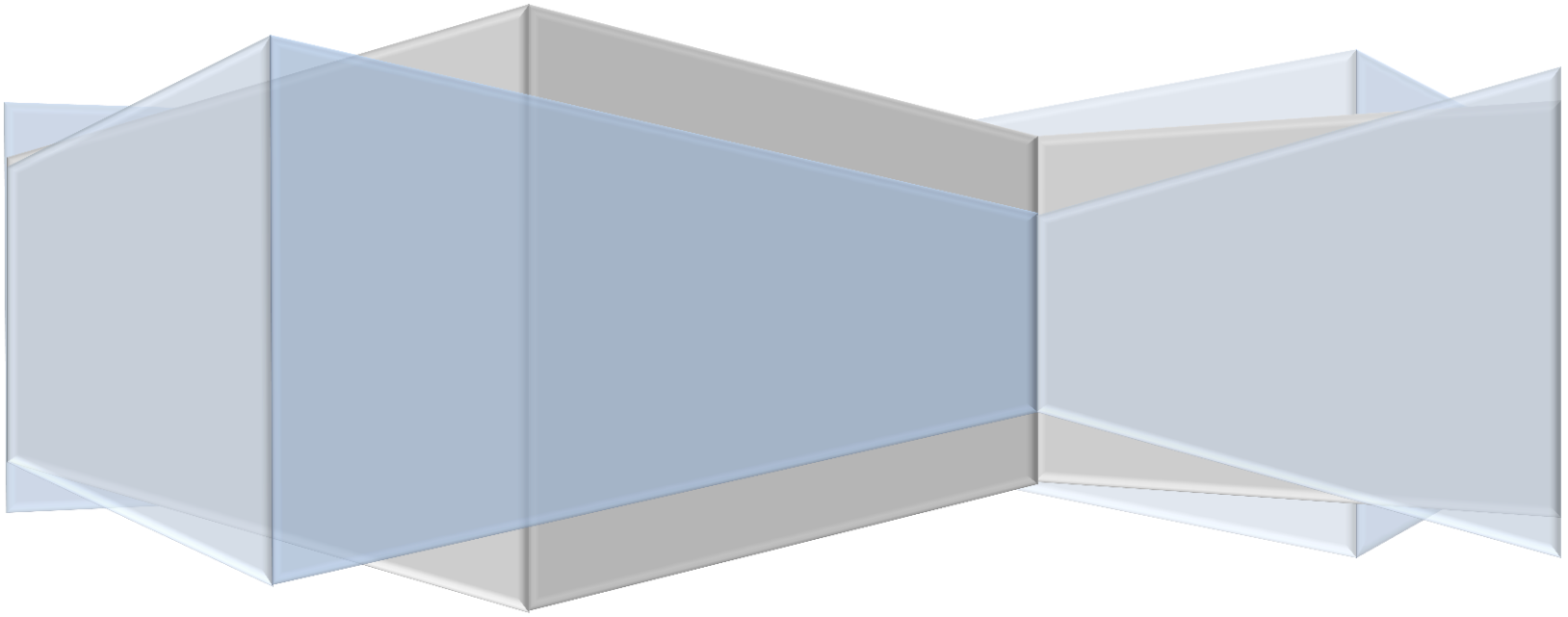
Heritage Health Provider Network:

Analyzing Time Series Data on Health Claims to Predict Days in Hospital

By: Tazein Fatma and Nathalie Pham | May 8, 2018



Github Repository

Please use the linked GitHub repository for accessing *most* of the data and all of the SPSS Modeler streams, R files, and Python notebooks used in the subsequent analysis for this project. The only data files that were not uploaded to the GitHub repository were the output files from the SPSS Modeler stream called *Stream1.str* which was used to join the relational tables. The data files were too large to upload to GitHub.

https://github.com/phamn3/heritage-health

Introduction

Hospitalization is touted as the largest component of health expenditure, and places a significant amount of financial strain on the U.S. healthcare system. The average cost on the hospital side per patient, depending on the state, can range anywhere from $1,000 to more than $3,000 for a typical hospital stay. The cost to patients ranges from $100 to $500 per day for an observation stay [1]. In 2016, the U.S. spent 17.8% of its GDP on healthcare, compared to the ten top high-income countries (Canada, Germany, Australia, U.K. Japan, Sweden, France, etc.) which spent only around 11.5% [2]. The rate of growth of healthcare spending far outstrips the rate of GDP growth in the U.S., and the country still performs lower on several population health outcomes when compared to other high-income countries [3]. The reality of it is that the United States has an unsustainable healthcare system.

Drivers of high and rising costs in healthcare include fluctuating prices of labor, goods, pharmaceuticals (drugs and medication), medical technology (medical equipment, knowledge, and devices), increasing severity of chronic diseases, administrative costs, and lack of management [3]. In particular, fragmented and uncoordinated data silos within hospital IT structures create inefficiencies within the system, and reduces timely, quality care to patients.

Due to the significant amount of pressure that hospitals are receiving, medical providers are keen on being able to accurately predict patients at a high risk of repeat hospital admissions. By employing predictive models, hospitals can work towards trying to understand what underlying factors in the care they provide help to reduce hospitalization rates across conditions, and can work towards improving the quality of care they provide. The idea is to identify these high-risk patients early to provide them with more personalized outpatient care, thereby helping to reduce unnecessary hospitalization admissions and medical spending overall. The healthcare industry contains a rich source of information on patients (demographics, medications, procedure summaries, lab tests, prescriptions, etc.) that can be de-identified and used along with machine learning algorithms to produce actionable insights.

The Challenge

The Heritage Health Prize competition was hosted by Kaggle in which the goal was to predict the number of days a patient would spend in the hospital in Year 4 given three years of historical claims data. De-identified data was provided by the Heritage Provider Network. In addition to claims data, demographic, drug count, and laboratory count data were provided. At the time, a $3 million dollar reward was offered to the team with the best predictive model.

