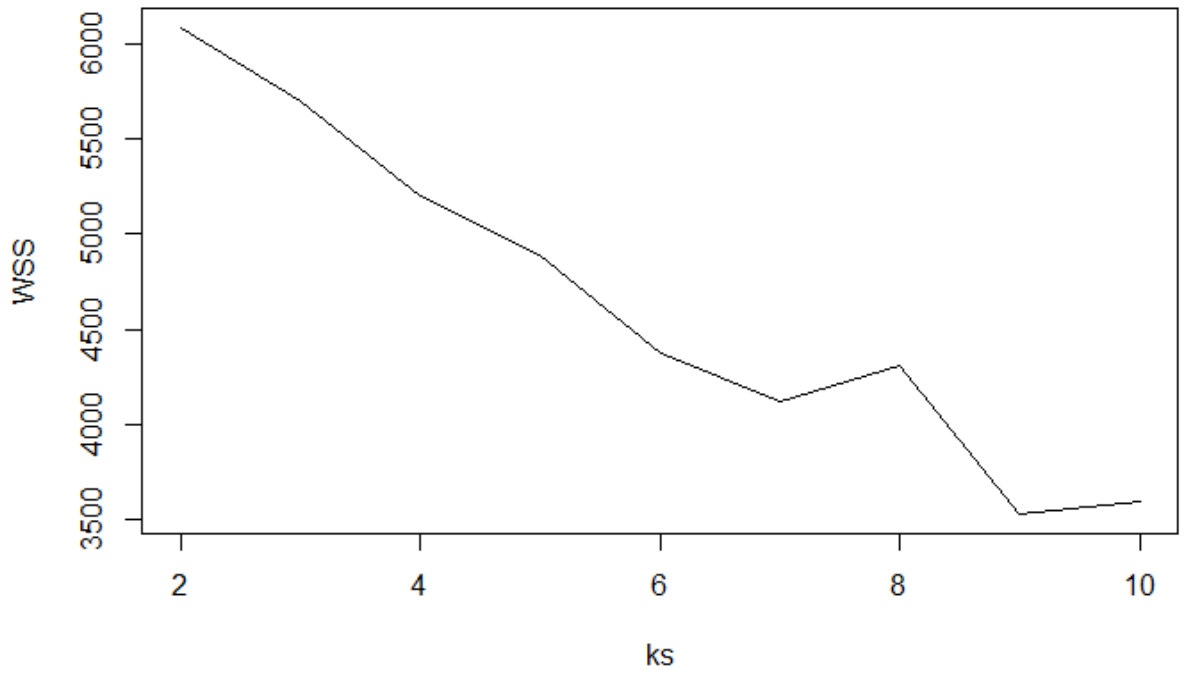

Figure 12 - Plot showing the result of Within sum squares for number of clusters ranging from 2 through 10 using the Gaussian Mixture Model based algorithm.

The internal validation of the clusters is done using the cluster.stats methods from the package fpc in R. The cluster.stats method is used with the Gaussian Mixture model based algorithm, result and some important metrics calculated are tabulated below.

Within cluster sum of squares	4346.635
Average silhouette width	0.07226543
Dunn Index	0.0158318
Average distance between clusters	0.07226543
Average distance within clusters	1.50671

The within cluster sum of squares is obtained based on the Euclidean distance matrix. It is desirable to have value for within sum of squares for good quality of clusters.
Silhouette width is used to measure the degree of similarity of an object in one cluster compared to other clusters. High value of Silhouette width indicates that the object is like the objects in the cluster and dissimilar to objects in other clusters.
Dunn Index is calculated as minimum separation / maximum diameter. The higher the Dunn Index, the better are the clusters.

