Exploratory Data Analysis of Heritage Health Claims Data

Rutvik Gavaskar

February 2020

Heritage Provider Network

I am pleased to present an exploratory data analysis report on Heritage Providers Network Claims Data. Drawing on the health care claims of 113000 Americans, one of the largest and most complete databases of its type, this report provides a one-of-a-kind view into health care use, for individuals over a span of 3 years.

The report showcases trends in gender, specialty, generalized place of service, length of stay, pay delay, Charlson index, drug count, lab count and primary condition group from the year Y1 to Y3.

The report relies on claims data from Heritage Provider Network. Note that because we rely on claims data, which is majorly categorical (non-numeric), variable frequencies are calculated to analyse and draw conclusions.

Some common statistical methods such as chi-square are used to test the relationship between 2 categorical variables. Entire analysis and plotting is done using Python and MS PowerBI.

Dataset:

The dataset provided by Heritage Provider Network (HPN) consists of 3 main tables:

Claims :

MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	LengthOfStay	DSFS	PrimaryCondition Group	CharIson Index	Procedure Group	SupLOS
96393713	7094351	347045	93075	Y3	Internal	Office	151	NaN	1- 2 months	METAB3	2-Jan	EM	0
57805129	3884005	523319	90756	Y1	Internal	Office	15	NaN	4- 5 months	SKNAUT	2-Jan	EM	0

Members :

MemberID	AgeAtFirstClaim	Sex
92806272	50-59	F
81827173	40-49	F

Days in Hospital:

MemberID	ClaimsTruncated	DaysInHospital
74032946	0	0
21964521	0	0

Dataset description:

For the purpose of data analysis, we will be only using the Claims and Members tables. The Days in Hospital table can be used to predict a patient's future days in hospital, which is not a part of this analysis.

The key dataset is the Claims Table that includes claims from 3 years of historical member membership, Y1, Y2, Y3.This dataset contains 2668990 records that are reports of patients over the three years.

The Members Table consists of Member information of their IDs, sex and their age range during the first claim. There are a total of 113000 unique records of members.

To further the simplify the analysis process, I have merged the Claims table with the Members Table based on MemberID field which is common in both tables. I will simply refer to it as Claims_Members table.

Data Understanding:

Data understanding is an important step in the analysis of data tasks, since it can help to get an idea of how to get started and which techniques and methods to use. In addition, it can also help from the beginning to make certain decisions about how to predict the results.

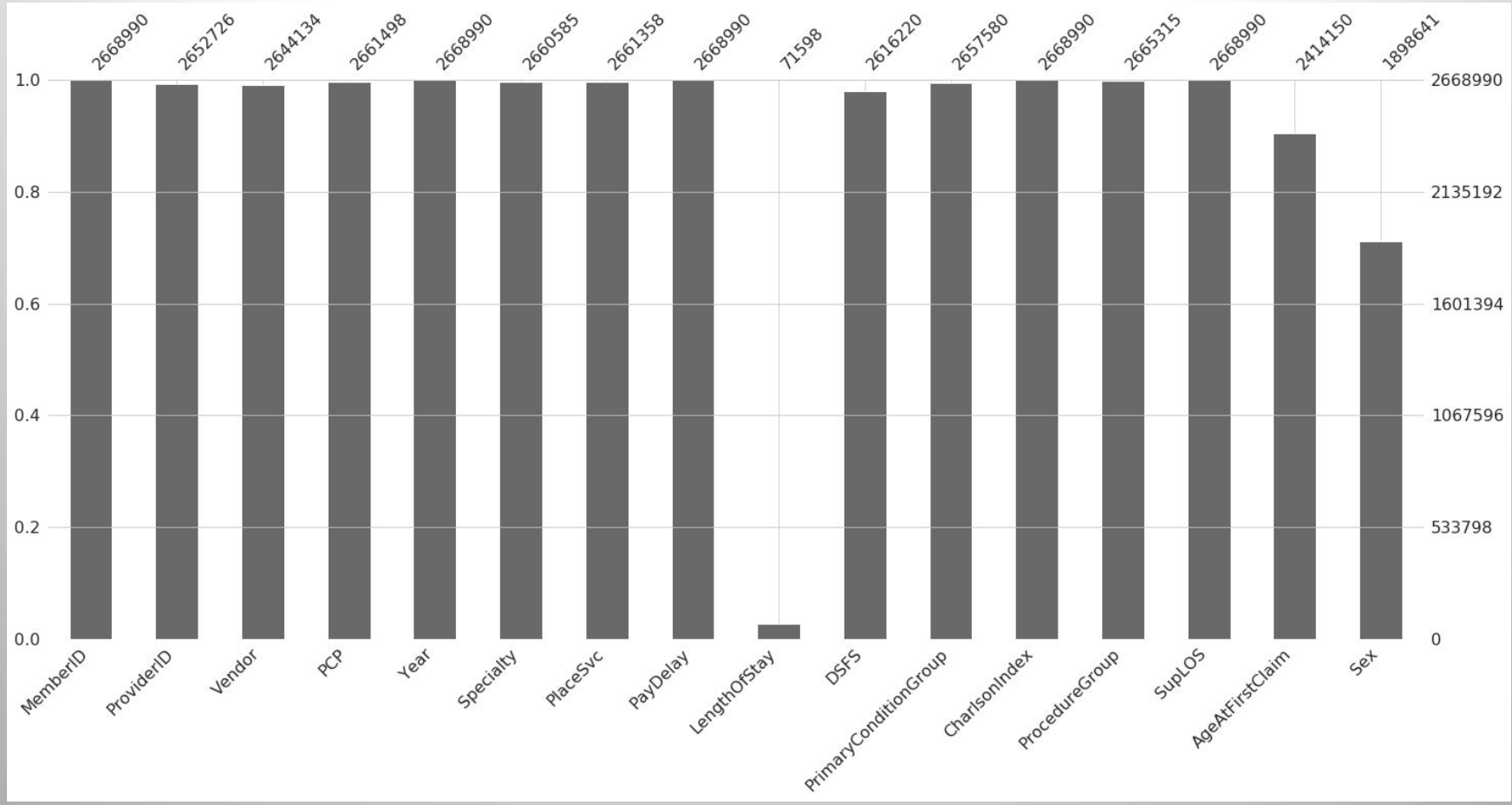
```
Int64Index: 2668990 entries, 0 to 2668989
Data columns (total 16 columns):
MemberID                int64    2668990
ProviderID              float64  2652726
Vendor                  float64  2644134
PCP                     float64  2661498
Year                    object   2668990
Specialty               object   2660585
PlaceSvc                object   2661358
PayDelay                int64    2668990
LengthOfStay            object     71598
DSFS                    object   2616220
PrimaryConditionGroup   object   2657580
CharlsonIndex           object   2668990
ProcedureGroup          object   2665315
SupLOS                  int64    2668990
AgeAtFirstClaim         object   2414150
Sex                     object   1898641
dtypes: float64(3), int64(3), object(10)
memory usage: 346.2+ MB
```

This table provides us with the type of data; numeric or non-numeric and total count of records in the Claims_Members Table, which includes the claims of historical participant from 3 year. As we can see there are 16 different fields and most of them are not numeric (nominal), will make it a bit difficult to plot.

It is also evident that some of these fields have a lot of missing entries.

Missing data:

Here is a chart that explains the severity of missing data in the dataset. This missing data is now treated to make this dataset more efficient.



Dealing with missing data:

ProviderID, Vendor and PCP :

Field	Description	Data	Distinct Values	Missing Percentage
ProviderID	IDs of doctor or specialist providing the service	[8013252., 9416979., 8511459., ..., 7676407., 6547259., 4875461.]	14699	0.60%
Vendor	company that issues the bill	[172193., 5166., 64764., ..., 246285., 898618., 663092.]	6387	0.89%
PCP	member's primary care physician	[37796., 5300., 91972., ..., 49539., 19128., 85137.]	1359	0.30%

These fields are IDs of individuals/companies and have an insignificant amount of records missing. These may or may not help in drawing any conclusions. For the completeness of data, we will keep these fields and replace the missing values with a new ‘unknown’ category for each of these fields.

