Southern Methodist University

CSE -7331: Introduction to Data Mining

Dr. Michael Hahsler

Project 1: Data and Visualization

Dmonty, Sevil Sanjao, 47568070
Pritheesh Panchmahalkar, 47524741

Contents

EXECUTIVE SUMMARY

In this project, we are going to analyze the Heritage Health prize hospital data. Every year around 70 million admissions are registered in the U.S. Around $30 million is spent unnecessarily on hospital admissions. Hence, we need to find a better way to ensure that the patients that are at risk get the right treatment with less expenses. Data mining helps us find relationships between various attributes in the data set and visualize them. The visualizations give an idea on what is happening and what could have gone wrong in the process of the hospitalization. We find most important attributes in the data that could help us predict the health status of a patient and cure the disease. The data helps various stakeholders like medical researchers, hospital healthcare providers and patients. Medical researchers use data mining to analyze the data which could help them refine the medical procedures which will in turn benefit patients with better medication. Hospital healthcare providers can use data mining techniques to accurately the most viable medical equipment's to be used. After analyzing the patient treatment data and their admission rate, administrators can use this data to upgrade hospital facilities. Patients these days do a lot of research before visiting a hospital because of growing healthcare costs. They analyze the best available insurance rate through easily available software's which use hospital treatment costs using data mining. And doctors can use the historical data to see patient family history for any significant health conditions and use that to treat them effectively.

Hence, we can say that, data mining techniques are being used in healthcare system across the globe.

INTRODUCTION

In 2016, America's hospitals treated 143 million people in their emergency departments, provided 605 million outpatient visits, performed over 27 million surgeries and delivered nearly 4 million babies. Every year, hospitals provide vital health care services like these to hundreds of millions of people in thousands of communities (aha.org report 2018-06).

For this project we have data from the Heritage Health. The Heritage Health prize was a competition for individual who want to gain more experience with data mining techniques. According to the American Hospital Association every year more than 71 Million individuals in the United states are admitted to hospital every year (HPN, 2012), studies have concluded that in the year 2006 well over $ 30 Billion was spent on unnecessary hospital admissions. By analyzing provided data we are going to see how this given data could be used to predict future about the members or claims files by members,

- how vendors are paying for claims?
- Is there any pay delay?
- Under What specialty member got admitted?
- What is the ratio of gender-based hospitalization?
- What is the readmission rate?

When a person falls sick he/she go to doctor/hospital, the doctor treats the sick person and for their provided medical services, the doctor/Hospital submits the medical billing insurance claim. In other words, the medical billing insurance claims process starts when a healthcare provider treats a patient and sends a bill of service to the medical insurance providing company (atena, IHC group, humana etc..). After submission of claim insurance provider then check the claim based on the plan on which the member has enrolled. As payer the insurance provider carefully reviews the documents for covered services and after review they release payment to the healthcare services provider(doctors/Hospital). By this one can understand that this is pretty straight forward process but in reality, it is not that straight forward. While submitting claims you could use one of the two available options by manually (on paper) or Electronically (by implementing proper HIPAA Transaction and code set rule). A study report says that on can save around 3$ per claim if filed electronically.

Why Data mining?

Electronic health records (EHR) are quickly becoming more common among healthcare facilities. With increased access to a large amount of patient data, healthcare providers can now optimize the efficiency and quality of their organizations using data mining. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. This could be a win/win overall. But due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining and analytic strategies.

Like analytics and business intelligence, the term data mining can mean different things to different people. The most basic definition of data mining is the analysis of large data sets to discover patterns and use those patterns to forecast or predict the likelihood of future events.

That said, not all analyses of large quantities of data constitute data mining. We generally categorize analytics as follows:

- Descriptive analytics—Describing what has happened
- Predictive analytics—Predicting what will happen
- Prescriptive analytics—Determining what to do about it

It is to the middle category—predictive analytics—that data mining applies. Data mining involves uncovering patterns from vast data stores and using that information to build predictive models.

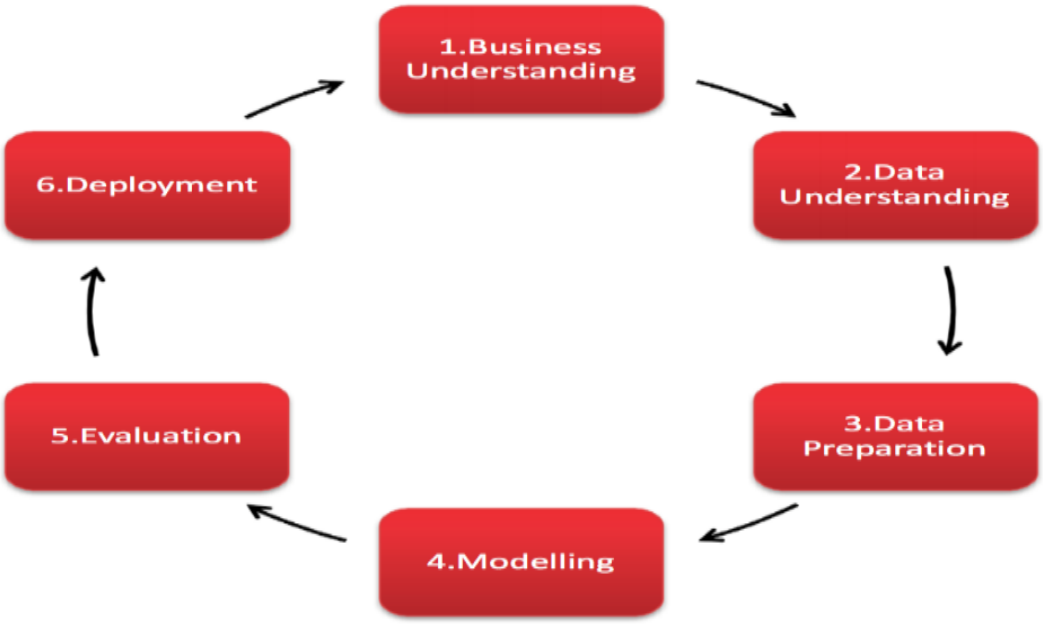
Many industries successfully use data mining. It helps the retail industry model customer response. It helps banks predict customer profitability. It serves similar use cases in telecom, manufacturing, the automotive industry, higher education, life sciences, and more.

However, data mining in healthcare today remains, for the most part, an academic exercise with only a few pragmatic success stories. Academicians are using data-mining approaches like decision trees, clusters, neural networks, and time series to publish research. Healthcare, however, has always been slow to incorporate the latest research into everyday practice.

