Southern Methodist University
CSE -7331: Introduction to Data Mining

Dr. Michael Hahsler

Association Rule Mining

Project 4

Sevil D'monty - 47568070
Pritheesh Panchmahalkar - 47524741

Table of Contents

Executive summary

In this project, we use the same dataset as used for the previous project. We also create new count variables to see if they could make any effect on the rules. The variables in the dataset are continuous and hence, need to be discretized in order for the Apriori algorithm to be run. We discuss about the discretization of different continuous variables and create transaction based on the discretized variables. These generated transactions are used as input for the Apriori algorithm to generate the Frequent item sets and Association rules. The most frequent items in the dataset like length of stay being zero and the office count being between 0 and four appear in most of the rules. The newly created count variables are 0 for most of the patients as they do not always make claims based on the primary conditions like heart, renal etc. But, if we observe that if we examine only the data of unhealthy patients, the newly created count variable come into play and become part of frequent item sets and Association rules. We observe that the groups created in the association rules are similar to results we observe with clustering and they seem to agree with each other based on conditions like Pregnancy or hip fracture.

Data Preparation

For this association rule mining project, we are going to use similar data as we have used for previous project, project 3 classification. There are lots of attributes within the given data which might be not necessary for this project, so we are not taking them into consideration. The list of useful attributes for this project are as follows.

Attribute	Description
Claim	Total number of claims filled by member within duration of year 1,2 and 3.
Paydelay	It is number of days between the date of service and date of payment
Charlson Index	A measure of the affect diseases has on overall illness, grouped by significance, that generalizes additional diagnoses.
DrugCount	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	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	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).
dih levels	The attribute days in hospital gives days spent in hospital by the member.
AgeAtFirstClaim	Age in years at the time of the first claim’s date of service computed from the date of birth; Generalized into ten-year age intervals.
Heart count	Number of claims filed by the member having primary condition under heart diseases.
Injured count	Number of claims filed by the member having primary condition under MISCELLANEOUS #2 which means that the member got injured by accident.
Renal Count	Number of claims filed based on the PrimaryCondition Renal2 which is related to chronic renal failures.
Member ID	Unique member identifier.
Cancer Count	Number of claims filed by the members having primary condition under cancer diseases
Office Count	The count of claims a member has made with place of service at the office.
Urgent Count	The urgent claims count attribute gives the count of claims a member has made with place of service as either ambulance, or urgent care.
Sex	Gender of the member either Male or Female
ClaimsTruncated	ClaimsTruncated (a flag for members who have had claims suppressed. If the flag is 1 for member xxx in DaysInHospital_Y2, some claims for member xxx will have been suppressed in Y1).

Claims: - The claims attribute gives the number of claims that each member has 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. As per the report on healthcare finance a new analysis by Change Healthcare released at the 2017 Healthcare Financial Management Association's ANI revealed compelling statistics about claims denials and their financial impact on hospitals. The finding showed that out of roughly $3 trillion in medical claims submitted by hospitals in the United States last year, around nine percent of charges were initially denied. That comes out to about $262 billion [1]. 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

Table 1: Charlson Index values and modified values

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.

Dih_levels: The attribute days in hospital gives days spent in hospital by the member. The data in this attribute ranging from 0 to 15 maximum. Let us consider the Histogram plotted using the data which omits the count when days in hospital is 0. It is clear that the frequency is very high for a day in the hospital and drastically low for the rest of the days except for the last value which is 15. It is possible that the value 15 is the result of encoding "15+ days" as 15.

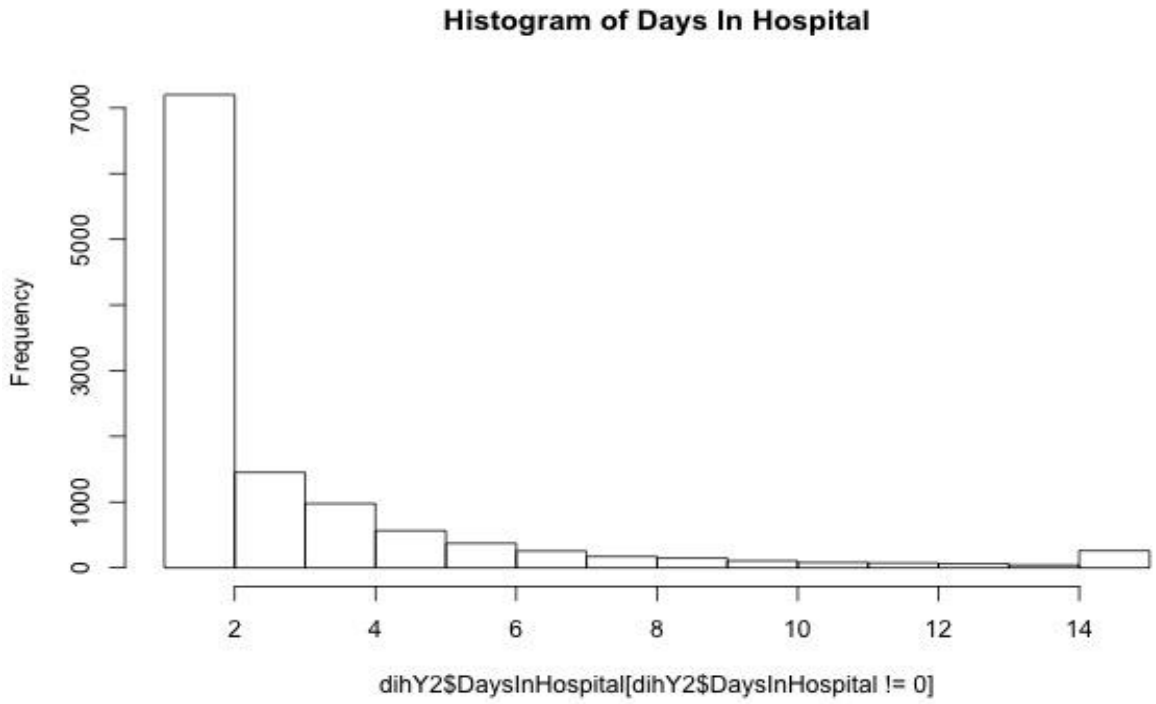

Figure 1: Histogram of Days in Hospital

We decided to divide this data in 3 classes namely as 'No Stay', 'Short Stay' and 'Long stay'. To divide the days in hospital among three classes we chose the median of days in hospital whose value is not zero. In this way, the classes could be nearly half way divided among short stay and long stay. If a member was admitted for 0 days then he/she is in No Stay level, same for the member who was admitted in hospital for 1 to 2 days then he is part of Short Stay level and lastly the member who were admitted for more than 2 days are categories as Long Stay. By this we can assume that the person who visits doctor for checkup or a surgery which do not require hospitalization comes under No Stay. The Member who need medical treatment up to 2 days in hospital are comes under Short stay such as Members who had minor accident, fractures etc. Long stay category there are members who might require long stay mostly more than 2 days such as member with very high charlson index and old age members or members who had serious accident, deliveries and might take long time for recovery. Approximately we got 83 % of data comes

under the class 'No Stay' and nearly 10% of data comes under 'Short Stay' and 7% of data comes under 'Long Stay'.