Dealing with missing data:

Specialty, PlaceSvc, ProcedureGroup and PrimaryConditionGroup :

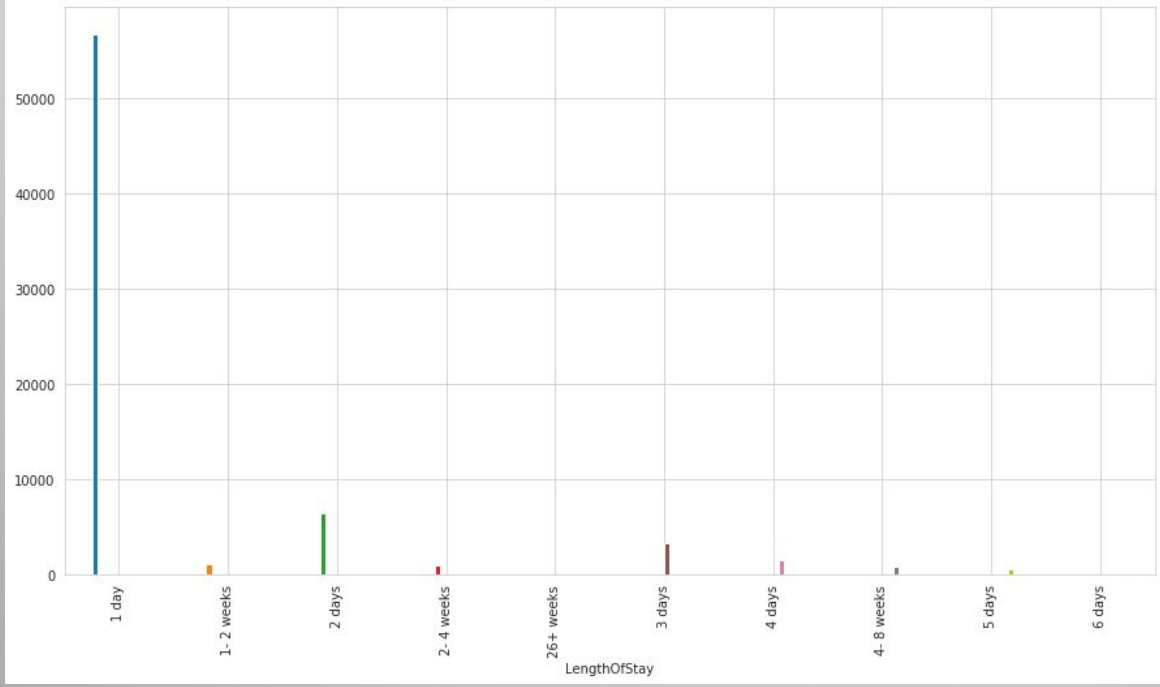
Field	Description	Data	Distinct Values	Missing Percentage
Specialty	Specialty of treatment	[Surgery, Internal, Other, nan, Laboratory, General Practice, Diagnostic Imaging, Pathology, Anesthesiology, Emergency, Obstetrics and Gynecology, Rehabilitation, Pediatrics]	12	0.30%
PlaceSvc	PlaceSvc Place where the member was treated	[Office, Outpatient Hospital, Independent Lab, Inpatient Hospital, Urgent Care, Other, nan, Ambulance, Home]	8	0.30%
PrimaryConditionGroup	A generalization of the primary diagnosis codes	[NEUMENT, MISCHRT, SKNAUT, GIBLEED, MSC2a3, ODaBNCA, METAB3, ARTHSPIN, HEMTOL, PNEUM, CANCRA, CATAST, RESPR4, GYNEC1, INFEC4, FXDISLC, COPD, UTI, TRAUMA, ROAMI, MISCL5, FLaELEC, SEIZURE, GYNECA, CHF, nan, NCRDZ, APPCHOL, AMI, HEART2, CANCRB, RENAL3, SEPSIS, GIOBSENT, HEART4, METAB1, PERVALV, RENAL2, HIPFX, STROKE, MISCL1, PRGNCY, LIVERDZ, CANCRM, RENAL1, PERINTL]	45	0.40%
ProcedureGroup	A generalization of the CPT code or treatment code	[MED, EM, RAD, SCS, PL, SIS, SDS, nan, ANES, SMS, SRS, SNS, SGS, SAS, SEOA, SUS, SO, SMCD]	17	0.10%

As shown in the table above these values are all polynomial and contain missing values. I have used the same method as for the previous one to deal with these missing values, which is treating the missing data as just a new category.

Dealing with missing data:

LengthOfStay, SupLOS:

Field	Description	Data	Distinct Values	Missing Percentage
LengthOfStay	Length of stay in hospital	[nan, '1 day', '3 days', '5 days', '2 days', '6 days', '1- 2 weeks', '2- 4 weeks', '4 days', '4- 8 weeks', '26+ weeks']	10	97.30%
SupLOS	A flag that indicates if LengthOfStay is null because it has been suppressed	[0, 1]	2	0

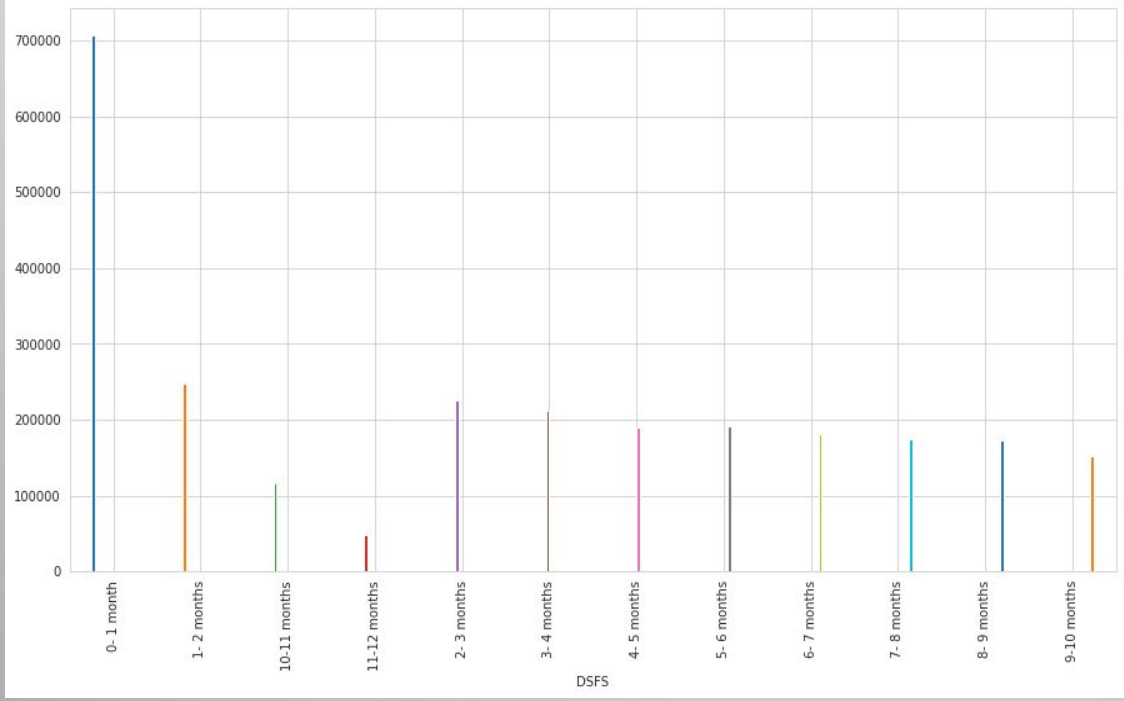


The field LengthOfStay has over 97% missing values. The distribution of its values can be seen in the graph. If there is a blank LengthOfStay and SupLOS is 0, then that is how it was when it came out of the HPN dataset, according to the data provider. When LengthOfStay is null and SupLOS is 1 then LengthOfStay has been suppressed.
As there are over 97% missing values, I have decided to delete this field from the dataset along with SupLOS field.

Dealing with missing data:

DSFS:

Field	Description	Data	Distinct Values	Missing Percentage
DSFS	Days since first service that year	['8- 9 months', '9-10 months', '0- 1 month', '6- 7 months', '1- 2 months', '7- 8 months', '10-11 months', '3- 4 months', '5- 6 months', '2- 3 months', '11-12 months', '4- 5 months', nan]	12	2%