CRISP-DM Model

As we follow the CRISP-DM framework for this project, this part will cover first three elements in the framework, business understanding, data understanding and data preparation. Since the project is term project we try to customize the framework and select what can be achievable during the term period, so we not follow every mini step mention in framework such as create terminology or construct a cost-benefit analysis.

Figure 1: CRISP-DM framework



BUSINESS UNDERSTANDING

The First concept in CRISP DM model is problem Business Understanding, we are going to ask our self some questions and try to understand how we can answer those questions.
How can we extract useful knowledge from provided data?
Who will be more interested in data?
What kind of prediction would be made using data?
How can we relate relationships with data?

1. How can we extract useful knowledge from provided data?

Here we have data which includes csv files named Claims_Y1.csv, Members_Y1.csv, DayInHospital_Y2.csv and Lookup PrimaryConditionGroup.csv. We will discuss some of the important columns of each files in detail in next section. It is very important to understand your data to extract useful information from it. Data mining is a great technique we will use to extract this useful knowledge and will see how we can utilize this information for predict health issues member will have, members readmission rates, what should be the research area for medical study to improvise existing medicinal practices or design new medication procedures etc.

2. Who will be more interested in data?

Data contains information about members who have claims, hospitals where member visited or got admitted, claims submitted by hospitals to vendor behalf of members, delay in payment by vendors, specialty under which member got admitted and so on, we will see type of data in next section. Above section gives some hint about the person of interest in these data, such as members, vendors, hospitals etc. we will now consider why members will be interested in data.
Member's interest: - Members as patients would be more interested in the data as they can have knowledge about the vendors, providers, physicians who could provide better service at reasonable prices. Member of particular insurance provider want to know some of the following question they have such as,

1. Member want to know which hospital cover his/her insurance or what are the hospitals under their insurance providers network?
2. How much payment delay is there from insurance provider.
3. What specialist does hospital provide.
4. Is the treatment covered in their plan?

Insurance companies interest: - Insurance companies will also be interested in data for number of reasons such as they might be interested in members having less readmission rate as well as members with less claims.

Provider interest: - provider of health services might be interested in members with both high and low claim rates as well as specific members with their specialty requirement. Health care providers can focus their research on Primary Condition Groups that are common among patients or that are life-threatening and improvise medical techniques.

3. What kind of predictions could be make?

By using given data, we could make few predictions such as what kind of treatment a member is getting from hospital. It also helps us predict about which member might be hospitalize in coming year. What will be the health condition of particular member will it be better or will it worst? We will also predict updated charlson index value. important of all based on gender-based analysis we might be able to predict how healthy each gender-based member will be.

4. How can we establish relationships with data?

As describe earlier we use data mining for predictive analysis which is possible only by making relations between available data. The data which is available with us contains information about insurance member and their hospitalization records and claims etc. this data

DATA UNDERSTANDING

Members Data

MemberID	Integer	nominal	Member pseudonym.
AgeAtFirstClaim	Factor(string)	Ordinal	Age in years at the time of the first claim's date of service computed from. This has values between 0-9, 10-19, 20-29 ….to 80+
sex	Factor	Nominal	Sex Biological sex of member: M = Male; F=Female.

Claims Data

MemberID	Integer	Nominal	Member pseudonym.
ProviderID	Integer	Nominal	Provider pseudonym. This is a ID of the doctor or specialist who providing services to member.
Vendor	Integer	Nominal	Vendor pseudonym. This is an ID of company who is issuing bill to the member
PCP	Integer	Nominal	Primary care physician pseudonym. It is members primary care physician.
Year	Factor (String)	Ordinal	Year in which the claim was made: Y1; Y2; Y3.
Specialty	Factor (String)	Nominal	Generalized specialty.
PlaceSvc	Factor (String)	Nominal	Generalized place of service. This have values such as Ambulance, urgent care, Office etc. total 8 place of services listed.
PayDelay	Integer	Ratio	Number of days delay between the date of service (the date the actual procedure was performed, or service provided) and date of payment. Values above 161 days (the 95% percentile) are top-coded as "162+".

LengthOfStay	Factor (String)	Ordinal	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).
DSFS	Factor (String)	Ordinal	Days since first claim, computed from the first claim for that member for each year, generalized to: [0-1] month, (1-2] months, (2-3] months, (3-4] months, (4-5] months, (5-6] months, (6-7] months, (7-8] months, (8-9] months, (9-10] months, (10-11] months, (11-12] months.
PrimaryConditionGroup	Factor (String)	Nominal	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates, that generalize the primary diagnosis codes total 45 conditions listed.
CharlsonIndex	Factor (String)	Ordinal	A measure of the affect diseases has on overall illness, grouped by significance, that generalizes additional diagnoses. Scores greater than zero are carried forward (for up to a year) in subsequent claims with a comorbidity score of zero [1,4].

PayDelay has 44623 number of NA's (missing values).

DayInHospital Data

MemberID	Integer	Nominal	Member pseudonym.
DaysInHospital_Y2	Integer	Ratio	Number of days spent in the hospital in the year Y2.

Primary Condition Group

PrimaryConditionGroup	Factor(String)	Nominal	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates, that generalize the primary diagnosis codes total 45 types listed.
Description	Factor(String)	Nominal	Gives a general description of the PrimaryConditionGroup

Quality of the data

1. There is missing data for the attribute length of stay which is nearly 95.83% of data, which we think could be missing by mistake in the length of stay attribute since there are some members with values such as 1 day, 2 days till 7 days and after that 1-2 weeks up to 26+ weeks. The missing data could be because the patients were in the office or in an ambulance for a surgery or similar place of service, where the patient spends less time and may not have been recorded.
2. There are missing values for attributes ProviderID (0.605 %), Vendor (1.006 %), Primary care Physician (PCP - 0.251%), Paydelay (6.921 %).
3. We noticed while analyzing relation of gender and primary conditions, there are males that were suffering from primary conditions which could be found in females such as gynecological issues and breast cancer which could be a result of wrong documentation.
4. We do not have any duplicate values in given data.

Simple appropriate statistics

1. 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 vendor (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.

Classification of claims of the patients based on Charlson Index

Charlson Index	0	1-2	3-4	5+
Frequency	369191	263524	10780	1211

An appropriate statistic for the Charlson Index would be mode as the scale of the data is ordinal. Mode for this attribute is 0, as the frequency of claims for Charlson Index '0' was the highest.