The scope of the original challenge fit the intrinsic motivations of the group to utilize health data for this project. Medical health data is notoriously messy and heterogeneous in

nature, and we wanted to see how we would fair in the end to end business value development process.

Problem Definition

Due to the fact that the Kaggle competition ended in 2012, the number of days in hospital in Year 4 was not released making it no longer possible to calculate an accuracy score for predictions. This project instead explored the use of Year 1-2 claims data to predict the days in hospital of Year 3.

The business value proposition was to try and see if predictors of days in hospital could be discerned from the data. Doing so would help to increase efficacy of population health management practices and reduce hospitalization costs.

Data Sources:

Although Kaggle has officially taken down the datasets used for the Heritage Health Prize competition from their website, an archived repository was found with all of the necessary components from the contest [4]. The archive contains contest data, rules, and supplemental information that was made available to contestants at the time of the competition.

Contest data was provided to contestants in three staggered releases throughout the duration. Only the data files contained in Release 3 were used for this project [4]. The Release 3 zip file contained separate relational tables for historical claims data in Years 1-3, members data (age and sex), drug count data, lab count data, and outcome data for Years 2-3 (days in hospital). A detailed description of data fields was found in the data dictionary [5] and was used to understand the data.

Data Processing

The first goal of pre-processing was to join all relevant tables into one main table and aggregate the data in such a way that each row corresponded to a patient rather than a claim. The claims table contained multiple instances of MemberID, which indicated that a particular patient revisited a hospital on more than one occasion (i.e. had multiple claims). The unit of analysis was changed to per patient to match the unit of analysis of the outcome variable (days in hospital per patient). Below outlines the processing steps taken on the data.

SPSS Modeler was used to join all tables:

1. Claims table contained historical claims data for Years 1, 2 & 3. This table was joined with Members data on the MemberID key to bring in demographic data such as Age and Sex information for each claim record.
2. DrugCount and LabCount data was also joined with the Claims table on MemberID and DSFS (Days Since First Claim).
3. The merged data file up to this point was filtered based on Claim Year to create two files. One file contained the joined Claims data for Y1 and the other contained joined Claims data for Y2.
4. DaysInHospital Year 2 and DaysInHospital Year 3 tables contain information about “ClaimsTruncated” and “DaysInHospital”. These tables were joined with the Claims tables. In particular, DaysInHospital Year 2 was joined with Claims data Y1 to form the

train dataset while DaysInHospital Year 3 was joined with Claims data Y2 to form the test dataset.

5. The final tables contained Claim level information based on MemberID. The Train data dimensions were 865,689 rows by 20 columns while the Test data dimensions were 898,872 rows by 20 columns.

R Studio: Data Cleaning

The same data cleaning steps were done on both the Train (DIH_Y1_Target) and Test (DIH_Y2_Target) datasets. R files for cleaning each respective file were named Agg_Y1.R and Agg_Y2.R respectively.

1. DSFS variable (days since first claim) was updated from “0-1 month”, “1-2 months”, and so on to integer values of “1”, “2” and so on.
2. Claim Year column was dropped since each files only contains claims data for a single year.
3. PayDelay variable was categorical and had values of “162+”; they were updated to an integer value of “162”.
4. LengthOfStay variable was updated to indicate the number of integer days. “1 day”, “2 days” and so on were changed to integer values of “1”, “2”, up to “6” days. Values of “1-2 weeks”, “2-4 weeks”, and so on were changed to the median value of “11”, “21”, etc. days. The highest category “26+ weeks” was updated to “182” days.
5. DrugCount had “7+” values which were changed to “7” integers. Missing values were changed to “0”, assuming there were no drugs prescribed. Overall the variable was changed to integer.
6. LabCount had “10+” values which were changed to integer “10” values. Missing values were changed to “0”, again assuming the patient did not have any laboratory tests conducted. Overall the variable was changed to integer.
7. AgeAtFirstClaim had values grouped in age groups of “0-9”, “10-19”, “20-29” and so on up to “80+”. Each age group was assigned its median value of “5”, “15”, “25” and so on with “80+” capped at “80”. Also this variable had missing values and they were imputed with value of “99”.
8. For all other categorical variables, missing values were replaced by “Missing_CategoryName” value (Sex, ProviderID, Vendor, PCP, Specialty, PlaceSvc, PCGroup, ProcedureGroup). Missing values were imputed in this way so that a dummy variables would be created correctly in the Python code.

At this point in the data processing, exploratory data analysis was conducted in R Studio. The EDA insights are highlighted in the next section titled “Exploratory Data Analysis”. Continuing the explanation of the process methodology, the data cleaning on the Train and Test datasets from R were outputted into csv files (Dummy1.csv and Dummy2.csv) and brought into Python to conduct aggregation (using the notebooks DIH_Y1_dummy_dtypeFix.ipynb and DIH_Y2_dummy_dtypeFix.ipynb).

Python in Jupyter Notebook: Dummy Variables and Aggregation using Pivot Tables

The next task was to aggregate the data by the MemberID field to reduce the unit of analysis from claims to patient level. Dummy variables were created for all categorical variables. Data was grouped by MemberID for aggregation and pivoted against the rest of the variables. All categorical variables were summed in the pivot except for the following fields:

- Count of MemberID to get the number of claims per patient
- Mean of AgeAtFirstClaim and DSFS
- Distinct count of ProviderID, Venor, and PCP

Resulting dataset for Train and Test shows aggregated data from Y1 and Y2 for each patient (one row per patient). The output datafiles were out_Agg_dummy_Y1.csv and out_Agg_dummy_Y2.csv respectively. The objective of the project henceforth was to predict DaysInHospital for the next year for each MemeberID (i.e. given claims data in one year predict DaysInHospital for the following year).

Excel: (Cleaning the Column Names in Aggregated Data)

Creating aggregated data in pivot tables using Python adds the prefixes “Sum”, “mean”, “count”, etc. to the column names. The text was brought in to Excel and manually changed to remove unnecessary unicode characters. The files were saved and renamed the same out_Agg_dummy_Y1.csv and out_Agg_dummy_Y2.csv.