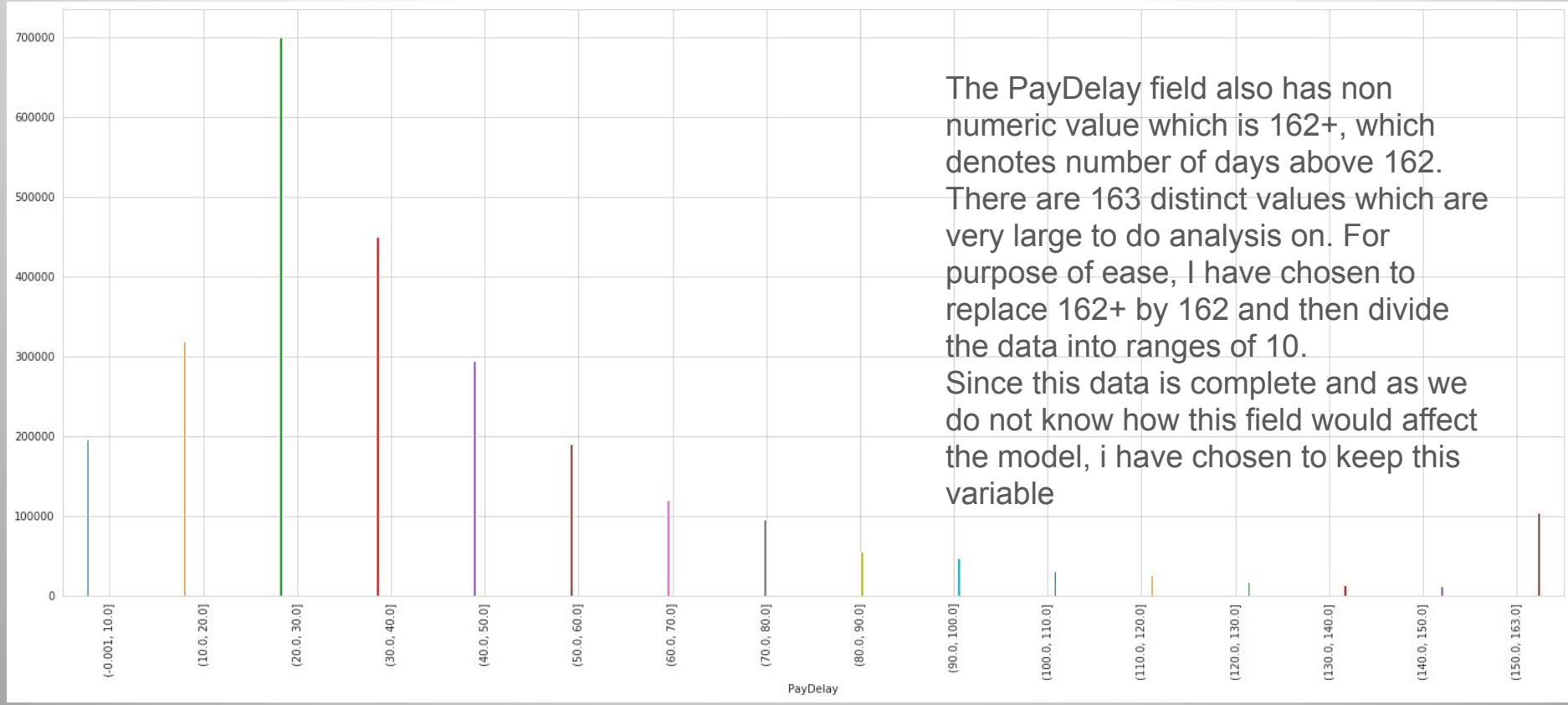


We can see from this histogram that the majority of DSFS are with value: "0-1 month" but some are different. One idea is to replace the missing DSFS attribute values with the most common value for this attribute that is "0-1 month." But as this can cause knowledge loss that can increase the number of days spent in hospital.
I will be using the same method as before to replace missing values with a new category called 'no-month'.

Data processing:

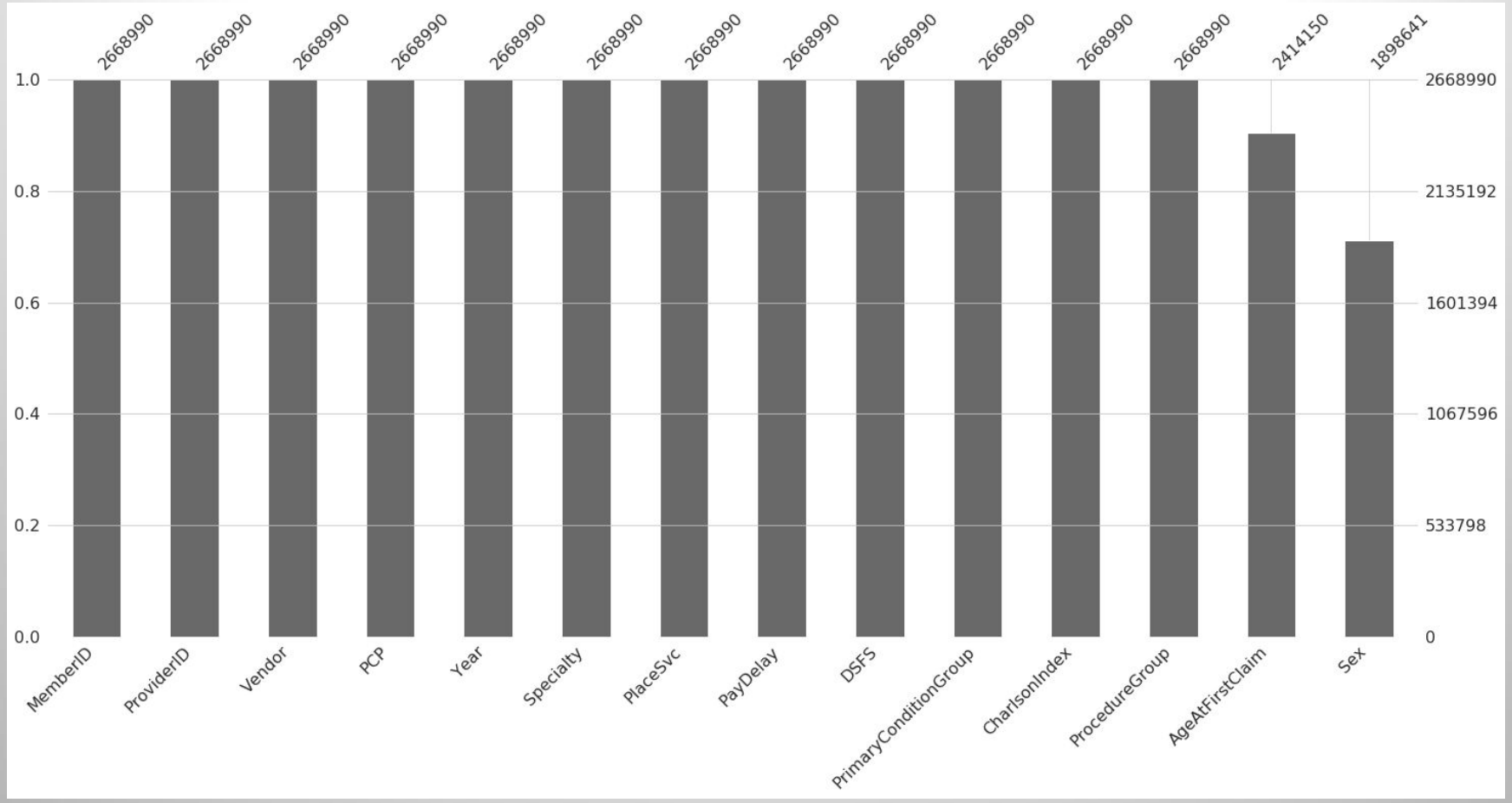
PayDelay:

Field	Description	Data	Distinct Values	Missing Percentage
PayDelay	The delay between the claim and the day the claim was paid for	[1, 2, 3, 4,....., 160, 161, 162+]	163	0%



The PayDelay field also has non numeric value which is 162+, which denotes number of days above 162. There are 163 distinct values which are very large to do analysis on. For purpose of ease, I have chosen to replace 162+ by 162 and then divide the data into ranges of 10.
Since this data is complete and as we do not know how this field would affect the model, i have chosen to keep this variable