AgeAtFirstClaim: This attribute gives age of the member when he/she filed for their first claim or in other words the age of the member at which he/she filed their first claim. This attribute has values with the interval of 10 years such as 0-9, 10-19 till 80+. This attribute is very complicated as you cannot be sure if the member who filed for their first claim what was his/her actual age as this attribute has range of ages.

Heart claims count: We have created this attribute named as Heart count. The Heart Claims count attribute gives the count of claims a member made when suffering from an ailment related to heart. To get the count of such claims the primary condition group could be AMI (ACUTE MYOCARDIAL INFARCTION), HEART4 (ATHEROSCLEROSIS AND PERIPHERAL VASCULAR DISEASE), CATAST (CATASTROPHIC CONDITIONS which includes cardiac arrest), CHF (CONGESTIVE HEART FAILURE), MISCHRT (MISCELLANEOUS CARDIAC), HEART2 (OTHER CARDIAC CONDITIONS), PERVALV (PERICARDITIS) or STROKE. For the models, as the objective is to predict the number of days a member spends in the hospital, it is possible that the members suffering from problems related to heart are likely to spend more in hospital, say ICU after the cardiac arrest. The mean of length of stay of the members based on claims on above mentioned PrimaryConditionGroups is 0.46.

Injured claims count: The injured claims count attribute gives the count of claims a member has made when suffering from injuries. The injured claims count is calculated by using the PrimaryConditionGroup 'MSC2a3' (Misc#2) which deals with external causes of injuries. The patient is likely to spend more time in the hospital due to injuries as the patient might have broken ribs, limbs etc. and needs to spend some time in the hospital.

Renal claims count: The renal claims count attribute gives the count of claims a member has made when suffering from problems related to kidney like chronic renal failure, end-stage renal disease and kidney transplants. These problems are likely to make the patient spend more time in the hospital due to dialysis, or the patient may be under observation for a few days etc.

Urgent claims count: The urgent claims count attribute gives the count of claims a member has made with place of service as either ambulance, or urgent care. Urgent care plays a critical role in saving a patient's life. Based on a report from CDC, majority of the patients that needed urgent medical care were males and age group less than 34. [3] We will see if this applies for our dataset too based on the generated rules.

Cancer Count: We have created this attribute using the attribute Primary condition group from the claims data. We have grouped together various types of cancer diseases such as CANCER A (Malignant neoplasms of respiratory tract and intrathoracic organs; leukemias, non-Hodgkin's lymphomas, and other histiocytic malignancies), CANCER B (All other malignant neoplasms not in Cancer A or gynecologic ones (including Hodgkin's disease); radiation therapy and chemotherapy encounters where cancer not specified). We all know that the mortality rate for the disease like cancer is very low.

Office claims count: The office claims count attribute gives the count of claims a member has made with place of service at the office. The frequency of visit of patients reflects their health. *Around 990.8 million visits have been recorded at a physician's office.* [2]

Based on figure 2, it is clear that most of the claims that the patients have made were in the office.

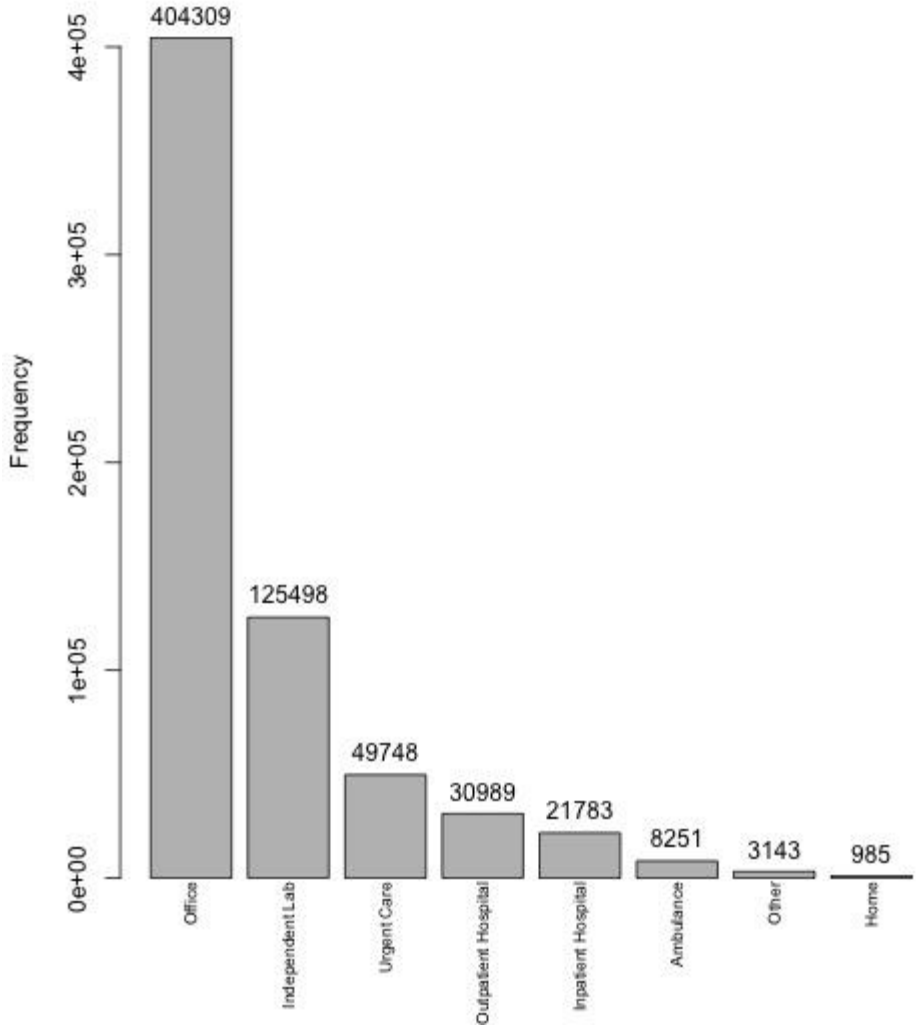

Figure 2: Barplot showing the frequency of different places of service

Discretization of continuous variables

In the context of variables, discretization is conversion of continuous variable into discrete variables. There are different ways in which continuous variables can be discretized. We will discuss equal-width discretization and equal-frequency discretization.