R Studio: (More Processing)

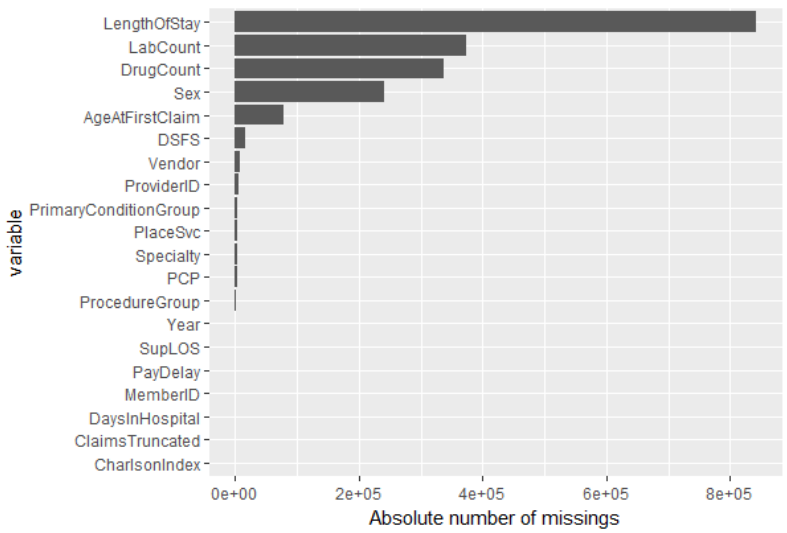
After bringing the aggregated dummy variable data files back into R Studio, further preprocessing was done. At this point the Train data dimensions were 75,832 rows by 107 columns while the Test data dimensions were 71,435 rows by 107 columns.

- Duplicate MemberID_Count column was deleted. We realized that Python automatically creates a distinct column on the aggregated key (MemberID) when running the code, so a duplicate column was created along with our original MemberID_Count column.
- Due to the fact that the Sex dummy variables were summed, fields indicating the sex of the patient had values greater than 1. This does not make sense on a high level, so fields greater than 0 within Sex_F, Sex_M, and Sex_Missing were imputed with a 1, else 0.

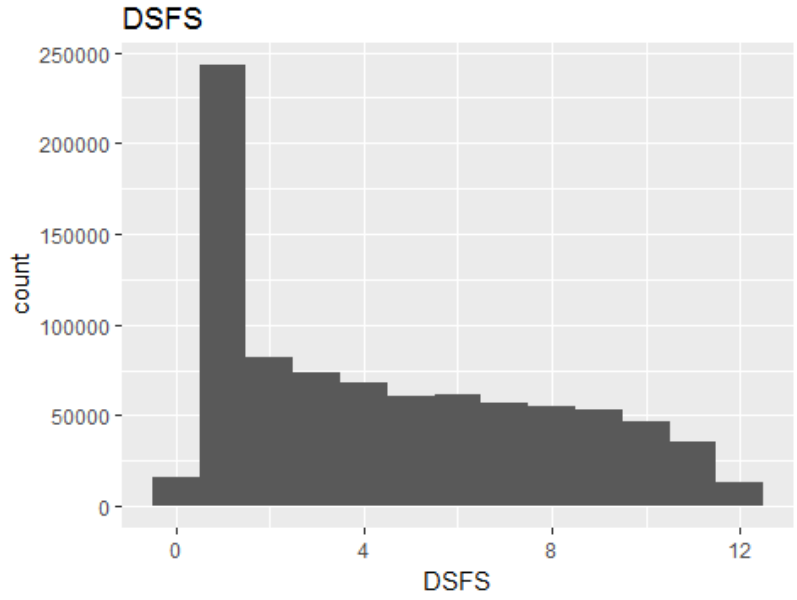
From here on out, the majority of modeling and analysis was done in R Studio. The Stream2.str file displays some attempts at association/market basket analysis and feature selection. Nothing of note was gleamed from this analysis since the Feature Selection Node did not reduce the 100+ variable dataset in the way that we had hoped. Only a few predictor variables were deemed as unimportant in predicting DaysInHospital. All variables were kept for further analysis. The market basket approach was also eliminated due to the fact that few insights were gleamed in relation to the outcome variable DaysInHospital. It was good to understand the associations between the categorical features in our dataset though.

Exploratory Data Analysis

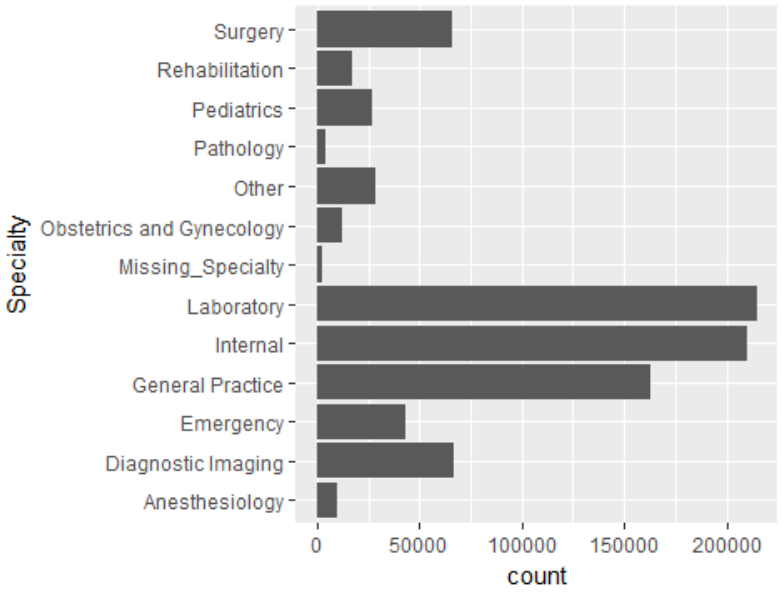
All exploratory data analysis was conducted on the Train dataset (Claims Y1 + DIH Y2) *before* being brought in to Python for aggregation/dummification.