2. Sex:

Based on the data, the value of sex could either be 'M' for Male or 'F' for female. The frequency of claims made by patients of different sex, the vendors (insurance companies) can make more profit if focused on patients who have less claims. There could be few conditions that are specific/common to a sex which could help the doctors understand what disease the patient is likely to suffer from based on their sex. Females are more likely to visit the hospitals more often after certain age and based on common Primary Condition groups when compared to males. "Females between ages 19-44 spent 66 percent more per capita in 2012 than did males in the same age-group. The significant difference in spending is largely associated with the costs for maternity care, and females spending over 46 percent more than males on retail prescription-drugs." This shows that females are more health-conscious when compared to males and visit the hospitals more often assuming the percentage of males and females in the population is nearly the same. Based on the members table, the total females are 54.9% and males are 45.1%. The percentage of females is relatively higher than that of males but not by a large value i.e., a difference of about 8000 in the male and female members vs a difference of 113000 for the claims of males and females.

	Males	Females
Percentage	41.21925%	58.78075%

The percentage of claims for females is relatively higher than males. The reason could be that the females may visit the hospital for regular checkups more often. Sex is a nominal attribute and mode would be an appropriate statistic. Mode for the claims data is 'F', as females accounted for the major percentage of the claims. There are no missing values with the attribute sex.

3. **Place of service**

Place of service gives general information on where the service was provided to the patient. This attribute has 8 places where the service has been provided according to the data. The places of services are Office, Independent Lab, Outpatient, Hospital, Urgent Care, Inpatient, Hospital, Other, Home, and Ambulance. Vendors will be less interested in the place of service that is expensive. Based on how the services are provided, the patients would be more interested in the hospitals or health care providers that have better and less expensive services. Place of service has on impact on how much the patients are willing to pay to the health care providers. "According to OIG, between 2010 and 2012, Medicare overpaid physicians by some $33 million based on an incorrect POS being listed on claims. In many cases, this was because the physician performed the procedure in a facility location (i.e. an ASC) yet identified the place of service as non-facility (usually as their office)". Hence, it is important not only to use the correct place of

service but also record it correctly to avoid the overpay or underpay. The appropriate statistic for the place of service is **mode** which is 'Office' which could be a primary consultation or a day visit with the doctor at their office.

4. **Pay Delay**

Pay delay is the delay between the claim and the day the claim was paid for. Pay delay is an integer which gives the value in days, the duration between the day of claim and the day of payment for the claim. There could be several reasons why payment gets delayed. Table below gives summary of pay delay attribute, which has minimum 0-day delay to maximum 161 days delay. The missing data is denoted as NA's.

The mean of the data is 46.59 which means that, on an average the payment is done within a month and a half. The missing values have been ignored for the calculation of the mean and median. The description of the data mentions that the values above 161 days (the 95% percentile) are top-coded as "162+". The value "162+" does not appear in the data. It is highly possible that NA's could be 162+. But if it is, the actual mean will account to a value much larger when compared to the mean which doesn't consider the NA's and would be difficult to calculate.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	28.00	37.00	46.59	58.00	161.00	44623

If the NA's are replaced with 162 (an approximation of value), then the summary of the pay delay is

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	28.00	37.00	46.59	58.00	161.00

5. Primary Condition

This is one of the important attributes as per the health care provider and insurance provider as this attribute provides details under which primary condition, a patient has been admitted. This attribute has 45 unique values (different types of disease). Primary condition is a nominal attribute and hence mode is an appropriate statistic. The mode for the Primary condition group is 'MSC2a3' which is around 17% of the total claims.

Data Visualization

1. Charlson Index:

The scale of Charlson Index data is ordinal. Hence, to visualize the Charlson Index, a bar plot is plotted with Charlson Index on X-axis and Frequency of claims on Y-axis shown in Figure 2. The above visualization gives clear idea of the frequency of the claims of patients based on Charlson Index. The claims for Charlson Index of 0 and 1-2 together constitute a little more than 98% of the total claims. The percentage of claims for Charlson index value of 5+ is almost 0.

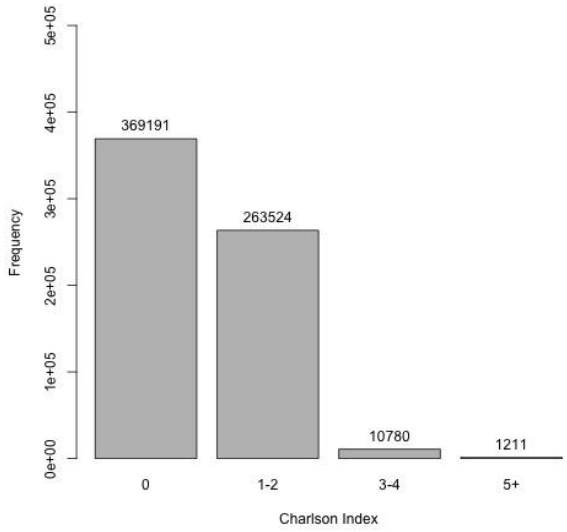


Figure 2: Charlson Index vs Frequency of claims

2. Sex :

A bar plot is chosen for the visualization of the sex attribute of the claims data, as it is a nominal attribute. Figure 3 depicts a bar plot that is plotted showing sex on X-axis and the frequency of claims on Y-axis. The percentage of females is relatively higher when compared to males based on the plot. The claims of females are around 113K higher than that of males.

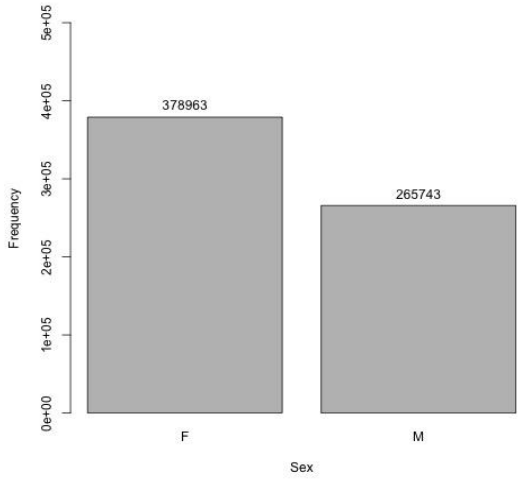


Figure 3: Sex vs Frequency of claims

3. Place of Service

A bar plot has been plotted to visualize the place of service since it is a nominal attribute. Figure 4 shows a barplot with different places of service plotted on X-axis and frequency of claims plotted on Y-axis. Among 8 places of service, more than 60% of the times, 'Office' has been used as the place of service for medication. It is highly unlikely that a medication could be at 'Home' based on the data.