Equal-Width Discretization

The equal-width discretization method divides the range of the continuous variable into n intervals which are of equal length.

Equal-Frequency Discretization

The equal-frequency discretization method sorts the values of the continuous variable and then divides them into n intervals such that each interval has approximately equal number of values.

We will also use a user defined interval discretization which defines the intervals in which the range of the continuous variable should be divided.

Let's examine the status of each variable we are going to use to mine association rules and discretize them based on different approaches.

Charlson Index

Based on figure 3, Charlson Index has high frequencies for lower CharlsonIndex which is good as it means majority of the patients are healthy. Based on the frequency of the data, we have divided CharlsonIndex into 2 groups, one with healthy patients (CharlsonIndex is zero) and other with not (so) healthy patients (Non-zero CharlsonIndex). This would divide the patients into two groups with almost equal frequency. From figure 4, after discretizing the CharlsonIndex, we can see that the patients have been divided into two groups with approximately 16000 healthy patients and 12000 unhealthy patients.

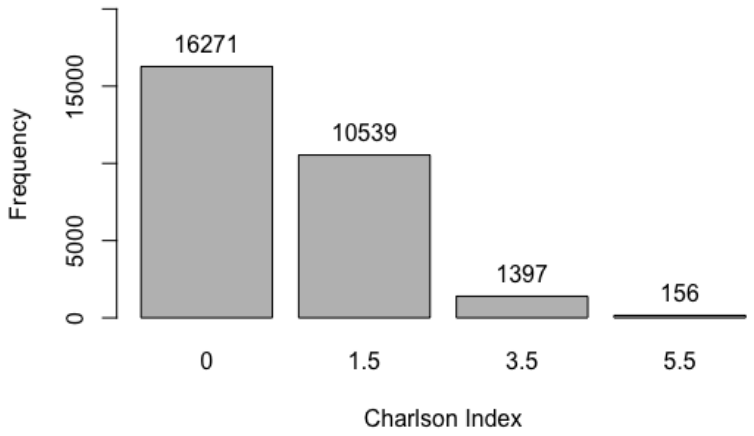

Figure 3: Bar plot showing the frequencies of different values of CharlsonIndex

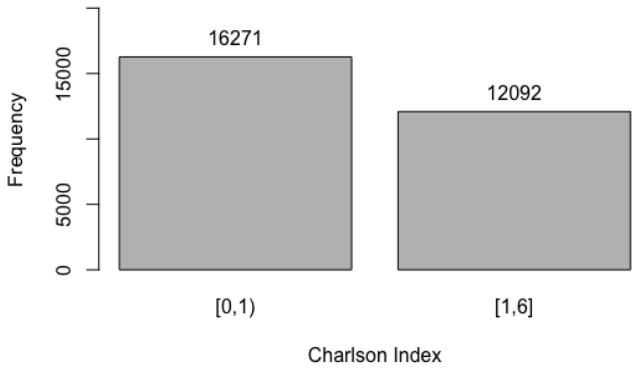

Figure 4: Bar plot showing the frequencies of different values of CharlsonIndex

Length of stay

Length of stay has a wide range of distribution with patients not staying in hospital till may be a whole year.

Mean	1.924
Median	1

We have divided the length of stay into 3 groups based on intervals 0-1 days, 1-2 days and 2-474 days. The interval end days are exclusive. It is because most of the times, patients do not have to stay in the hospital unless in serious condition. The one day stay in hospital could be because the patient could be kept under observation and released the next day it could turn out that the patient is actually healthy. Based on figure 5, we can see that there a lot of patients (~82%) with 0 length of stay.

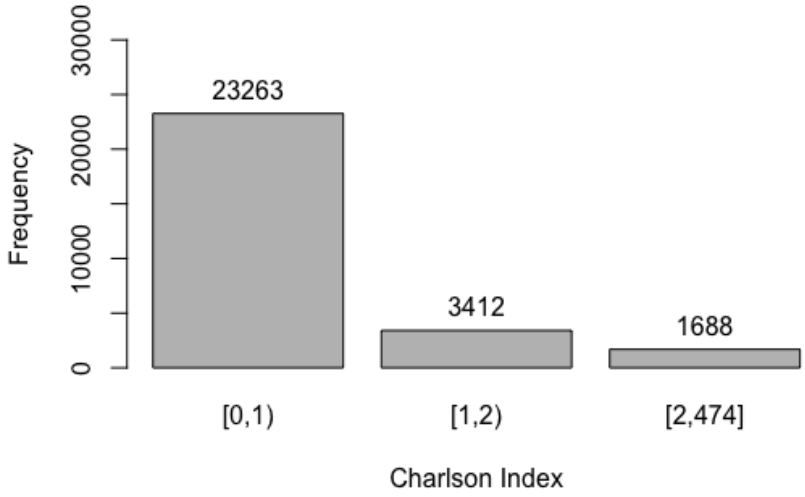

Figure 5: Bar plot showing the frequency of new Charlson index

Renal Count, Cancer count, Urgent count

The patients suffering from renal problems are very less and hence most of the patients do not have any claims based on renal primary condition. Because of the vast difference in the renal counts, we have divided the patients into 2 groups, one having claims based on renal condition and the other without any claims on renal condition. Similarly, the count variables like cancer count and urgent count were divided into 2 groups. These variables would be interesting to observe to compare the rules with dataset having only the patients with these counts > 0.