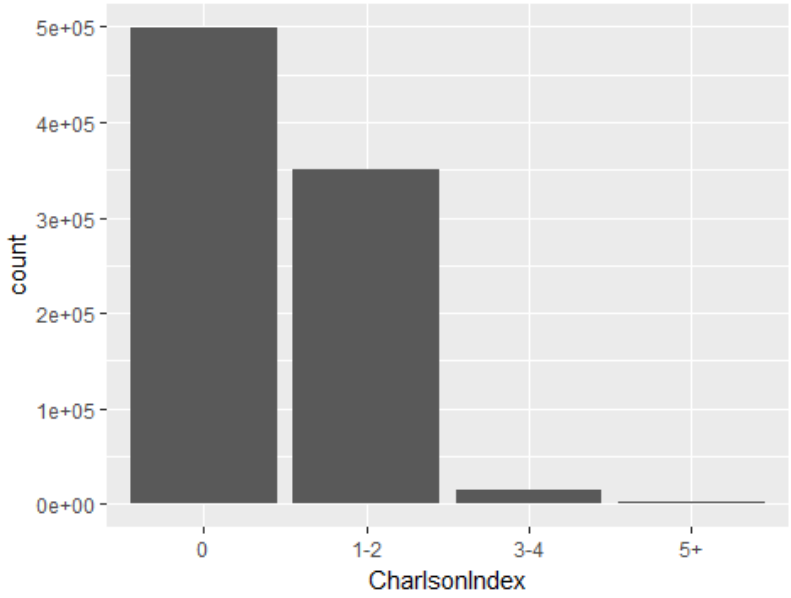
Breakdown of Missing Variables: This shows that most of the missing values pertain to “LengthOfStay”, “LabCount”, and “DrugCount”. Many records do not have patient’s Sex specified. There are values missing in AgeAtFirstClaim and few other variables as depicted.



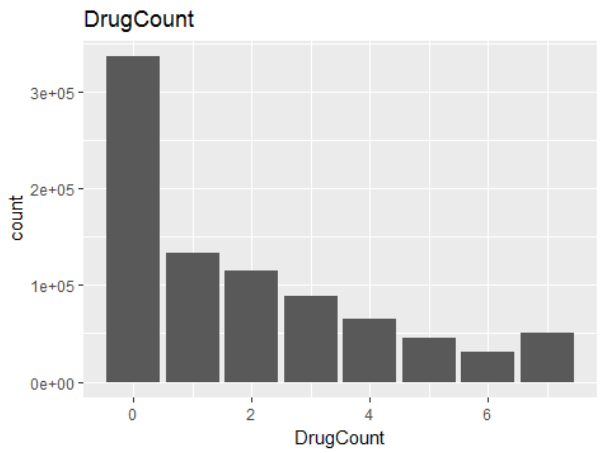
Histogram of DSFS (Days Since First Service): The majority of claims were within 1 month of the first claim of Year 1. This indicates that either there were a lot of repeat visits to the hospital within a short period of time per patient or that some patients revisted a second time after the first visit for maybe a check-up hospital appointment.



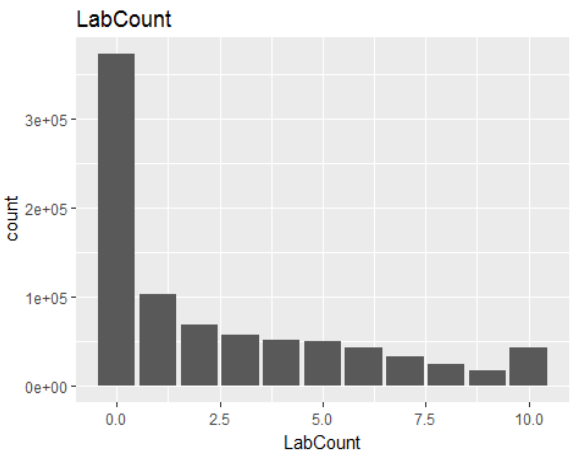
Bar Graph of Specialty: The majority of patients that visited the hospitals went to Laboratory, Internal Medicine, and General Practice specialty groups. It also seems that an equal amount of patients visited for Surgery and Diagnostic Imaging as well.



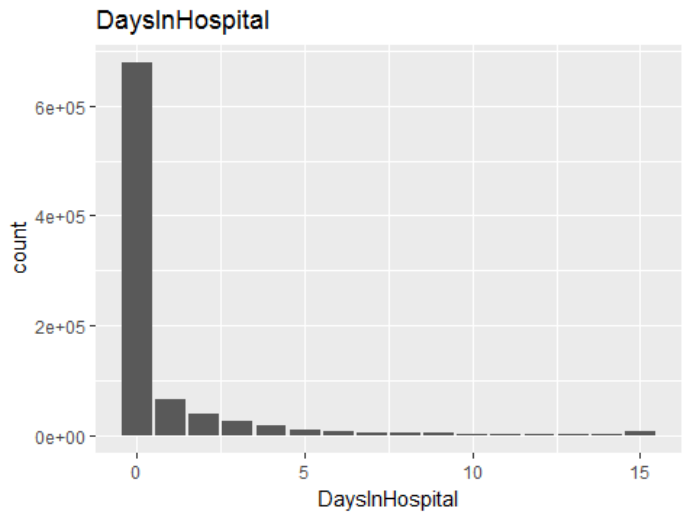
Bar Graph of Charlson Index: Patients do not typically visit the hospital for severe diseases (majority have an index of 0 while the next majority have an index of 1-2). There are very few patients that come in with a disease ranking index of 3-4 and 5+. This index measures the affect diseases have on overall illness.