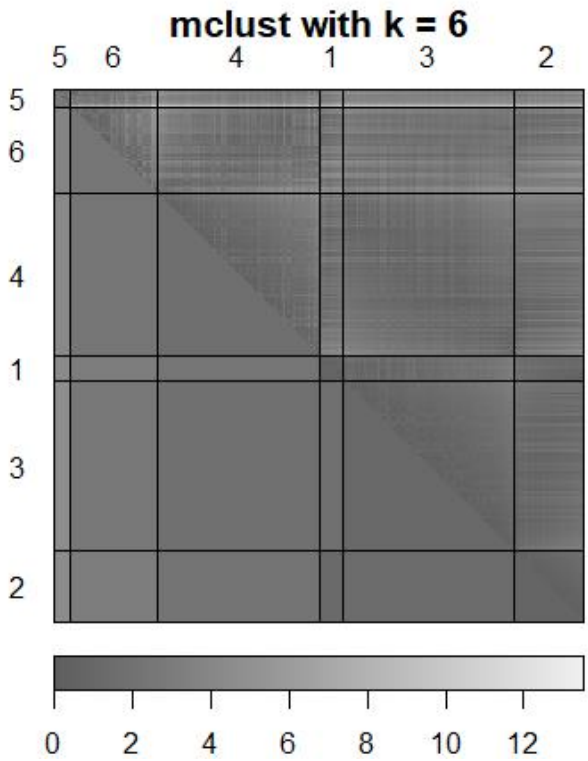

Figure 13 - *Dissplot for Gaussian Mixture Model clustering algorithm on patients with hip fracture for 6 clusters*

External Validation

The external validation is done on the gaussian mixture model using the Charlson index. The pregnant patients had just two Charlson Index values in the claims table which were 0 and 1.5, so, there are only two levels in the Charlson Index. The Charlson Index value is determined for each cluster based on the LengthOfStay, DrugCount and LabCount. An accuracy of 85.56% was obtained when validated using the CharlsonIndex.

Evaluation and conclusion

The dataset mostly contains nominal data and it is difficult for the clustering algorithms to cluster the data. Hence, only the numeric data and the newly created columns were chosen from the dataset for modeling. However, even with the numeric data, the clusters formed were unstable, i.e., the clusters formed do not have a good degree of similarity for the objects in them, but few facts could be observed. The data was very diverse and hence it was nearly as similar as the random data for clustering algorithms. Based on K-means, the patients in the cluster#1 were found be possibly healthy and had relatively less claims, labcount and drugcount. These patients could have received good treatment and hence, they got healthier soon or it could be possible that these patients have had claims in the end of the year.

Based on Hierarchical clustering, for which the data of patients having primary condition group as hip fracture was considered, the Cluster#1 had patients with less drugcount, labcount and claims. It could be possible that these patients could belong to young age group and they got healthier very soon or they could be of either age group and had their claims at the end of the year. Cluster#2 had patients with less drugcount, less labcount but relatively higher claims. It could be possible that these patients have been under diagnosis for longer duration as they may be of older age group.

Based on the Gaussian mixture model, for which the data of patients having primary condition group as pregnancy was considered. It has been observed that it is divided 50-50 in less lab count, drug count and Length of stay, from which we thought that there are nearly 50% healthier pregnant members and nearly 50% pregnant members who need medical care than other.

The dataset mostly contains nominal data and it is difficult for the clustering algorithms to cluster the data. Hence, only the numeric data and the newly created columns were chosen from the dataset for modeling. However, even with the numeric data, the clusters formed were unstable, i.e., the clusters formed do not have a good degree of similarity for the objects in them, but few facts could be observed. The data was very diverse and hence it was nearly as similar as the random data for clustering algorithms.

Based on the analysis of various clusters obtained using the clustering algorithms on different subsets of the data, the health care providers and the insurance companies can focus their interest. The outliers in the clusters could be analyzed and conclude whether it is an abnormality or just a wrong entry in the data. Common patterns in the data can be observed and it would help the health care providers understand the need for specific equipment in the hospital. It would also help the medical scientist observe patterns and design medicines according to the need and seriousness of a condition. The medical stores also could do cluster analysis based on drugs and primary condition to predict what kind of medicines could be used by the patients.

Reference

https://www.aha.org/statistics/fast-facts-us-hospitals

https://journals.lww.com/jbjsjournal/Abstract/2005/03000/Early_Mortality_After_Hip_Fracture___Is_Delay.1.aspx

https://datasciencelab.wordpress.com/tag/gap-statistic/