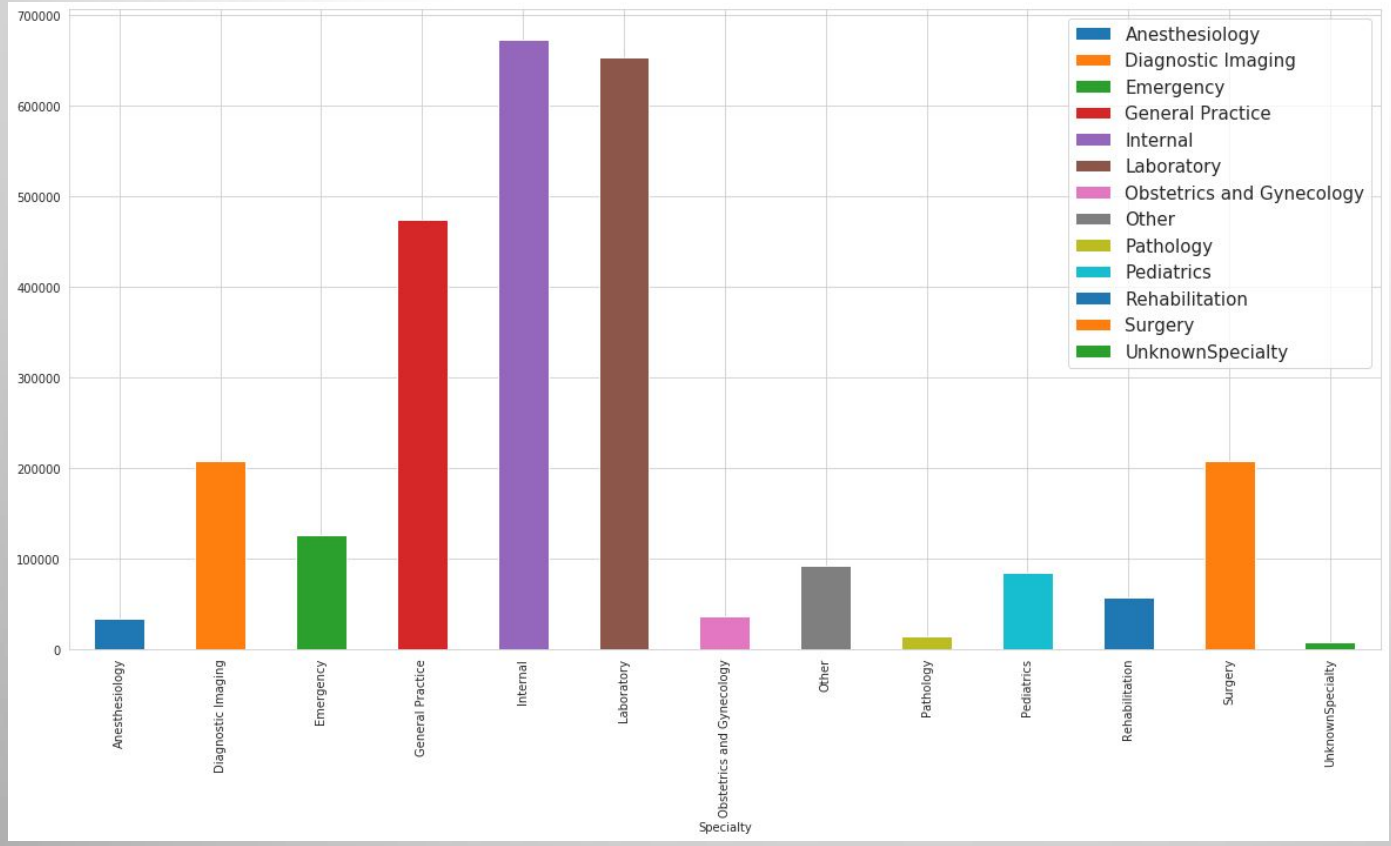
Clean data:



I have chosen to keep AgeAtFirstClaim and Sex data as is as they are important contributors in the analysis. The data now looks complete and ready for analysis.

Some Important Variables:

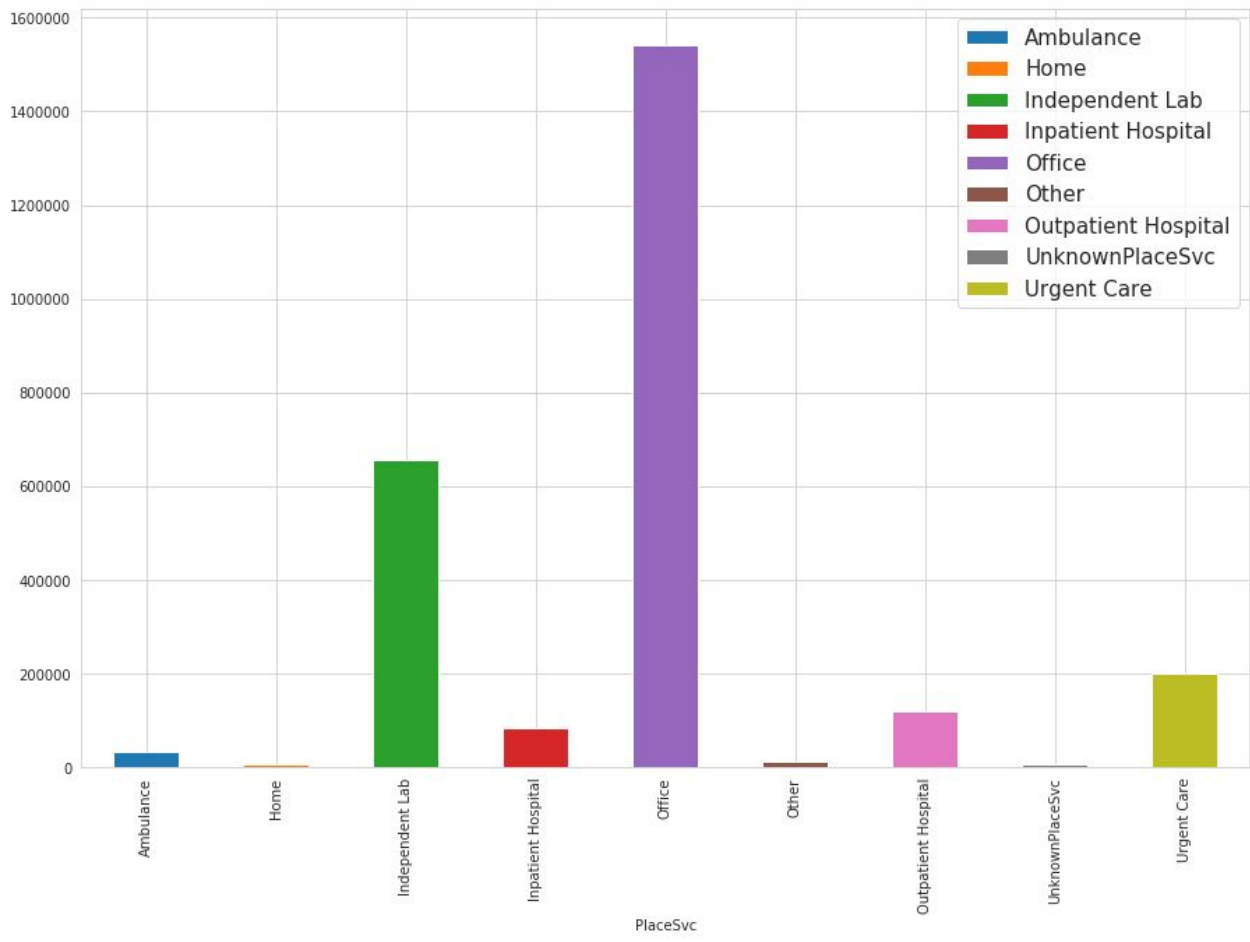
Specialty:



From the Frequency graph, it can be seen that most claims have been made against Internal (~670000), Laboratory (~650000) and General Practice (~480000) specialty groups over a period of 3 years from Y1 to Y3.

Some Important Variables:

PlaceSvc: Place of Service is where the member was treated.

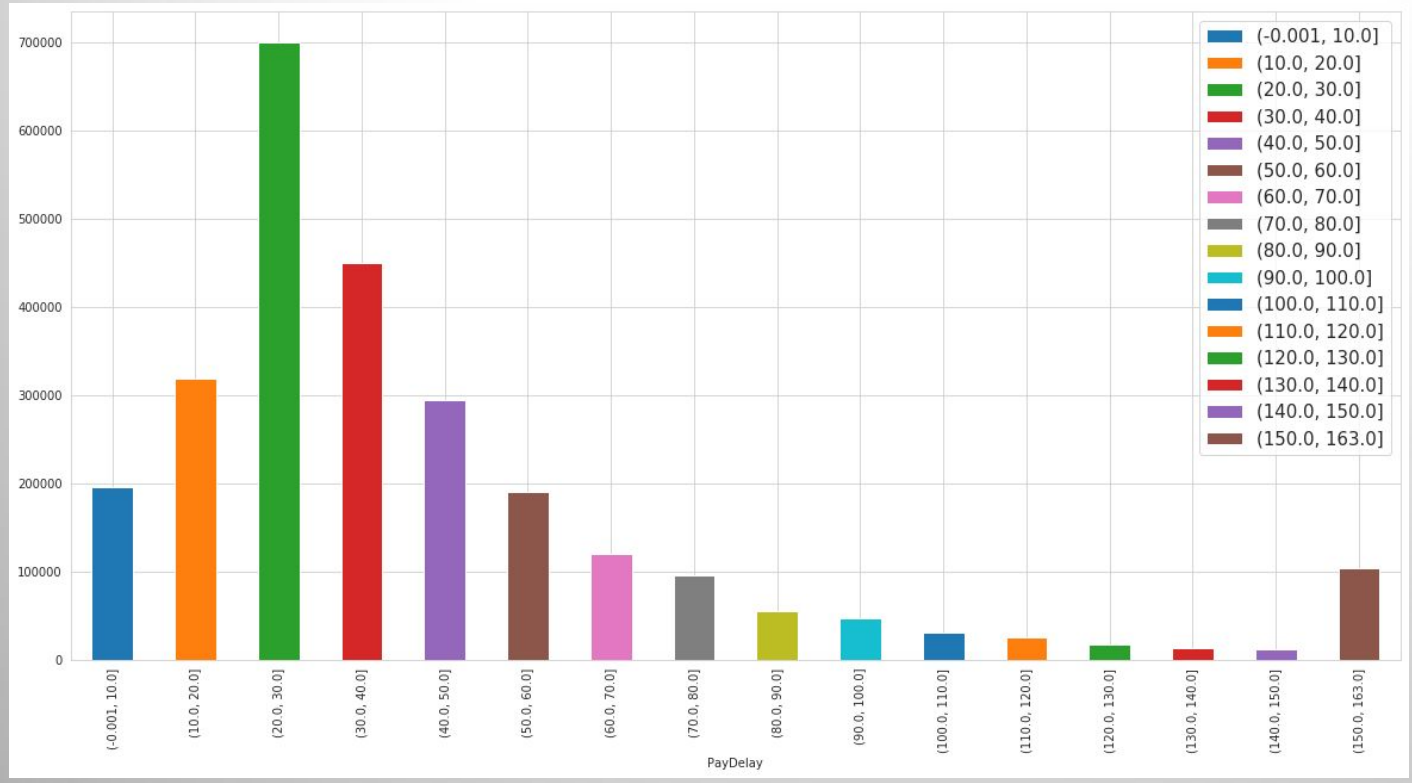


It is quite evident that most of the members (~1550000) visited a healthcare provider's office for treatment or consultation than any other place of service followed by independent lab (~650000), over a period of 3 years.