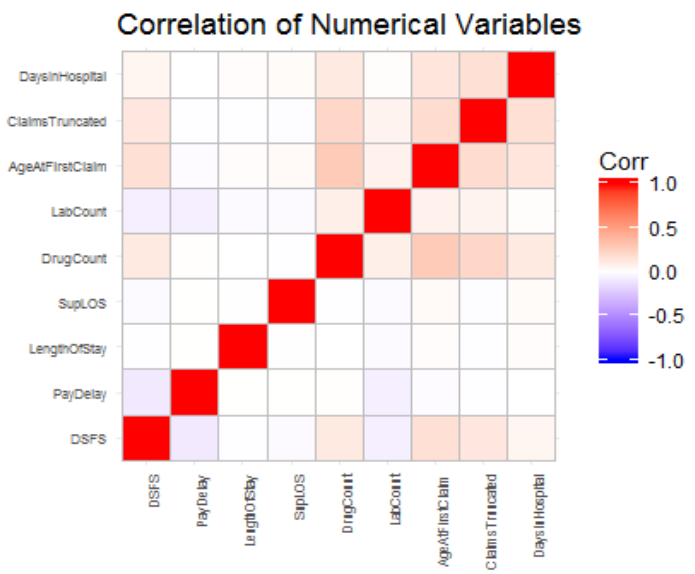
Bar graph of Drug Count: Majority of patients were prescribed no drugs.



Bar Graph of Lab Count: Majority of patients did not get lab tests done.

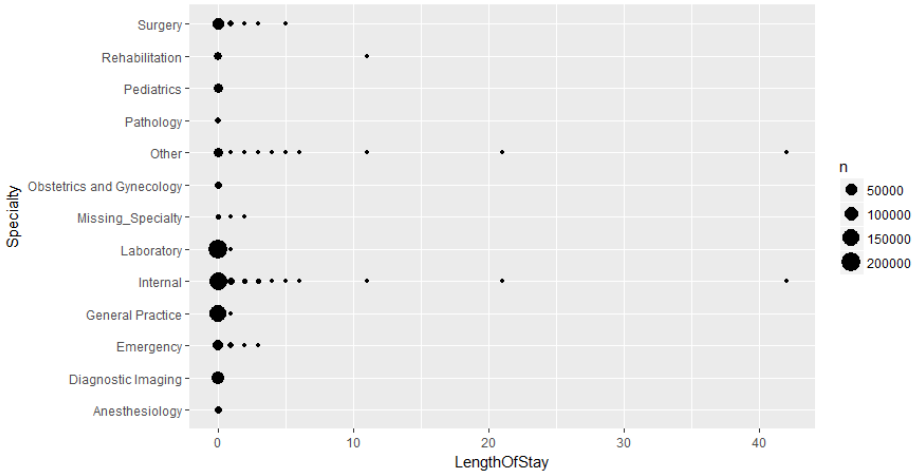


Bar Graph of Outcome Variable (DaysInHospital): Over 75% of the data (more than 600,000 records) had instances of patients that did *not* visit the hospital the following year in Year 2 when they had visits in Year 1.

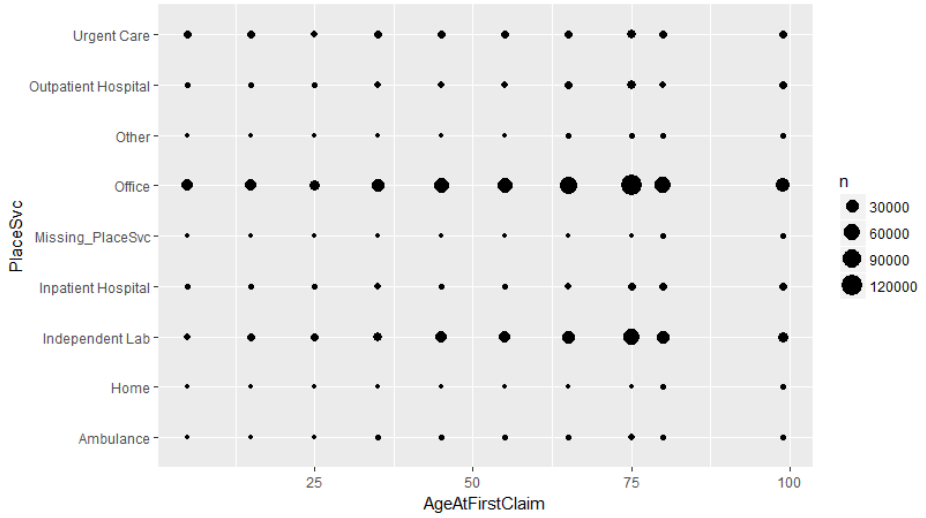


Correlation Plot of Numerical Variables: No evidence of strong multicollinearity between numerical values since the highest Pearson's correlation coefficient was 0.27255 (between AgeAtFirstClaim and DrugCount).

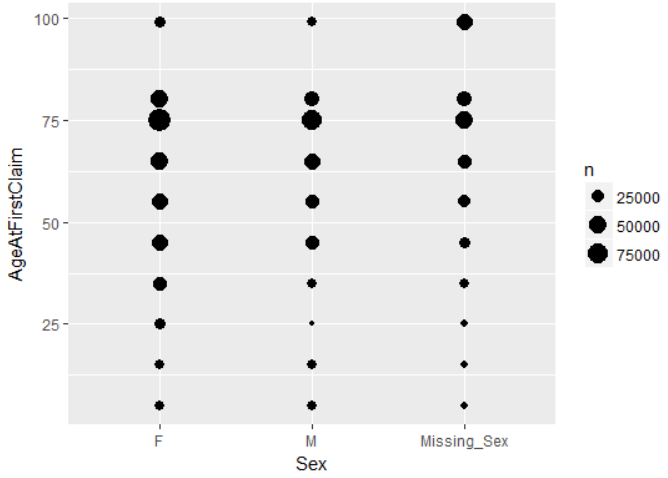
Correlation plots of Categorical Variables:



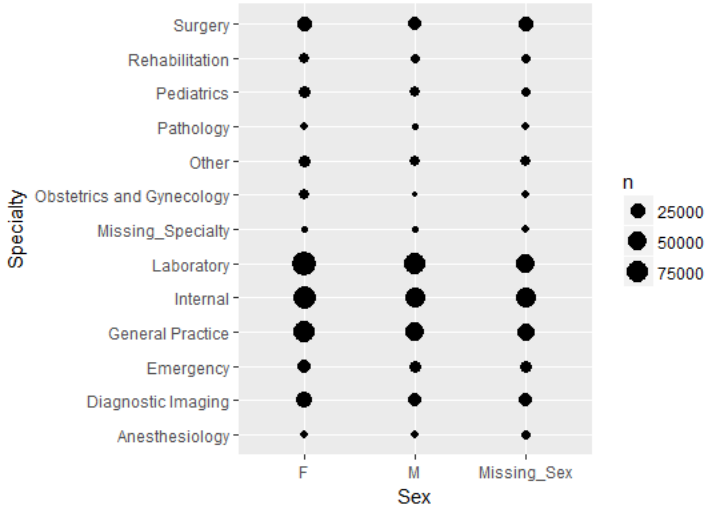
Length of Stay and Specialty: This plot shows that most of the claims were associated with 0 LengthOfStay. The records where LengthOfStay is more than zero are associated with "Emergency", " Internal Medicine", "Surgery", "Other", and few records of "Rehabilitation" as specialty.



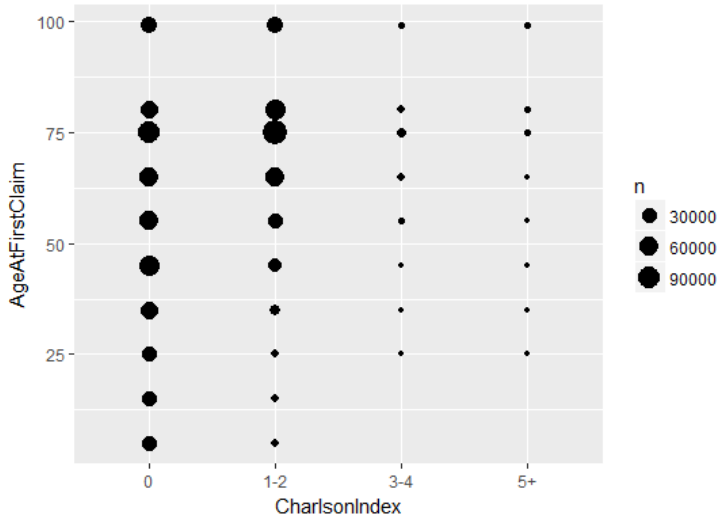
Age At First Claim and Place Of Service: This plot gives a breakdown of Place of Service and patients' age at the time of first claim. It shows that as patient's age increase the office visits & Independent Lab visits also increase. With office visits, the highest counts are for the "75" year age group. It is to be noted that "urgent care" visits have no relation with age. Every age group goes equally to urgent care. Here we have to specify that values of "99" age are to be ignored as they represent missing values.



Gender and Age At First Claim: This plot shows that number of claims increase for both Females and Males as their Age increases. The bubbles are slightly bigger for Females, so it means that Females have more claims as the age increases as compared to Males. It is to be noted that many claims do not have "Sex" specified in the records.

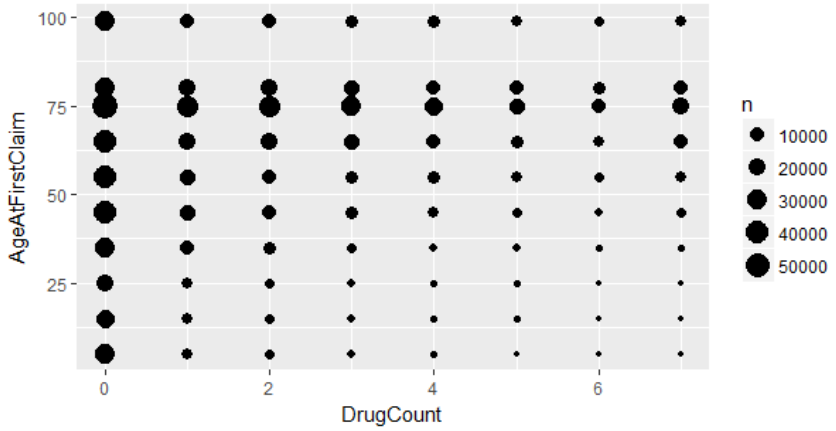


Gender and Specialty: This plot shows similar pattern for Females and Males with Females having slightly higher number of claims compared to Males within each Specialty. “Diagnostic Imaging” for Females has more claims than Males.



Definition of CharlsonIndex as per data dictionary:
A measure of the affect diseases have 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].

CharlsonIndex and Age: This plot shows that most of the claims have “CharlsonIndex” as “0” as the patients age increases. For “CharlsonIndex: 1-2” more patients belong to a higher age group (specifically age group of 75 has the most records). There are few records for “CharlsonIndex” of 3-4 and 5+, but these groups relatively have more records for higher age groups, as expected.



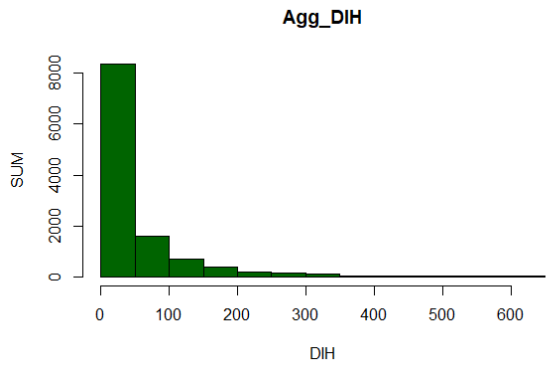
Age and DrugCount: This plot shows that majority of patients have 0 DrugCount and as the patient's age group increases the DrugCount also increases. Specially for the age group of 50-75, DrugCount is typically higher per claim.