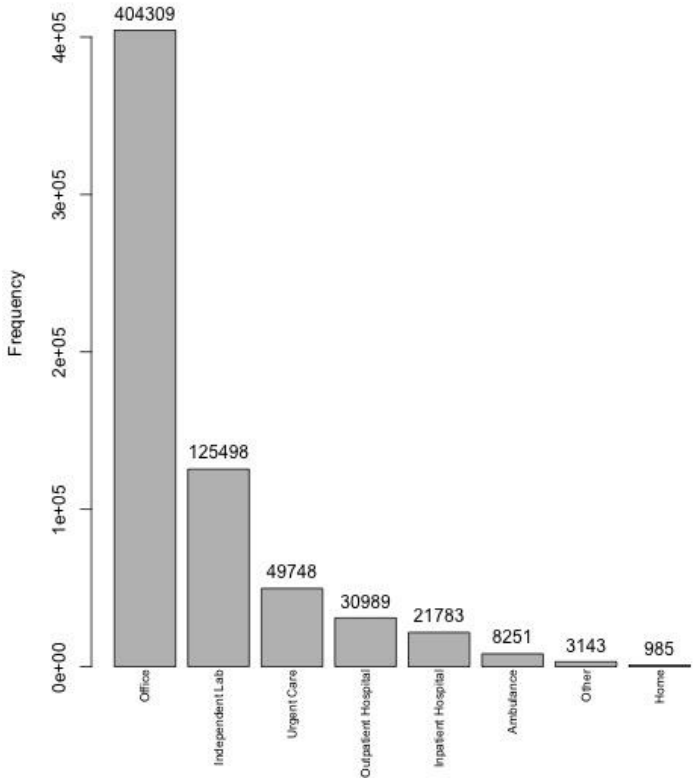


Figure 4: Place of Service vs Frequency of claims

4. Days Since First Service:

To visualize the dsfs (Days Since First Claim) attribute of the data, a bar plot has been chosen, as dsfs is an ordinal attribute that takes any of the 12 values. Figure 5 shows a bar plot with the period of days since first service on X-axis and the frequency of claims on Y-axis. Around a quarter of claims were made in a duration of less than a month since first claim. The frequency of claims keeps decreasing as the duration increases.

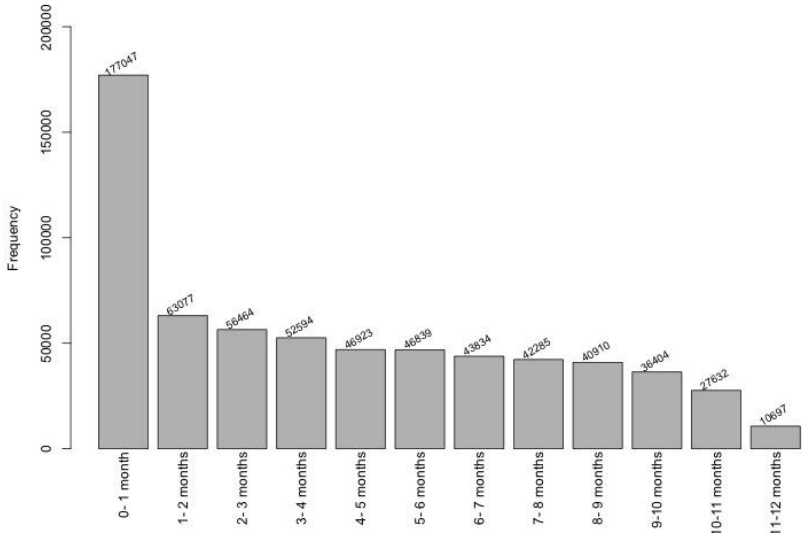


Figure 5: Days since first service vs Frequency of claims

5. Pay Delay:

The measure of scale for the pay delay attribute is ratio and hence, a histogram has been plotted to visualize the attribute. Figure 6 is a histogram plotted with pay delay on X-axis and frequency of claims made on Y-axis. Most of the payments were made between 30 to 50 days which could be standard payment terms by vendor after verifying all documentation about claims filed by members. Also, there are several payment delays lasts up to 161 days which might be because of incorrect documentation or uncovered tests made by doctor on members or some providers do not charge until the patient is treated which is unlikely.

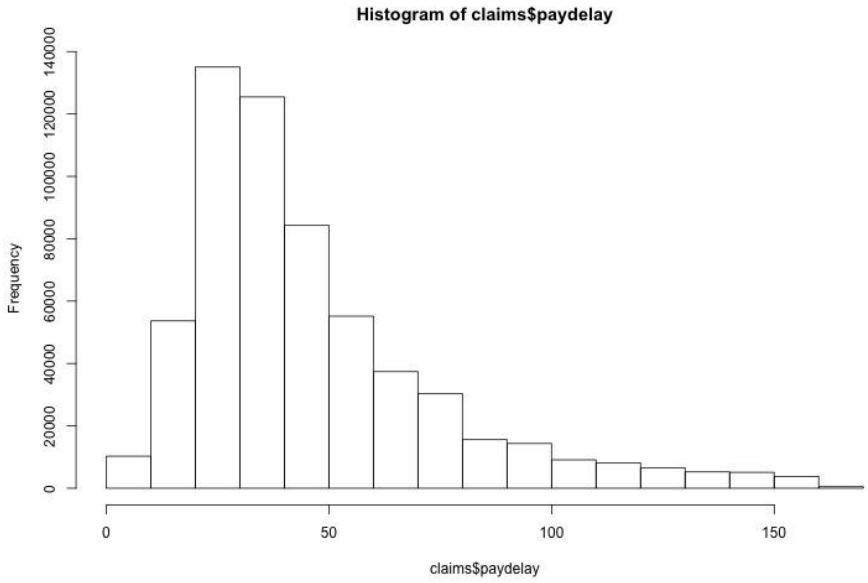


Figure 6: Pay delay vs Frequency of claims

6. Primary Condition group

A bar plot has been plotted to visualize the Primary Condition group as it is a nominal attribute. Figure 7 shows a bar plot of top 15 Primary Condition groups sorted based on the frequency of claims with Primary Condition groups on X-axis and frequency of claims on Y-axis. The graph gives us detailed distribution of members by their primary condition, it will also help researchers/scientists focus their research on the most common primary conditions and develop medicine. Doctor will get help by identifying members by their specialty requirement and also helps insurance provider to understand which member will submit claim more frequently. Around 40% of the claims were based on the Primary Condition Groups like 'METAB3', 'MSC2a3' and 'ARTHSPIN'.

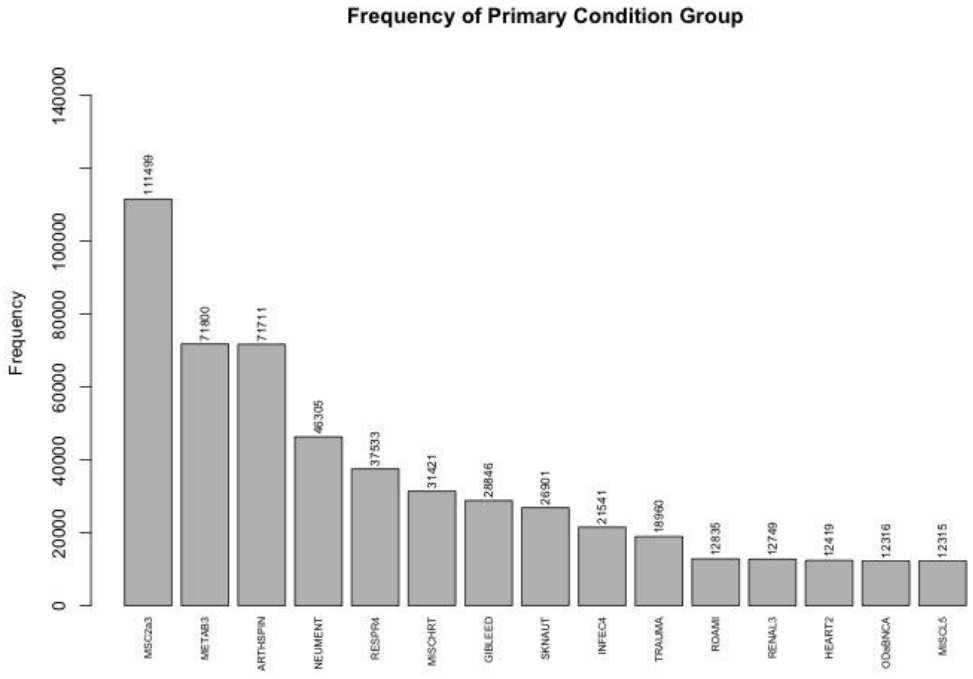


Figure 7: Primary Condition Group vs Frequency of claims

Relationships:

Charlson Index and dsfs

The table below shows relation between charlson index and days since first service. This relationship helps us understand how many claims were filed monthly basis with the range of charlson index. The range of DSFS got decreased as range of charlson index increases. This means that most of the patients do not have a life-threatening condition, based on Charlson index. Since, most of the members have repeated claims, it could also mean that the patient's health improved from higher Charlson index to lower Charlson Index as they are being treated. Table 1 shows the cross tabulation of frequency of claims and Days since first service.

DSFS	Charlson index			
	0	1-2	3-4	5+
0- 1 month	137933	37571	1416	127
1- 2 months	38886	23248	849	94
10-11 months	10084	16741	712	95
11-12 months	3445	6849	366	37
2- 3 months	31524	23798	1019	123
3- 4 months	27931	23617	959	87
4- 5 months	24215	21674	921	113
5- 6 months	23280	22560	876	123
6- 7 months	20780	22032	955	67
7- 8 months	19603	21757	823	102
8- 9 months	17368	22466	949	127
9-10 months	14142	21211	935	116

Table 1: Charlson Index vs DSFS

Pearson's Chi-squared test
data: CharlsonIndex and DSFS
X-squared = 53710, df = 33, p-value < 2.2e-16
The p-value is very less, hence we can reject null hypothesis and confirm that the CharlsonIndex and DSFS are not independent of each other, hence related.

Sex and Charlson Index

The relationship between gender and charlson index give us number of members with their relative sex and criticality of their diseases. Table 2 shows number of females and males based on their charlson index. We can see the relatively very few numbers of members have critical health issues, but we can see that index value 5+ has almost same value of both gender members followed by index 3-4. With the help of the table we also able to see that females are healthier than males even though there are more females than males.

	0	1-2	3-4	5+
F	228559	144724	5072	608
M	140632	118800	5708	603

Table 2: Charlson Index vs Sex

Pearson's Chi-squared test

Data: Sex and CharlsonIndex

X-squared = 3761.5, df = 3, p-value < 2.2e-16

The p-value is very less, hence we can reject null hypothesis and confirm that the CharlsonIndex and sex are not independent of each other, hence related.

Sex Vs. primary group

Male and female are alike in many ways, however there are important biological and behavioral differences between the two genders. They affect manifestation, epidemiology and pathophysiology of many widespread diseases and the approach to health care. Despite our knowledge of these crucial differences, there is little gender-specific health care; the prevention, management and therapeutic treatment of many common diseases does not reflect the most obvious and most important risk factors for the patient: sex and gender. To discuss and address properly the differences in health and health care between men and women, it is necessary to distinguish between sex and gender and their respective effects on health. Sex differences are based on biological factors. Table below shows gender based primary conditions, this relationship helps us understand total number of members distributed by their gender (M and F) and with their primary conditions. While checking data we come across primary condition GYNES1 and GYNECA which are Female illness but there are some male members which we think data mistake.

Primary Condition	F	M	Primary Condition	F	M	Primary Condition	F	M
AMI	3155	5736	HEART2	6194	6225	PERVALV	393	457
APPCHOL	2425	2616	HEART4	3686	3391	PNCRDZ	161	98
ARTHSPIN	44387	27324	HEMTOL	3944	2250	PNEUM	1313	1268
CANCRA	509	592	HIPFX	780	243	PRGNCY	7649	184
CANCRB	2452	7287	INFEC4	12163	9378	RENAL1	53	71
CANCRM	169	62	LIVERDZ	392	358	RENAL2	760	1343
CATAST	240	205	METAB1	534	490	RENAL3	3916	8833
CHF	1588	1584	METAB3	39634	32166	RESPR4	21384	16149
COPD	6408	5298	MISCHRT	18128	13293	ROAMI	6806	6029
FLaELEC	825	422	MISCL1	786	515	SEIZURE	3084	2078
FXDISLC	4605	4077	MISCL5	7451	4864	SEPSIS	65	55
GIBLEED	17860	10986	MSC2a3	70029	41470	SKNAUT	15706	11195
GIOBSENT	1592	1239	NEUMENT	27586	18719	STROKE	1153	996
GYNEC1	10818	416	ODaBNCA	7898	4418	TRAUMA	9966	8994
GYNECA	2349	20	PERINTL	98	63	UTI	7869	2286

Table 3: PrimaryConditionGroup vs Sex

Pearson's Chi-squared test
testdata: sex and primary condition group
X-squared = 29475, df = 44, p-value < 2.2e-16
The p-value is very less, hence we can reject null hypothesis and confirm that the sex and PrimaryConditionGroup are not independent of each other, hence related.