We can say that Office and Independent Labs are the places more likely visited by patients for treatment.

Some Important Variables:

PayDelay:



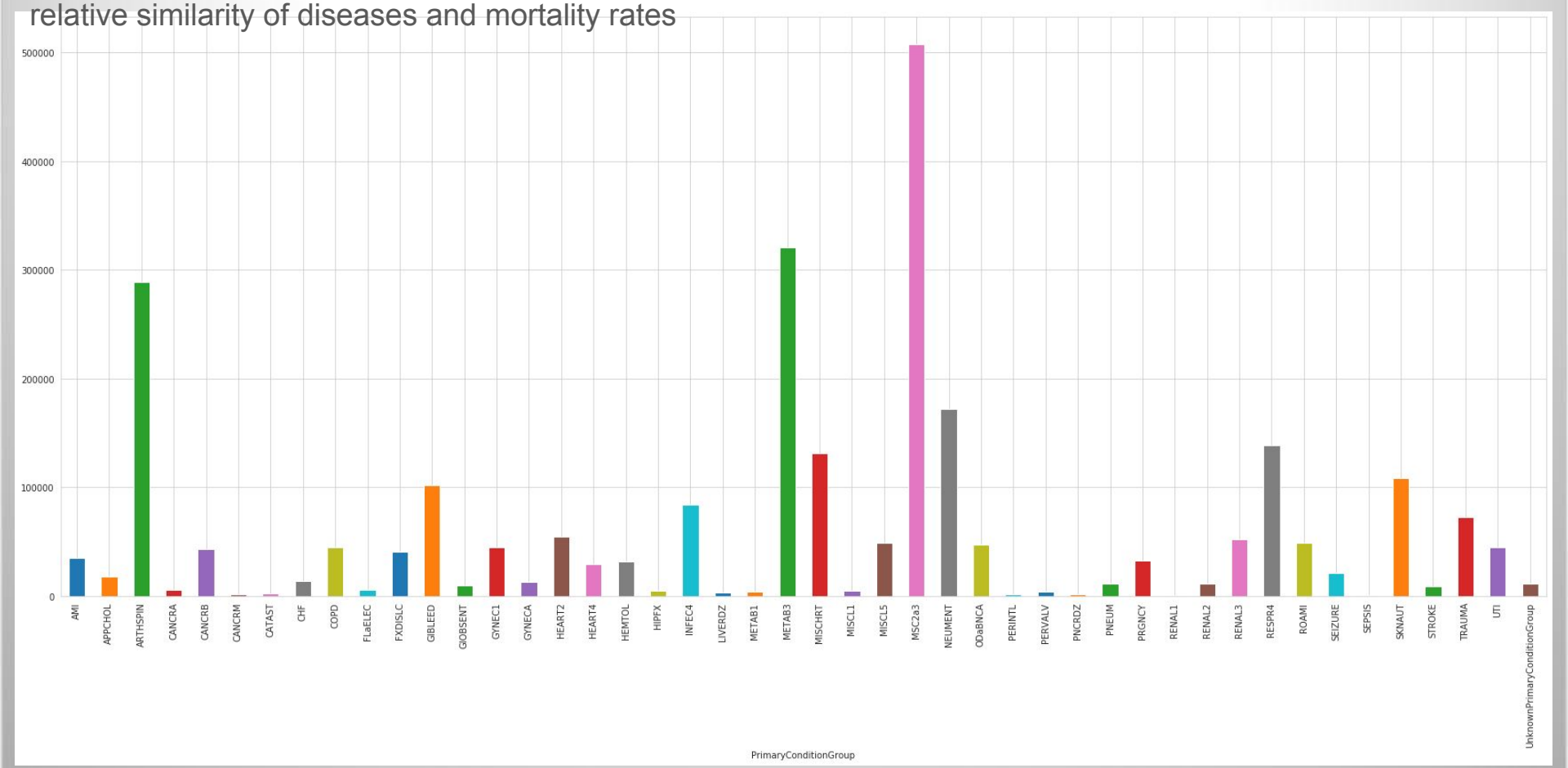
Most of the payments were made between 20 and 30 days.

There is a sudden increase in number of payments made after 150 days, which are well over 100000.

Note that these insights are based on data of all 3 years together.

Some Important Variables:

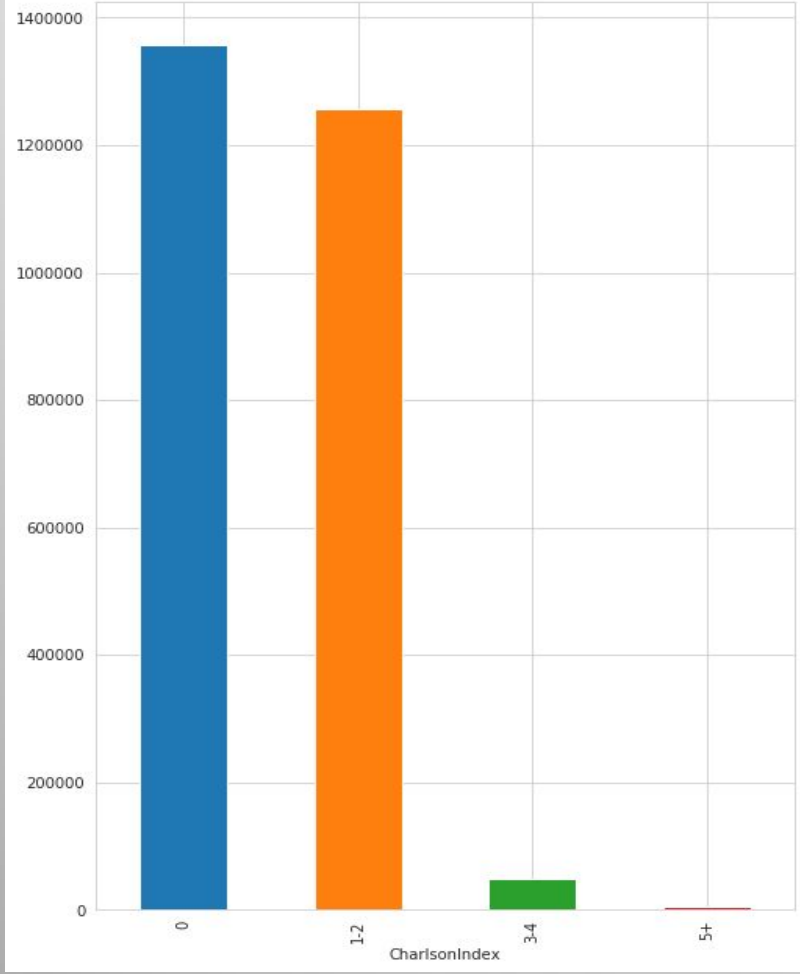
PrimaryConditionGroup: Primary conditions refer to broad diagnostic categories, which are based on the relative similarity of diseases and mortality rates



Most of the members obtained treatment for MSC2a3 (~500000) which are external causes of injury, followed by METAB3 (~320000) which are endocrine, metabolic or immune disorders. Approximately, 280000 claims were made for ARTHSPIN which are arthropathies and spine disorders.