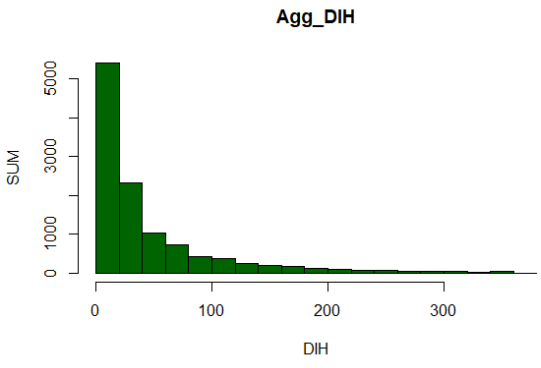
Other initial findings not outlined in the graphs:

- Vast majority of patients went to an Office for their place of service.
- Majority of PayDelay's was between 20-30 days (typically 1 month benchmark).
- Somewhat 50/50 split between patients who were Females versus Males, although there were more females. The number of missing variables Sex variables is roughly the same as the number of defined males.
- Over 200,000 patients were between the ages of 65 to 75.

Looking at the Outcome Variable after Aggregation and Dummification:

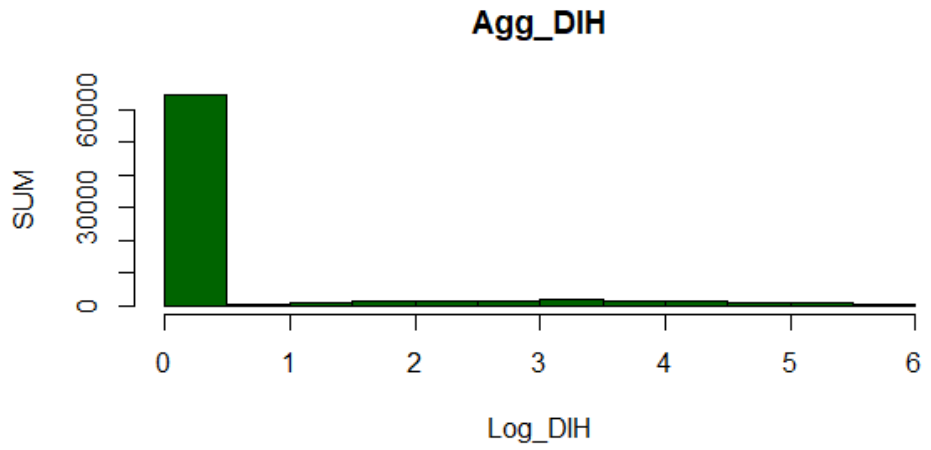
- We see the distribution is right skewed.





DaysInHosptial : This is the distribution of dependent Variables “DaysInHospital” and we note that most of the values lie between “0” & “200” but there are some values that go as high as “600” after the aggregation process. We decided to filter values which were above 365 as it is not possible to have more than 365 DaysInHospital in a year. It seems there are some erroneous entries which shows sum of hospital admission days that go over 365 days.

DaysInHospital after filtering out >365 days: After we removed records where total number of DaysInHospital was more than 365, we get the above distribution.



LogDaysInHosptial: This is the distribution of dependent Variables once we log transformed it to get around the skewness. Since there are many “0” values in DaysInHospital and log of 0 is not advisable, we added 1 to get around the issue : log(df1_Agg$DaysInHospital + 1)
This results in a lot of 0 values but the rest of the distribution is now somewhat close to a normal distribution. One consideration was to filter out 0 values which drastically improves the shape of the distribution, but then that meant that we would have ended up predicting for only those records which have some values in DaysInHospital. A log transformation of DaysInHospital was added to both the Train and Test datasets.

Modeling:

The below modeling approaches outlines our attempt at predicting DaysInHospital for Year 3 based on training the models on Claims Y1 and DIH Y2 data and predicting on Claims Y2 data.

1. Baseline Model: No Patients Return

The assumption here was that the number of days in hospital will be 0 (aka no patients return). RMSE value was calculated using the code below.

```
#### testing a base model1 with DIH = 0
rmse(log(0 + 1), df2_Agg$LogDaysInHospital)
```

This gives an RMSE = **1.374595**

2. Second Baseline Model: Mean of LogDaysInHospital

The assumption here is that the number of days in hospital in Year 3 will be the mean of the previous year's DaysInHospital (Year 2).

```
# Testing a base model2 with DIH_Y2 = mean(DIH_Y1)
rmse(mean(df1_Agg$LogDaysInHospital, na.rm = TRUE), df2_Agg$LogDaysInHospital)
```

This gives RMSe = **1.282541**

3. Linear Regression Model:

In the linear regression model, all features were to predict for DaysInHospital in Year 3. The linear model had an R squared value of 0.1381 which was very low.

RMSE = **1.187255**

Predictions are log values since we transformed our dependent variable "DaysInHospital" to a log scale. The log values of the predictions needed to be transformed back into normal values and this is the summary of the predicted DaysInHospital :

```
summary(dihY2_lm)
   Min.    1st Qu.   Median    Mean    3rd Qu.     Max.
 0.0000   0.0000    0.0000    0.6981   1.0000   148.0000
```

This shows that the mean value of DaysInHospital for Y3 was 0.6981 and that the maximum number of days a patient can be hospitalized was 148 days.

4. Lasso Regression Model:

In this model, all features were used. We optimized "Lambdas" to refine the results, so we used: lambdas <- 10^seq(3, -2, by = -.1)

RMSE = **1.259825**

```
summary(dihY2_lasso)
   Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
  0.000   0.000    0.000      0.118   0.000     44.000
```

Summary results indicate that the mean value of DaysInHospital for Y3 was 0.118 and that the maximum number of days a patient can be hospitalized was 44.

5. GLM (Generalized linear model) :

All the features were included and it was observed that RMSE score increased a bit compared to previous models.