Interval	Renal count	Cancer count	Urgent count
[0,1)	27985	22123	20424
[1, 20)	378	6240	7939

Heart count

Based on figure 6, we can see that there were relatively larger claims based on heart conditions. Heart count has been divided into 3 groups (i.e., groups with 0 heart claims, 1-2 heart claims and 3 or greater heart claims). The patients with 0 heart claims do not have any heart-based conditions. The patients with 1-2 heart condition-based claims have become healthier after the medical treatment, or the claims could have been at the end of the year or the worst might have happened to the patient. The table below shows the distribution of frequency of claims.

Interval	[0, 1)	[1, 3)	[3, 38]
Frequency	17061	5580	5722

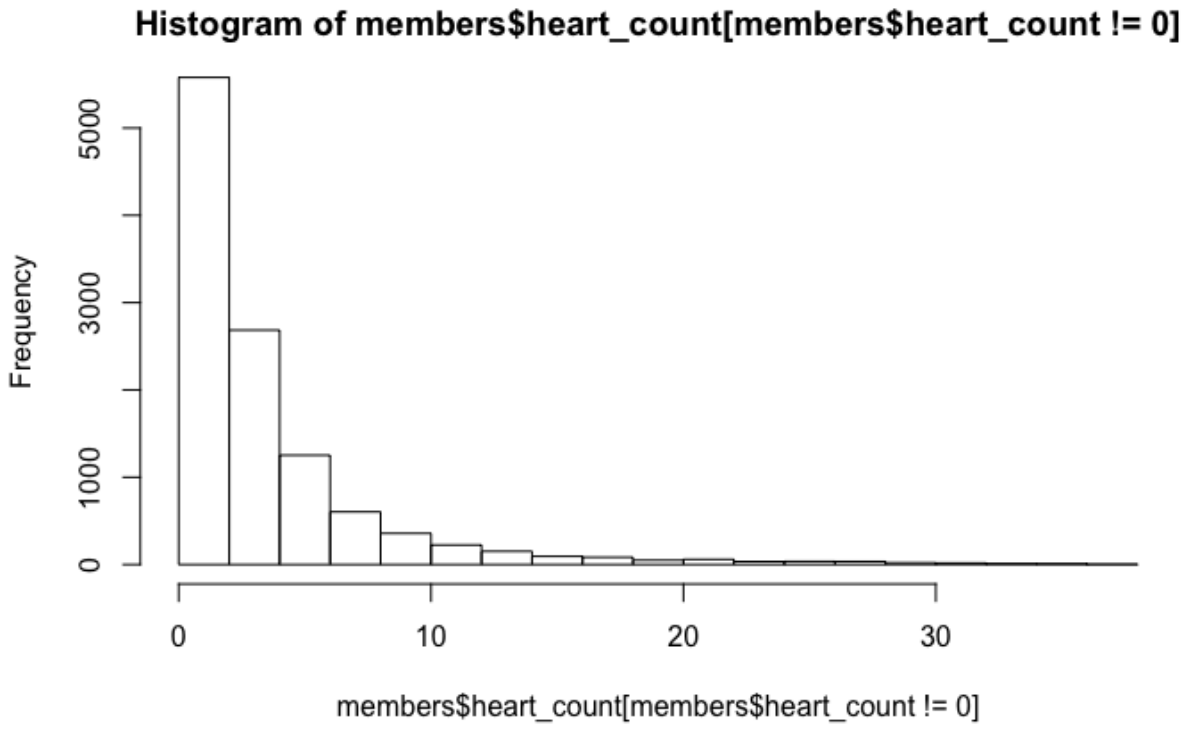

Figure 6: Histogram showing frequency of heart count with heart count claims > 0

The other variables like AgeAtFirstClaim, claims, PayDelay, DrugCount, LabCount, injury count, office count have been discretized using the discretizeDF method in arules package in R. The default has been used for all the parameters, i.e., the value of methods is frequency (The variables are divided into groups based on the frequency), breaks = 3 (divide the variable into 3 different groups). The data for these attributes has a nice distribution and hence, could be a good idea to divide them into 3 groups (say low, medium and high) based on frequencies.

The data after discretization looks as shown in the image 1. As we have prepared the data, that is, we have discretized all the continuous variables first, we are now ready to create the transactions. Transactions have been created by coercion from this data frame.

	AgeAtFirstClaim <fctr>	Sex <fctr>	claims <fctr>	PayDelay <fctr>	CharlsonIndex <fctr>	LengthOfStay <fctr>	DrugCount <fctr>	LabCount <fctr>
1	[74.5,85]	F	[8,16)	[57.2,162]	[1,6]	[0,1)	[6,19)	[11,69]
2	[54.5,74.5)	F	[16,43]	[0,41.1)	[0,1)	[0,1)	[19,84]	[11,69]
3	[4.5,54.5)	F	[8,16)	[57.2,162]	[0,1)	[0,1)	[6,19)	[1,6)
4	[4.5,54.5)	M	[1,8)	[0,41.1)	[0,1)	[0,1)	[6,19)	[6,11)
5	[74.5,85]	F	[1,8)	[0,41.1)	[0,1)	[0,1)	[6,19)	[6,11)
6	[4.5,54.5)	F	[16,43]	[41.1,57.2)	[0,1)	[2,474]	[6,19)	[11,69]

	dih_levels <fctr>	renal_count <fctr>	cancer_count <fctr>	heart_count <fctr>	injury_counts <fctr>	office_count <fctr>	urgent_count <fctr>
1	no_stay	[0,1)	[0,1)	[3,38]	[3,43]	[4,9)	[0,1)
2	no_stay	[0,1)	[0,1)	[3,38]	[3,43]	[9,40]	[1,26]
3	no_stay	[0,1)	[1,38]	[0,1)	[1,3)	[9,40]	[0,1)
4	no_stay	[0,1)	[0,1)	[0,1)	[3,43]	[0,4)	[0,1)
5	no_stay	[0,1)	[0,1)	[0,1)	[1,3)	[0,4)	[0,1)
6	no_stay	[0,1)	[0,1)	[0,1)	[3,43]	[0,4)	[0,1)

6 rows

Table 1: Result of head of the new dataset

Summary of transactions

28363 rows (elements/itemsets/transactions) and

41 columns (items) and a density of 0.36

Modeling

We are going to find important association rules for created data set. The association rule mining is usually performed on the data using the Apriori algorithm which makes it easy to find rules.

Rule

A rule is a notation that represents which attribute datasets is frequently used with what attribute dataset. Formulae are shown as below [4]

Attribute set A => Attribute set B

i.e. the attribute set on right is most frequently used along with the attribute set on the left.

The Apriori algorithm generates the most relevant set of rules from given data. It also shows support, confidence and lift of those rules. These three measures can be used to decide the relative strength of the rules shown below.

Let's consider the rule **A => B** in order to compute the following metrics.

$$Support = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} = P(A \cap B)$$

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P(A \cap B)}{P(A)}$$

$$Lift = \frac{Confidence}{Expected\ Confidence} = \frac{P(A \cap B)}{P(A).P(B)}$$

We use the Apriori algorithm to generate the rules using the generated transactions. We will generate frequent item-sets and rules on the dataset and on the subset of the dataset which includes the patients who had at least one claim on injuries. Based on the generated rules and frequent item-sets we compare the results for both the datasets.

Frequent item sets

Frequent item sets have been generated using the Apriori algorithm on the dataset. Figure 7 shows the relative frequency of the items in the item sets. Based on the bar plot it is obvious that the frequent item sets are based on the claims least made like renal, cancer, and heart. The facts like most of the patients do not stay in the hospital, majority of the patients are female, majority of the patients are healthy as CharlsonIndex is in the range [0, 1), and around 40% of the patients made at least 3 claims based on injuries are preserved.

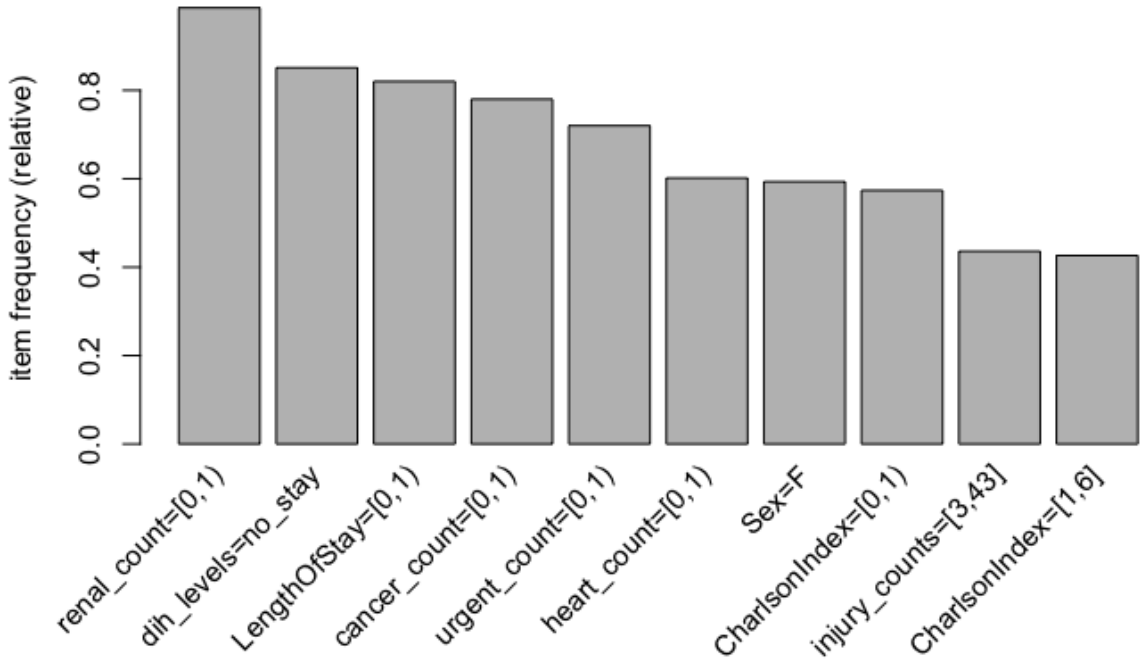

Figure 7: Bar plot showing relative frequent item sets

Now, we set the minimum support to 0.003, time to the maximum time limit and length of frequent item sets as the total number of items in the transactions and generate the frequent item itemsets using the Apriori algorithm. The minimum support value was chosen based on the value that the item has happened on at least 100 patients. The following bar plot has been generated using the frequent itemsets generated. The plot looks like it is almost normally distributed. This is because of the lattice where there are more items in the center and less items as we move towards the top or bottom end of the lattice. The generated frequent itemsets has 2 million itemsets.

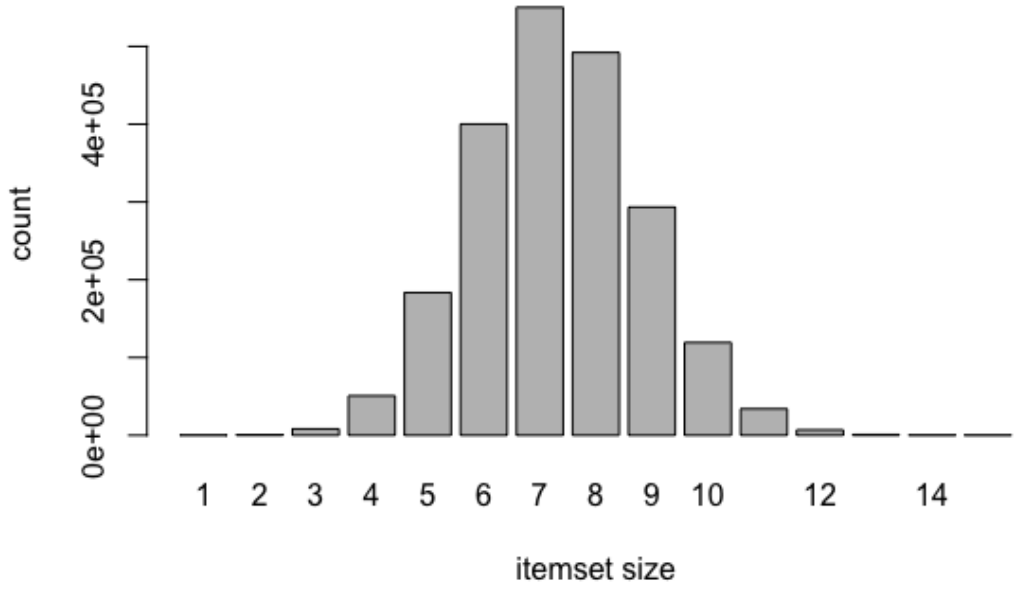

Figure 8: Bar plot showing the frequency of itemset size

Figure 9 compares the count of the itemsets generated for frequent, closed and maximal itemsets. The maximal itemset is a subset of closed itemset and the closed itemset is a subset of the frequent itemset.