Primary group and Charlson Index:

When person visit their primary care physician it is not the case that the individual suffering from one specific illness, that individual might have multiple symptoms also knows as multimorbidity. There is increasing interest in the concept of multimorbidity, which is the co-occurrence of multiple diseases or primary conditions within 1 person. The multimorbidity is important in case such as primary care where family practitioners (PCP) act as the first point of contact for people with wide range of conditions. The table below shows relation between primary group and Charlson index. Based on this we can see that a member could have multimorbidity with Charlson index value 0 but as the Charlson index values goes higher value on primary condition reduces dramatically. The table 2 shows data for 15 PrimaryConditionGroups that had the most claims.

	0	1-2	3-4	5+
MSC2a3	73323	36664	1326	186
METAB3	25377	43698	2618	107
ARTHSPIN	47116	23687	824	84
NEUMENT	28912	16665	663	65
RESPR4	25122	12089	295	27
MISCHRT	17983	12923	470	45
GIBLEED	18042	10423	319	62
SKNAUT	14795	11474	588	44
INFEC4	14253	6737	527	24
TRAUMA	14229	4594	120	17
ROAMI	6780	5889	160	6
RENAL3	7860	4686	178	25
HEART2	5661	6502	236	20
ODaBNCA	8092	4047	138	39
MISCL5	8180	3986	133	16

Table 2: Charlson Index vs PrimaryConditionGroup

Pearson's Chi-squared test
testdata: CharlsonIndex and PrimaryConditionGroup
X-squared = 110810, df = 132, p-value < 2.2e-16
The p-value is very less, hence we can reject null hypothesis and confirm that the Charlson Index and PrimaryConditionGroup are not independent of each other, hence related.

Primary group and Age at first claim

The United States — and the world — are aging. The number of Americans aged 65 and older is projected to double from 46 million to more than 98 million by 2060. It will be the first time in history that the number of older adults outnumbers children under age 5. In addition, older adults will live longer than ever before: One out of every four 65-year-olds today will live past age 90 (a report by American psychological association). We have made relation of primary group with age of the members, the reason for doing this is to take an overview of above mention report of APA. By the table we have chosen top 10 primary conditions and took percentage of illness each age group. We can see that till the age of 40 there are not much high rate of primary condition but after age of 40 there is increase in the rate of primary condition.

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
ARTHSPIN	0.72	3.4	2.82	6.76	12.74	14.8	17.73	26.79	14.24
INFEC4	15.23	9.16	6.54	9.5	11.98	10.57	10.43	16.22	10.37
METAB3	0.2	0.71	1.34	4.17	10.16	13.68	21.32	33.11	15.31
MISCHRT	0.25	0.25	0.64	2.75	7.31	11.38	19.08	34.92	23.42
MSC2a3	6.91	4.3	4.16	7.13	11.8	12.16	15.7	25.11	12.71
NEUMENT	4.5	3.96	2.83	5.5	9.09	9.79	14	29.1	21.23
RESPR4	13.45	10.06	5.42	8.71	13.27	10.77	11.55	17.24	9.54
SKNAUT	3.64	6.8	3.78	5.83	10.26	11.59	14.79	26.32	16.99
TRAUMA	7.53	15.01	6.8	9.74	13.69	10.6	10.28	15.18	11.17

Table 3: PrimaryConditionGroup and Age at first claim

Place of service and primary condition group

We have chosen relationship of primary condition and place of service, table below shows percentage of place of service with top 10 primary conditions. We can see that the majority of the patient visits are at doctor's office this might be because of the first-time consulting with their primary care physician or any specialist recommended by the patient's primary care physician. Second highest place of service is independent lab where different types of tests on patient are done, followed by outpatient hospital and so on.

	Ambulance	Home	Independent Lab	Inpatient Hospital	Office	Other	Outpatient Hospital	Urgent Care
ARTHSPIN	1.23	0.07	2.08	2.22	81.74	0.28	7.07	5.31
GIBLEED	5.41	0.06	8.69	6.22	54.12	0.35	8.06	17.09
INFEC4	0.29	0.15	14.22	2.33	65.36	0.32	1.54	15.79
METAB3	0.02	0.16	45.38	0.71	52.27	0.25	0.28	0.92
MISCHRT	0	0.24	13.81	1.63	80.51	0.51	0.89	2.41
MSC2a3	0.58	0.1	41.4	1.05	48.86	0.28	6.88	0.85
NEUMENT	5.52	0.25	2.68	0.74	81.2	0.53	2.81	6.27
RESPR4	0.28	0.03	4.07	1.52	76.85	0.21	2.96	14.08
SKNAUT	0.17	0.18	8.32	0.37	84.37	0.27	1.02	5.3
TRAUMA	0.46	0.07	0.95	1.62	53.81	0.86	3.91	38.3

Table 4: Place of service vs PrimaryConditionGroup

Data Preparation (New Attributes)

Claims per members:

We have created new attribute to find total claims filed by a member. Figure 8 depicts the number of claims on X-axis and the frequency on Y-axis. The reason of creating this attribute is to get idea about the claims filed by a member within given time duration. By the graph given below we can see that there are members who has claimed up to 37 times in given time and many members who just claimed once. There may be different claims for services at different places like ambulance, hospital, office etc. for the same kind of treatment at once. It can be inferred that the more claims a person has, the longer a patient is suffering or is under treatment for a disease which takes longer to be cured and has to be seen by a doctor regularly.

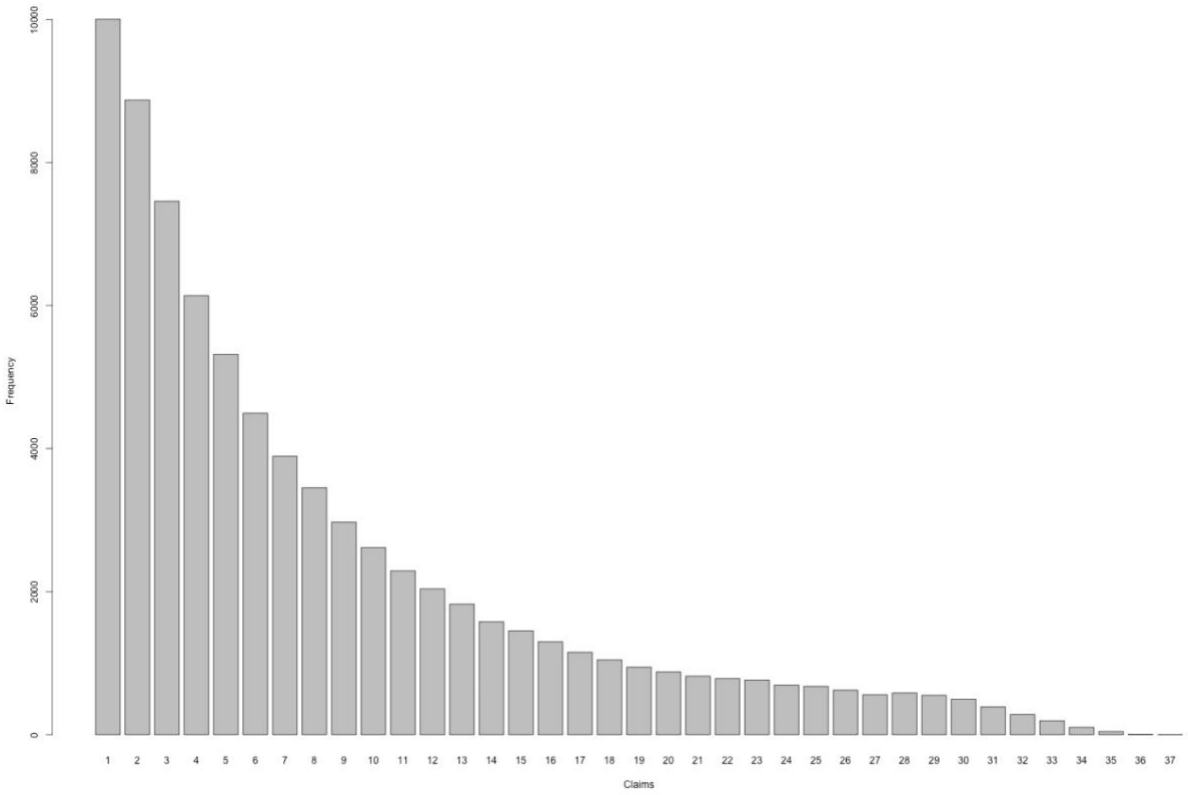


Figure 8: Number of claims vs Frequency of members

Average Pay delay

This is one of the very important relation we are thinking of where we calculate average of pay delay with members and we have plotted as shown below. Figure 9 shows a histogram with average pay delay plotted on X-axis and frequency of members on Y-axis. We have already discussed pay delay in earlier section. There are several reasons of pay delay such as non-payment by member, or incomplete documentation or service which is not covered under members insurance plan etc.

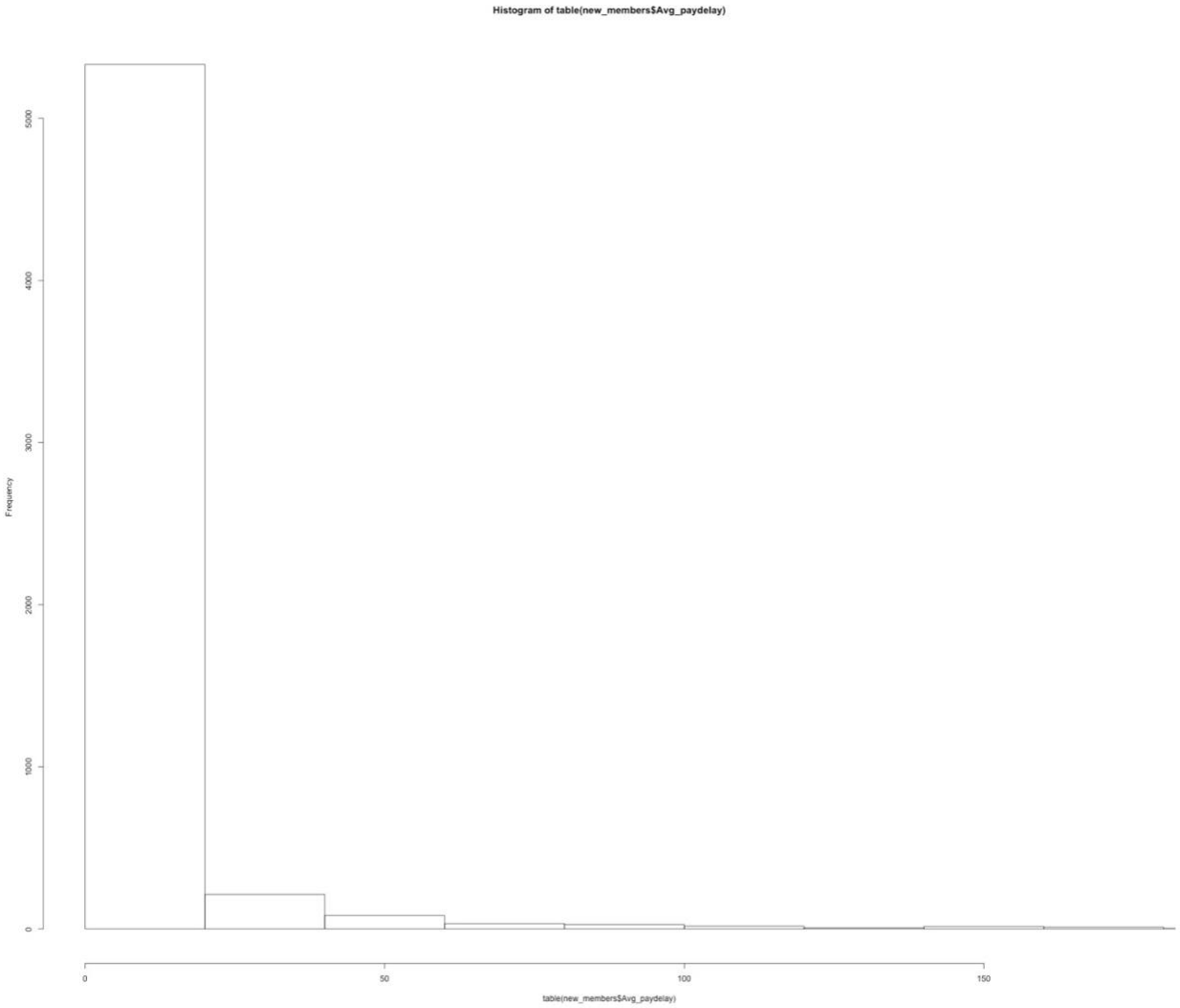


Figure 9: Average Paydelay vs Frequency of members

Difference in Charlson Index

The scale of Charlson Index attribute is converted into a numeric attribute and based on the claims of a member, the latest Charlson Index and the earliest Charlson Index are computed. Using this, the difference in Charlson Index is calculated which could give knowledge about the patient's health i.e., whether the patient's health is getting better or worse. Figure 9 depicts a barplot with difference in Charlson Index ranging from -5 to +5 and frequency of members on Y-axis. From the graph it is clear that the status of most of the patients is the same. This could be based on the fact that most of the patients had just one claim in the claims data. If the status of the health of the patients with just one claim is recorded at the end of the year, it could solve the mystery of the status of health.