Some Important Variables:

Charlson Index:

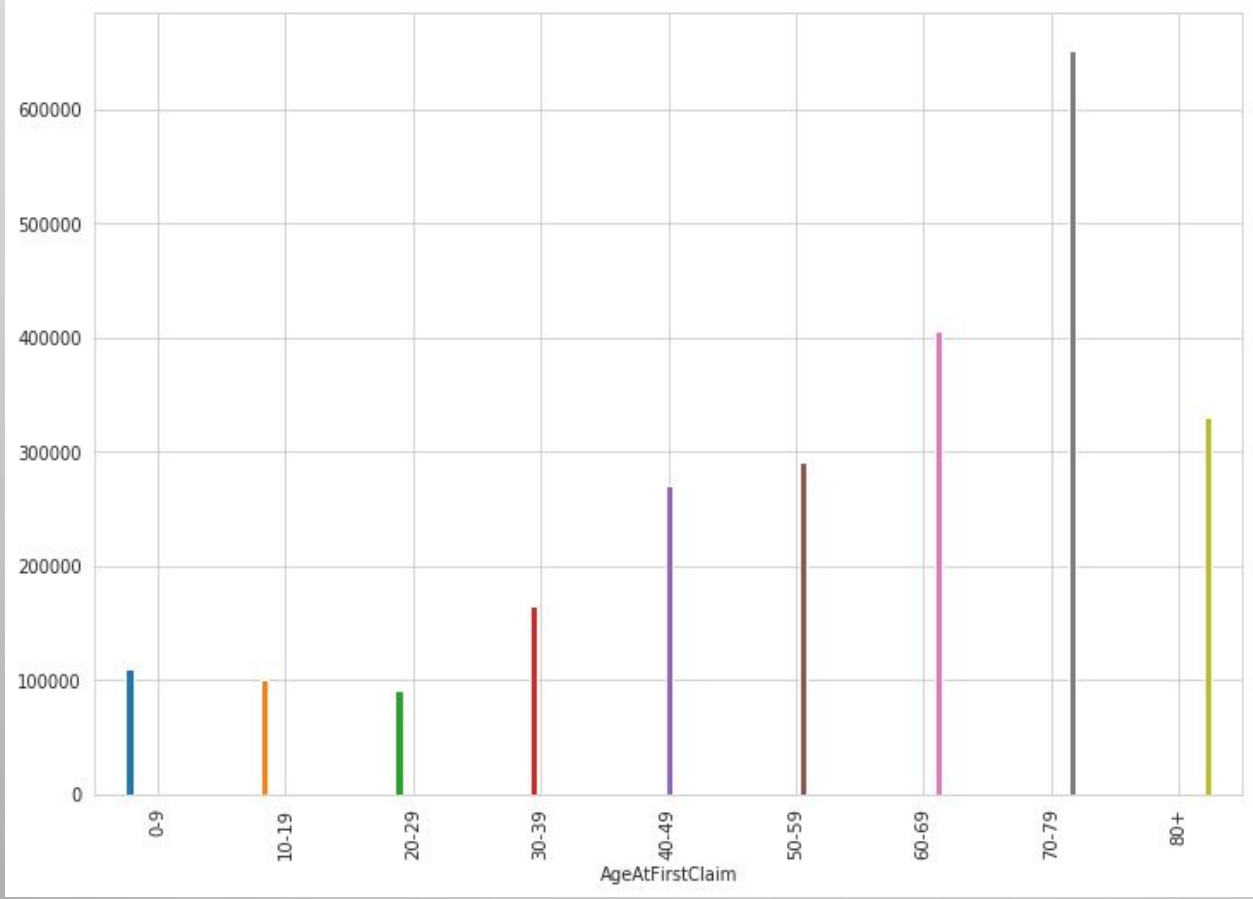


Most of the the patients have comorbidity score of less than 3.

Score	Condition
1	Coronary artery disease
	Congestive heart failure
	Chronic pulmonary disease
	Peptic ulcer disease
	Peripheral vascular disease
	Mild liver disease
	Cerebrovascular disease
	Connective tissues disease
	Diabetes
	Dementia
2	Hemiplegia
	Moderate-to-severe renal disease
	Diabetes with end-organ damage
	Any prior tumor (within 5 y of diagnosis)
	Leukemia
	Lymphoma
3	Moderate-to-severe liver disease
6	Metastatic solid tumor
	AIDS (not only HIV positive)

Some Important Variables:

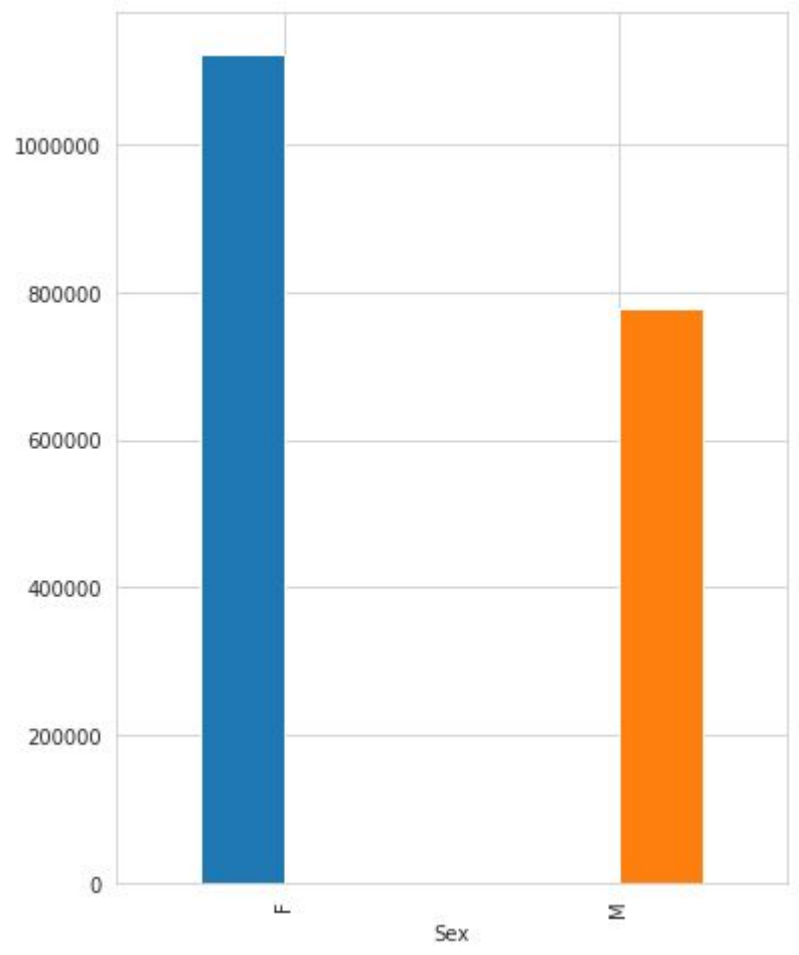
AgeAtFirstClaim:



Highest number of claims were made by the 70-79 age group, followed by 60-69 age group.

Some Important Variables:

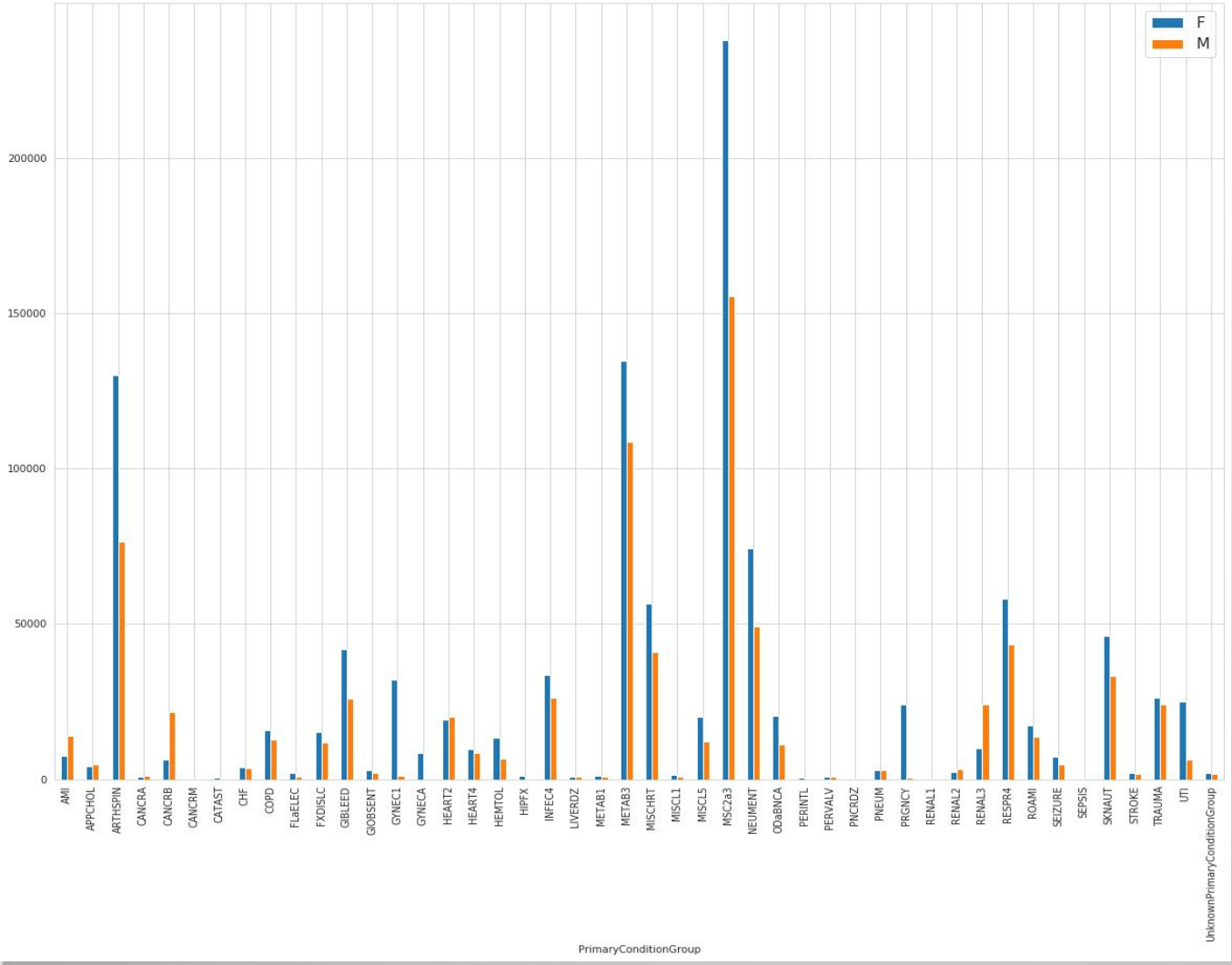
Sex:



1 million more claims have been made by females than males each year, despite removing the pregnancy factor from the Primary condition group during calculation. Nothing can be said about this since there is a significant amount of missing data in the ‘Sex’ column, but the same trend is seen over 3 years separately.

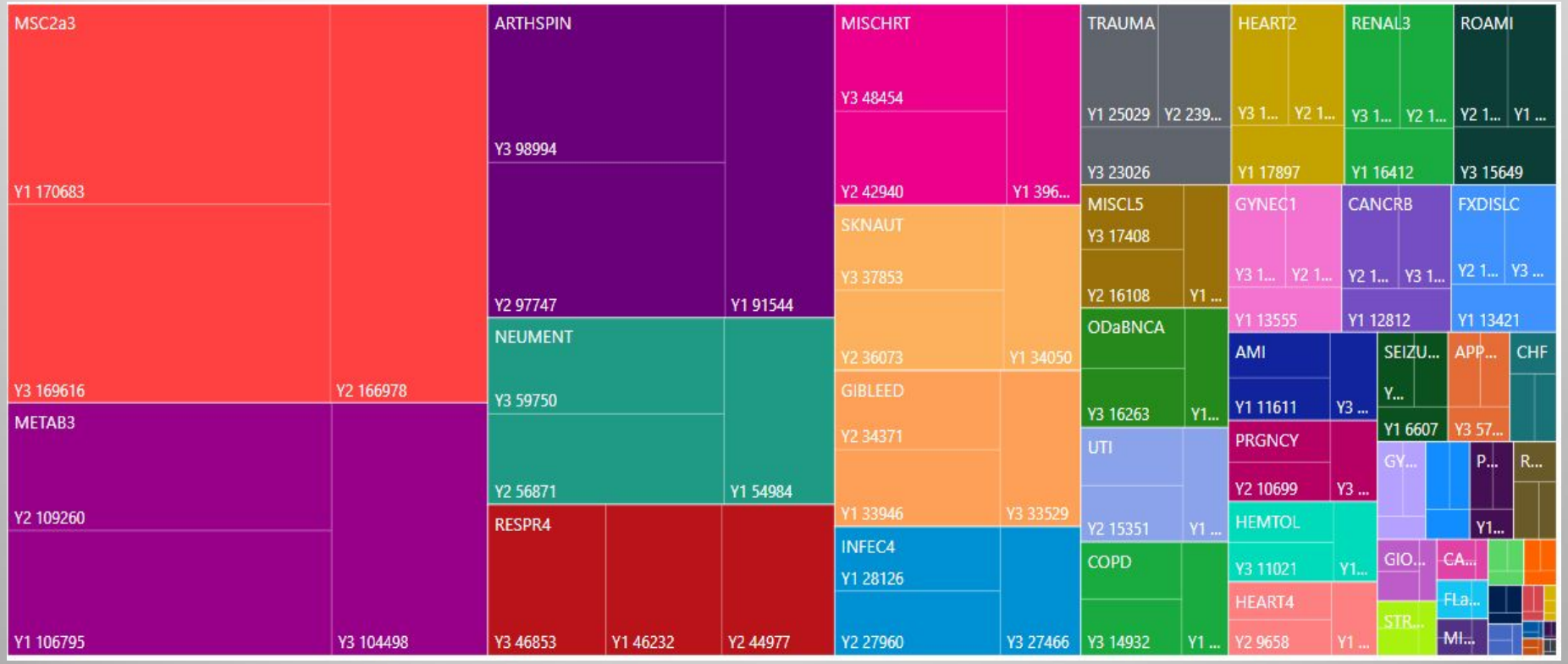
Key Variable Pairs:

Primary Condition Group and Sex (all 3 years):



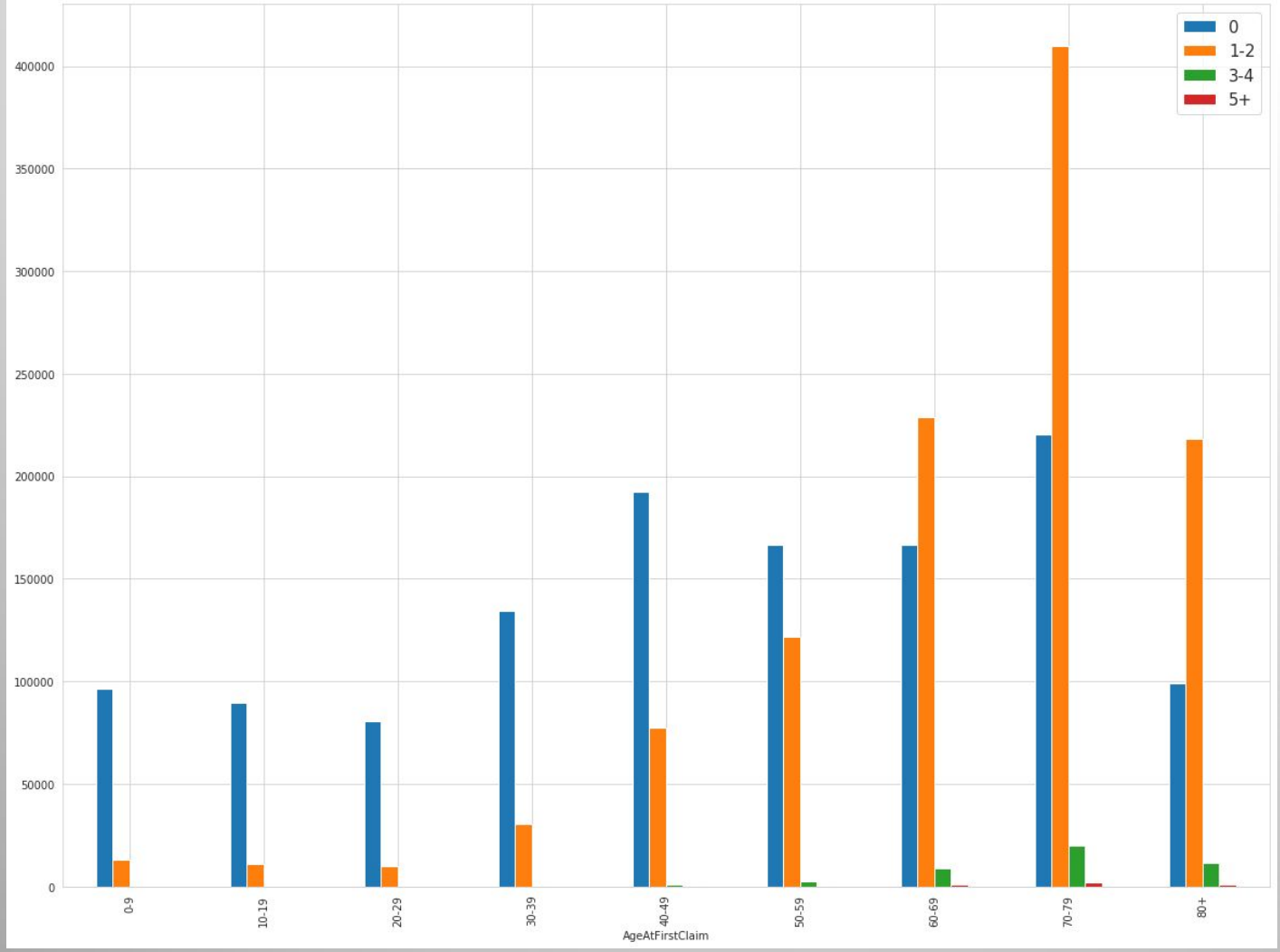
From this pair frequency graph it is evident that significantly more females have obtained treatment than males in all diagnostic categories except AMI - Myocardial infarction, APPCHOL - Appendicitis, hernias, cholecystitis, and cholangitis, Cancer, Diseases of pulmonary circulation, and, cardiac dysrhythmias,and Chronic renal failure.

Tree Map displaying year wise segregation of services obtained from a PrimaryConditionGroup



Key Variable Pairs:

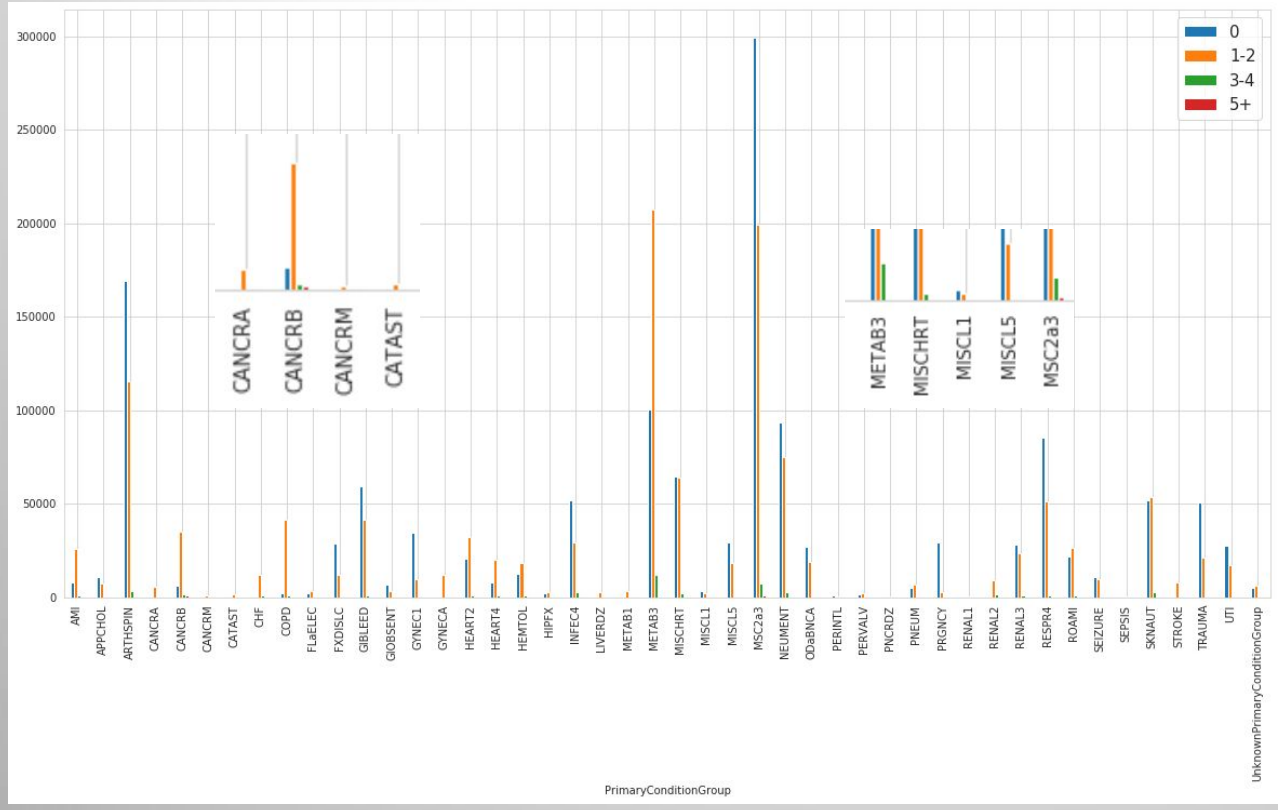
AgeAtFirstClaim and Charlson Index (all 3 years):



Here it can be seen that higher the age, higher is the Charlson Index score. It can be said that most of the claimants below the age of 40 obtain treatments for external injuries or non-severe diseases, while the members above the age of 40 are likely to be obtaining treatments for severe diseases.

Key Variable Pairs:

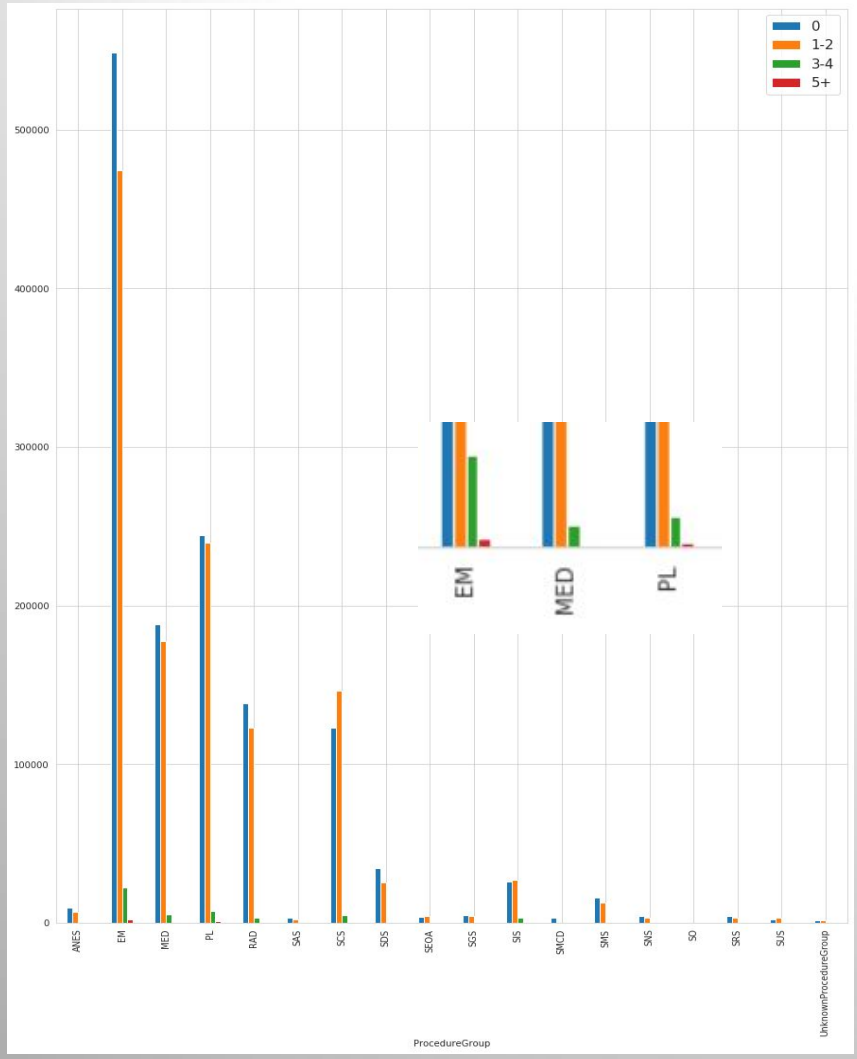
PrimaryConditionGroup and Charlson Index:



From this plot it can be clearly seen that higher the severity of disease, higher is the Charlson Index. For example Cancer Benign patients also have a score of 5+.

Key Variable Pairs:

ProcedureGroup and Charlson Index:



Most Patients with high charlson index visit EM or PL i.e. Emergency services or Pathology and Laboratory. Hence higher the charlson index , more likely the patient obtains service from a critical group.

Executive Summary

- Some interesting facts such as the top Providers, Vendors and Primary Care Physicians can be tracked.

Provider	Cases		Vendor	Bill Count		PCP	Cases
7053364	293866		240043	293868		91972	73772
1076052	143520		140343	193442		32724	46724
4107701	107100		251809	143520		20893	40845

- Despite removing the Pregnancy factor, approximately 1 million more females obtain medical services each year, than males.
- The age group of 70-79 are the highest seekers of medical services followed by 60-69 and 80+ each year.
- Top causes of obtaining medical services continue to be :
  - External causes of injury
  - Endocrine, metabolic and miscellaneous immune disorders
  - Arthropathies and spine disorders
- Currently, there are 943 members with Charlson Index over 5.

Recommendations:

- The cause behind females obtaining more care services than males, is still unknown, more research should be done here to find the cause.
- Charlson Index is a factor to determine a member’s future needs.
- Members in the age group of 70-79 should be given more attention as they are highly likely to obtain care services each year.

References:

- Exploratory data analysis in Python.
- Predicting hospitalization for patients
- Python Code colab link
- 1.1 Description of Data Fields