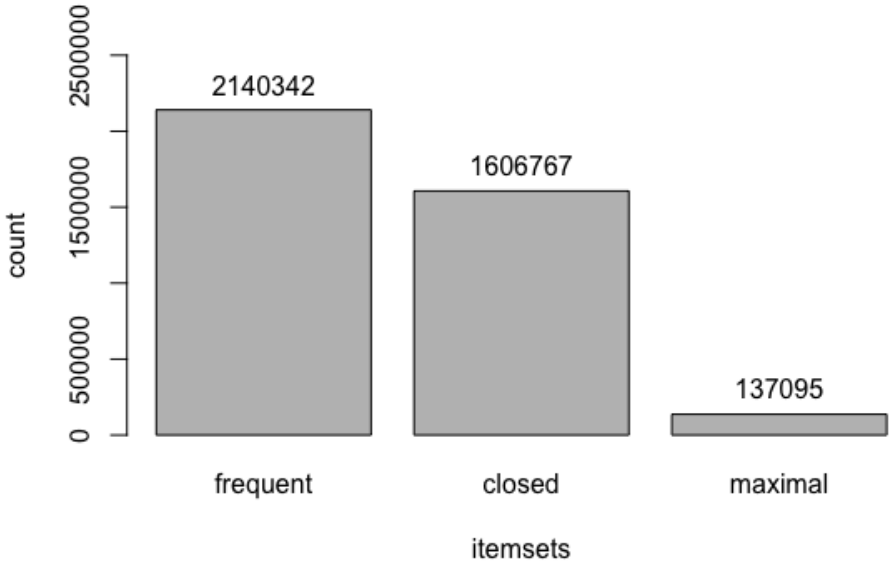

Figure 9: Bar plot showing the frequency of itemset size

The frequent items show patient groups from older age group and none of them stayed in the hospital. Two of the three patient groups could be unhealthy and did not make claims based on renal, cancer or urgency. One of the patient groups has a high drug count.

```
    items                                 support count
[1] {AgeAtFirstClaim=[54.5,74.5),
     Sex=M,
     PayDelay=[0,41.1),
     CharlsonIndex=[1,6],
     LengthOfStay=[0,1),
     dih_levels=no_stay,
     renal_count=[0,1),
     cancer_count=[0,1),
     heart_count=[0,1),
     urgent_count=[0,1)}          0.006804640   193
[2] {AgeAtFirstClaim=[54.5,74.5),
     Sex=F,
     CharlsonIndex=[0,1),
     LengthOfStay=[0,1),
     DrugCount=[19,84],
     dih_levels=no_stay,
     renal_count=[0,1),
     cancer_count=[0,1),
     heart_count=[0,1),
     urgent_count=[0,1)}          0.006522582   185
[3] {AgeAtFirstClaim=[74.5,85],
     Sex=M,
     CharlsonIndex=[1,6],
     LengthOfStay=[0,1),
     LabCount=[1,6),
     dih_levels=no_stay,
     renal_count=[0,1),
     cancer_count=[0,1),
     urgent_count=[0,1)}          0.006487325   184
```

Figure 10: Inspecting First 3 frequent item sets from the maximal frequent item set sorted by support

Now, let's examine frequent item sets for only unhealthy patients by excluding the counts that equal to 0 or length of stay is 0 or CharlsonIndex is 0. Figure 11 shows a bar plot of relative frequency items after exclusion of these items. The majority frequent items contain the patients who are female, their claims based on injury is greater than 3, they are older, made a lot of claims in a year, and have high drug and lab counts too.

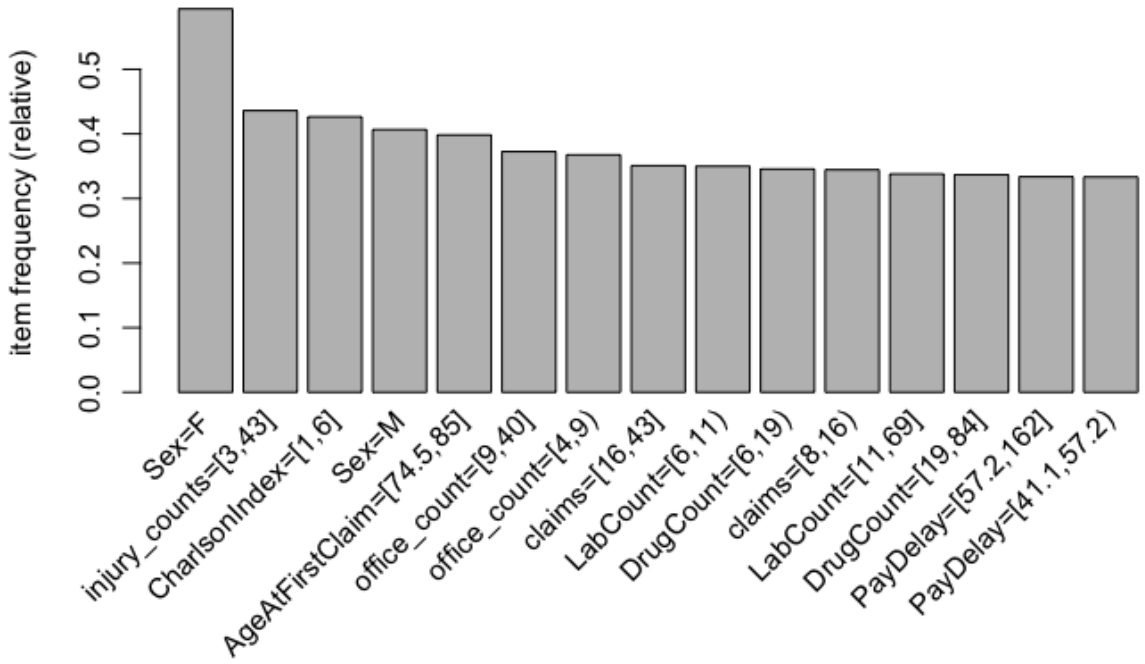

Figure 11: Bar plot showing relative frequency of items after excluding some items.

After exclusion of items related to healthy patients, we see that many itemsets have been dropped and the frequent item itemset is almost equal to closed item itemset.

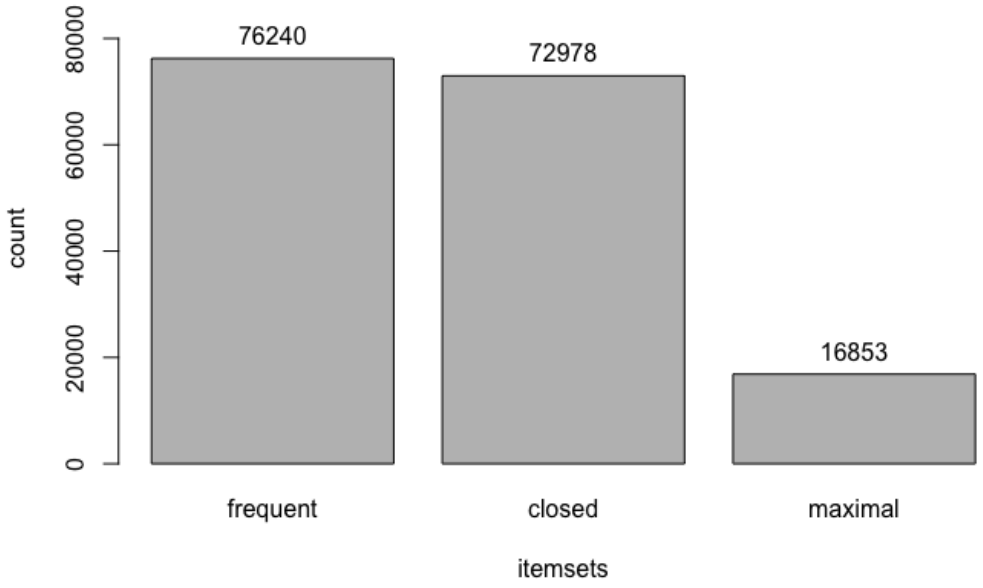

Figure 12: Bar plot showing the frequency of frequent, close and maximal itemsets after exclusion of items

Association Rules

A support of 0.003 has been used to create the association rules as mentioned above. Total of 4.6 million rules were generated by the Apriori algorithm based on the generated transaction set. The following table shows the result of inspect method applied of the first 6 rows of the rules created.

lhs <fctr>	<fctr>		**rhs** <fctr>	**support** <dbl>	**confidence** <dbl>	**Lift** <dbl>	**Count** <dbl>
[1]	{}	=>	{LengthOfStay=[0,1)}	0.82	0.82	1.00	23263
[2]	{}	=>	{dih_levels=no_stay}	0.85	0.85	1.00	24141
[3]	{}	=>	{renal_count=[0,1)}	0.98	0.98	1.00	27985
[4]	{renal_count=[1,20]}	=>	{CharlsonIndex=[1,6]}	0.01	0.98	2.30	371
[5]	{LengthOfStay=[2,474]}	=>	{renal_count=[0,1)}	0.05	0.98	0.99	1659
[6]	{dih_levels=long_stay}	=>	{renal_count=[0,1)}	0.06	0.96	0.96	1714

Association rules sorted based on lift

```
lhs                rhs                support                confidence lift     count
[1] {claims=[16,43],
    PayDelay=[0,41.1),
    CharlsonIndex=[1,6],
    LengthOfStay=[0,1),
    heart_count=[0,1),
    office_count=[0,4)}   => {ClaimsTruncated} 0.003596235  0.9807692 11.16723   102
[2] {claims=[16,43],
    PayDelay=[0,41.1),
    CharlsonIndex=[1,6],
    LengthOfStay=[0,1),
    renal_count=[0,1),
    heart_count=[0,1),
    office_count=[0,4)}   => {ClaimsTruncated} 0.003596235  0.9807692 11.16723   102
[3] {claims=[16,43],
    PayDelay=[0,41.1),
    CharlsonIndex=[1,6],
    heart_count=[0,1),
    injury_counts=[3,43],
    office_count=[0,4)}   => {ClaimsTruncated} 0.003560977  0.9805825 11.16510   101
[4] {claims=[16,43],
    PayDelay=[0,41.1),
    CharlsonIndex=[1,6],
    heart_count=[0,1),
    office_count=[0,4),
    urgent_count=[0,1)}   => {ClaimsTruncated} 0.003560977  0.9805825 11.16510   101
[5] {claims=[16,43],
    PayDelay=[0,41.1),
    CharlsonIndex=[1,6],
    LengthOfStay=[0,1),
    LabCount=[11,69],
    heart_count=[0,1),
    office_count=[0,4)}   => {ClaimsTruncated} 0.003560977  0.9805825 11.16510   101
```

We can see that it has generated some rules which around about 100 transaction obeying them. All these transactions have the claims between 16 and 43. That is, most of the patients have higher claims in a year.

As observed from the frequent item sets, the rules have length of stay as 0 and the count variables as 0. Looks like these set of patients are also unhealthy as their CharlsonIndex is not 0. Their office count is between 0 and 4, which could probably mean that these patients might eight have been treated well and they became healthy as only the latest CharlsonIndex of the patients were considered. A patient might not visit the hospital again after he turns healthier. It is however possible, that the patient could have been treated elsewhere say Inpatient Hospital or ambulance etc. The fifth group of patients had a high lab count. This is similar to the cluster which had high LabCounts and claims when working on the clustering project. This group of patients could be suffering from diabetes and required regular lab checkups. All these rules generate claims truncated as RHS. We will try to examine the rules without claims truncated.

```
   lhs                rhs              support     confidence lift   count
[1] {claims=[1,8),
    LabCount=[6,11),
    cancer_count=[0,1),
    injury_counts=[1,3),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.003525720  0.9708738 3.733816 100
[2] {claims=[1,8),
    LabCount=[6,11),
    renal_count=[0,1),
    cancer_count=[0,1),
    injury_counts=[1,3),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.003525720  0.9708738 3.733816 100
[3] {claims=[1,8),
    CharlsonIndex=[0,1),
    LabCount=[6,11),
    injury_counts=[1,3),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.003419949  0.9700000 3.730456  97
[4] {claims=[1,8),
    CharlsonIndex=[0,1),
    LabCount=[6,11),
    renal_count=[0,1),
    injury_counts=[1,3),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.003419949  0.9700000 3.730456  97
[5] {Sex=M,
    claims=[1,8),
    CharlsonIndex=[0,1),
    LabCount=[6,11),
    dih_levels=no_stay,
    urgent_count=[1,26]} => {office_count=[0,4)} 0.003314177  0.9690722 3.726887  94
```

Now after the claims truncated variable is removed, we can now see that office_count becomes the target for RHS as it has higher frequency too. But the groups of patients look similar to the ones created above. However, urgent count takes a place in the LHS. This means that if a person makes a claim, he visits the hospital. But it could have been urgent care or ambulance. It is possible that different claims could have been recorded for office visit and the ambulance or any other urgent care service. The patients are healthier group of patients who have CharlsonIndex of 0 unlike the other case where ClaimsTruncated was considered. They have higher LabCounts and InjuryCount too. May be the patients with too high claim rate have claims truncated set as true.

Now, let's consider the transactions that use only the unhealthy patient's data. Only 36000 rules have now been created based on these transactions unlike millions in the previous case.

```
   lhs                rhs              support     confidence lift   count
[1] {claims=[1,8),
    LabCount=[6,11),
    injury_counts=[1,3),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.004054578  0.9663866 3.716559 115
[2] {AgeAtFirstClaim=[4.5,54.5),
    claims=[1,8),
    DrugCount=[1,6),
    LabCount=[6,11),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.003596235  0.9532710 3.666119 102
[3] {Sex=M,
    claims=[1,8),
    LabCount=[6,11),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.004230864  0.9523810 3.662696 120
[4] {AgeAtFirstClaim=[4.5,54.5),
    claims=[1,8),
    LabCount=[6,11),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.005253323  0.9490446 3.649865 149
[5] {claims=[1,8),
    DrugCount=[1,6),
    LabCount=[6,11),
    urgent_count=[1,26]} => {office_count=[0,4)} 0.005182809  0.9483871 3.647336 147
[6] {claims=[1,8),
    LengthOfStay=[1,2),
    LabCount=[6,11)}     => {office_count=[0,4)} 0.003878292  0.9401709 3.615738 110
```

Based on the above rules sorted based on lift measure, the most common RHS is office count of 0 through 3 which is the same case as before. All the 6 groups had high lab counts and two of them had low drug count which could be the case where patients had to do regular lab checkups. It's possible all the 6 groups of patients could be suffering from diabetes. The last group of patients had to stay in the hospital for a day. May be, they were placed under observation for a day or the patients could be pregnant except for the second group. Most of these groups look like they have one frequent itemset in common which is {claims=[1,8), LabCount=[6,11),urgent_count=[1,26]}

Now let us examine the rules when sorted based on the confidence.

```
   lhs                rhs                support     lift     count confidence
[1] {AgeAtFirstClaim=[4.5,54.5),
    Sex=M,
    PayDelay=[0,41.1),
    DrugCount=[1,6),
    LabCount=[1,6),
    injury_counts=[1,3),
    office_count=[0,4)} => {claims=[1,8)}    0.003490463  3.279339    99        1
[2] {LengthOfStay=[2,474],
    heart_count=[3,38],
    office_count=[9,40]} => {claims=[16,43]} 0.013856080  2.851412   393     1
[3] {LengthOfStay=[2,474],
    LabCount=[11,69],
    office_count=[9,40]} => {claims=[16,43]} 0.017381800  2.851412   493     1
[4] {LengthOfStay=[2,474],
    DrugCount=[19,84],
    heart_count=[1,3),
    office_count=[9,40]} => {claims=[16,43]} 0.003984064  2.851412   113     1
[5] {LengthOfStay=[2,474],
    LabCount=[11,69],
    heart_count=[1,3),
    office_count=[9,40]} => {claims=[16,43]} 0.003419949  2.851412    97     1
[6] {LengthOfStay=[2,474],
    heart_count=[1,3),
    injury_counts=[3,43],
    office_count=[9,40]} => {claims=[16,43]} 0.003772521  2.851412   107     1
```

The rules now have a confidence of 1, which means if RHS occurs together with LHS in all the transactions. Large group of patients turn up when sorted by confidence. These group of patients have pretty high claims. It is possible that these group patients could have the highest claims made in the year. These patients have high length of stay, drug count and lab count. Some of the groups also have higher newly created variables, heart counts and injury counts. Group 4-6 have similar items based on heart counts. It could be possible they could form subsets for a larger itemset containing all these items in these groups with high confidence (may not be 1).

Note: The inspectDT method was slow as there were a lot of rules being generated and hence, we just sorted the data based on lift and confidence and examined the rules.

Evaluation and Deployment

We have discussed about the generation of frequent item sets and association rules. It is interesting that the first few groups sorted based on confidence and lift were similar to the results generated during the clustering. A healthy cluster where there were less lab counts, drug counts and claims are similar to the groups generated using the association rules and frequent item sets. It is also possible that these group of patients made these claims based on pregnancy on which clustering was done and they seem to agree with each other. The health care providers based on the generated rules can decide how long a patient can stay based on their previous data and can have a clear idea for the beds to be made available for the patients. Based on the count variables like heart counts, injury counts, renal counts and office counts, the rules could be examined and certain equipment necessary for these conditions could be known and handled appropriately. The office counts could also give an idea about what kind of primary care physicians are required in the health industry. The insurance providers can also be profitable based on the analysis of the claims and the count variables and the cost estimates could be calculated and only the patients who have less cost estimates could be the primary focus (which is not correct and humane).

Conclusion

The frequent item sets generated were based mostly on the office counts and length of stay as 0. This was because this most common item in the transaction and had high frequency of claims with these items. However, the unhealthy patient analysis brought some interesting facts into foreground. The newly generated count variables also helped in generating the rules and were present in the top 5 rules based on confidence.

Reference

[1] https://www.healthcarefinancenews.com/news/change-healthcare-analysis-shows-262-million-medical-claims-initially-denied-meaning-billions

[2] https://www.cdc.gov/nchs/fastats/physician-visits.htm

[3] https://www.cdc.gov/niosh/topics/ems/data.html

[4] http://r-statistics.co/Association-Mining-With-R.html