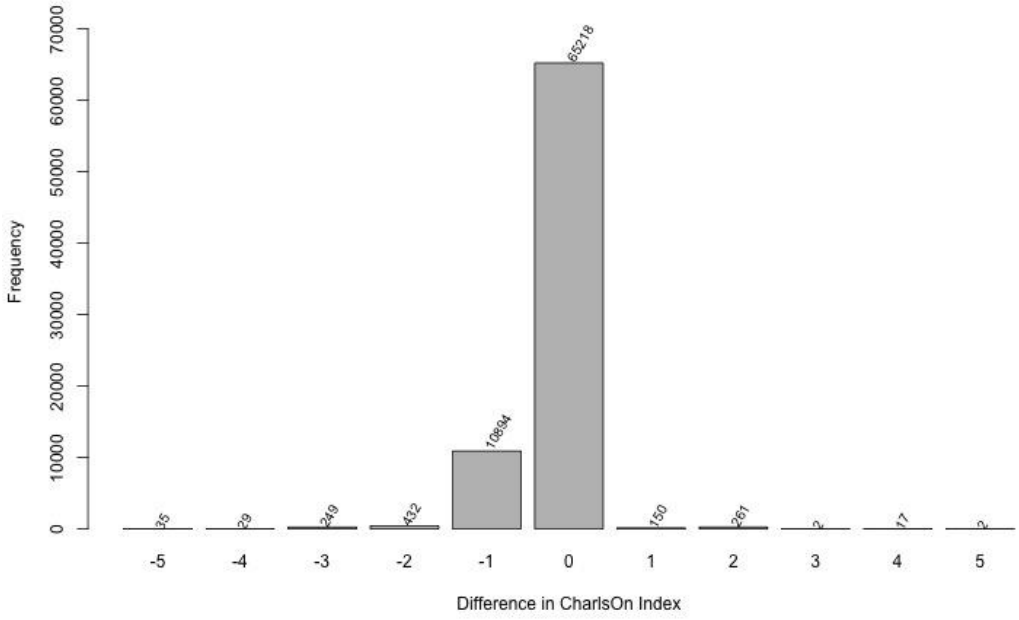


Figure 10: Difference in Charlson Index vs Frequency of members

Timeline of Charlson Index of two patients

Assuming that the claims have been recorded based on time, Figure 11 depicts a timeline of Charlson Index of two different patients (one in blue, other in black) who had highest number of claims in the data. For the sake of simplicity, let's say the blue line indicates the timeline of patient 1 and black line indicates the timeline of patient 2. The Charlson index of patient-2 keeps changing between 1 and 3 but during the last 6 claims, it has remained consistent at 1. Based on this consistency, we could say that his health is the same and didn't get worse.
Patient 1 has Charlson Index 1 until 21 claims, but his/her health conditions seems to have worsened and Charlson Index increased to 3 and after some fluctuation comes back to 3. The health condition of Patient 1 has not improved based on the timeline.

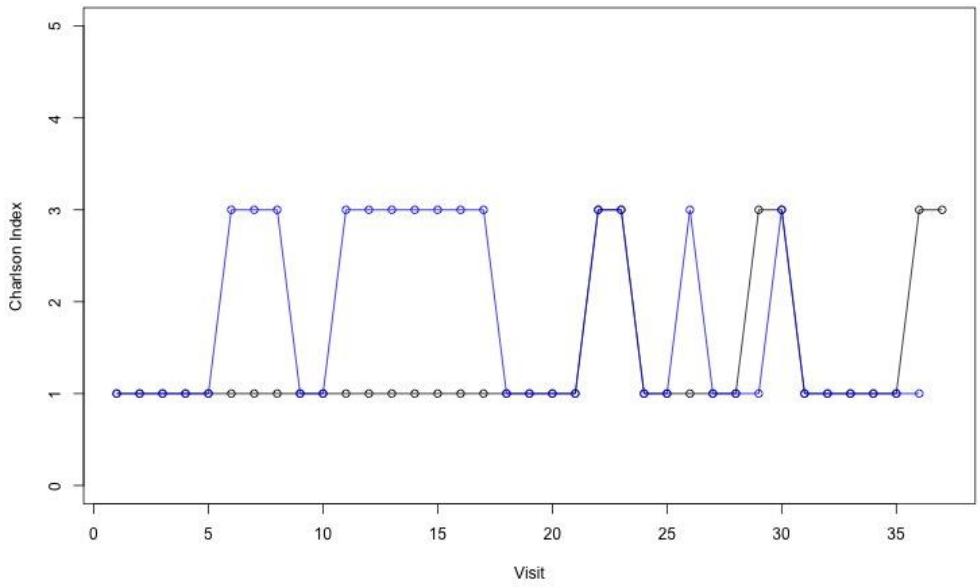


Figure 11: Timeline of Charlson index of two patients based on their claims

Conclusion:

During the course of our project we used important attributes like Primary Condition group, Sex, Charlson Index and others. With these attributes we will be able to predict the health condition of a patient accurately. The important attributes and relation between the attributes have been presented in a refined visualized format by using Bar Plots, Histograms and Cross Table. We created new attributes like number of claims files per person, average pay delay per person, their latest charlson index and the difference in their earliest and latest charlson index and creating appropriate visualizations for them to present in this report.

After plotting timeline of changing charlson index we can predict the future values of charlson index, this will be beneficial for the patient for their future treatment.

References

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3388783/
https://www.healthcare.gov/choose-a-plan/plan-types/
https://www.aha.org/system/files/2018-06/econ-contribution-2018_0.pdf
https://www.aha.org/system/files/2018-01/medicaremedicaidunderpmt%202017.pdf
http://www.apa.org/pi/aging/resources/guides/older.aspx
https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/2012AgeandGenderHighlights.pdf
https://www.racmonitor.com/wherever-you-go-there-you-are-the-importance-of-place-of-service