RMSE = **1.3209**

summary(dihY2_glm)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00000	0.00000	0.00000	0.01572	0.00000	3.00000

6. Ensemble Model:

In order to improve the RMSE score, an ensemble modeling approach was attempted. In this case a weighted average of predictions from previously run models was used which included Linear Regression, Lasso Regression, and GLM. Several combinations of weights were tested and below weights were able to improve RMSE just a little bit compared to the previous best score that was achieved using linear regression (**1.187255)**

```
pred_ensemble<- 0.1*lasso.pred  + 0.04*pred_glm + 0.86*pred_lm
```

This gives RMSE = **1.186385**

summary(normal_pred)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	0.0000	0.5468	0.0000	114.0000

Insights and Conclusions

Model Summary table of all of the models run on the Heritage Health Provider Network datasets.

Model RMSE Scores					
Baseline 1	**Baseline 2**	**Linear**	**Lasso**	**GLM**	**Ensemble**
1.3746	1.2825	1.873	1.2598	1.3209	1.1864

- Overall, the best performing model for this project was the Ensemble model which weighted the performances of Linear Regression, Lasso Regression, and GLM. There is low confidence in the performance of these models, and as such the generalizability of the models is questionable. If we were to approach the Health Provider Network with these findings and they base changes in their system off of the conclusions, there will be large sources of error in predicting which patients return to the hospital the next year. This then can potentially increase the costs of hospitalization, which is what we want to reduce (not our objective!). More iterations need to be conducted (as stated in the potential considerations for future work) before a good working data analysis and model can be constructed for predicting DaysInHospital.

- Significant features based on Linear Regression:
    - DrugCount & LabCount : If patient is on more drugs/ lab visits then it indicates severity of condition which affects hospitalization.
    - PlaceSvc_Ambulance, PlaceSvc_Inpatient_Hospital, PlaceSvc_Office: If patient's place of visit was recorded as Ambulance, Inpatient_Hospital & Office then this is significant in predicting hospitalization.
    - Following Primary Condition Groups lead to higher hospitalization:
        - Acute myocardial infarction
        - Arthropathies
        - CANCER A
        - Congestive heart failure
        - Gastrointestinal bleeding
        - Gynecology
        - Hip fracture
        - All other infections
        - Liver disorders
        - Neurological
        - Ingestions and benign tumors
        - Chronic renal failure
        - SEIZURE
    - Following Specialty groups lead to higher hospitalizations.
        - Diagnostic_Imaging
        - Emergency
        - Pathology
    - Vendor is also highly significant, indicating that if a patient has many Vendors(insurance providers) associated to the claims that affects hospitalization.
- Considerations for future work:
    - Reduce the number of predictors on the dataset. With the way that the Train and Test datasets were constructed, there were more than 100 predictor variables used for predicting DaysInHospital in Year 3. The Boruta Package is a well-known feature reduction technique in R that utilizes Random Forest for feature ranking, but is very time intensive. Past projects have used Boruta on a dataset with 40 predictor variables and it took over 11 hours to run. There is no telling how long it would take on a dataset with over 100 variables. Other feature ranking packages such as forward and backward selection might be considered.
    - Potentially combine Year 1 and Year 2 Claims data into one dataset to predict on Year 3 DaysInHospital in order to conduct a different take on analysis instead of the version this project considered. Potential problems with this though are that the data might lose the historical level of detail it originally contained in the claims data.
    - Filter the dataset to predict maybe on specific Primary Condition Groups that patients are a part of. Chronic diseases such as diabetes, arthritis, hypertension, and renal failures contribute to high costs of care within the U.S. healthcare system. It might be pertinent to pick out patients diagnosed with these diseases and do a drill down analysis to predict their DaysInHospital. This triage approach is more targeted of high-cost patients, and may reduce costs in the long term. Of course, in order for the entire healthcare system to really be reformed, an analysis of overall patient care across all Primary Condition Groups will be needed to provide hospitals insights into what changes needed to be implemented in the care they give.

    - Data was found for the potential life expectancy of Males and Females given their age. It may be interesting to add that to the datasets to see if the probability of dying the following year has any effect on increasing the predictability of DaysInHospital. A lot of patients are in the elder age group, and a lot of patients do not return the following year, so this is one biased assumption that could be made.
- Limitations of Project:
    - Due to the fact that there is always inherent bias in how data analysts conduct work, there are potential limitations with the way we conducted this project. A lot of our biases were attempted to be mitigated though by using correct analysis and statistical measures.
    - The way that we aggregated our data files (i.e. distinct count, sum, or mean) was based on our choice and is potentially biased. It may have reduced the level of detail needed to predict for the outcome variable.
    - It would have been better to join time series data over two years to get a more robust high level look at how patients interacted with the hospital over time. This would have required us to combine Claims Year 1 and Year 2 data as well as DaysInHospital for Year 2 and Year 3 and predict on Year 4 data (which we did not have). We are unsure here then how well our models and analysis actually captures patient claims data *over time*, which is the point. It may be hard then to generalize our findings to an actual real world forecasting setting.
    - There is very limited literature that we could find on the usage of historical claims data for creating a general model to forecast the next year's DaysInHospitals. As stated in the previous point then, it may be hard to extend the findings of our project outside of our black box setting.
- What we learned:
    - Health data is very messy and heterogeneous! A lot of considerations need to be taken when trying to tackle a health dataset. The vast majority of our time was used to try and get the data in the actual format that we wanted (row per patient). It was also very tricky to try and justify biases and assumptions when the point is to try and be as objective as possible.
    - Interesting insights were gleamed from the project though!
    - Might be prudent to look at time series data a bit more ☺

References

1 https://medium.com/verisfoundation/one-example-of-rising-costs-in-the-us-healthcare-system-an-executive-view-62feef15c98f

2. https://www.cnbc.com/2018/03/22/the-real-reason-medical-care-costs-so-much-more-in-the-us.html

3. http://www.allhealthpolicy.org/wp-content/uploads/2017/03/Alliance_for_Health_Reform_121.pdf

4 https://foreverdata.org/1015/index.html

5 https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf