Feature Selection and Exploration for the Heritage Health Prize

Dana DiVincenzo
BANA.780

Introduction to the Dataset

The Heritage Health Prize competition was an open competition for data scientists to try to “identify patients who will be admitted to a hospital within the next year using historical claims data.”

The historical claims data used in this report was real data that was wiped of any personal information in order to follow HIPAA guidelines. The factor “Days in Hospital” is the dependant variable that participants tried to predict, using other information given about thousands of members who had received medical treatment.

In this report, I discuss the data manipulation and feature engineering I performed on the dataset, to allow for better understanding and prediction of the dependant variable.

Initial Feature Selection

After an initial inspection of the data and column descriptions, there were features I knew I would not want to include in a predictive model:

- **Provider ID** - far too many individual providers, and it's unlikely the provider will cause any change in length in hospital (if anything, it may be the other way around)
- **Vendor ID** - far too many individual vendors, and it seems unlikely that the vendor would cause any change in length in hospital
- **PCP** - far too many individual primary care physicians, and it seems unlikely that the vendor would cause any change in length in hospital
- **Length of Stay** - is 96% NA values, brings very little information to the model
- **SupLOS** - this is just a binary indication of if Length of Stay was given for the member's claim, so we can drop this as well (we know it is 96% "no")

Data Explained

I was given multiple tables which held information about the members. Here is the data dictionary for more specific information about each table.
https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf

- Members Data
- Claims Data
- Drug Count Data
- Lab Count Data
- Outcome Data
- Primary Condition Group Data
- Procedure Group Data

Splitting Claims Data by Year

- I was given data from year 1, year 2, and year 3 in the claims, drug, and labs tables
- I was only given outcome data from year 1 and year 2, and each year was in a separate outcome table
- The data in the claims, drug, and labs tables were filtered and split into separate tables based on the year, for easier analysis
- Throughout this report, I will be focusing on the data from year 1. The data in year 2 could potentially be used as a testing dataset for a predictive model later on

Initial Data Manipulation for the Claims Table

- I dropped any rows that had an NA or blank value in all columns (besides the ones I initially decided to ignore)
  - I was able to confirm that no NA or blank values were present using frequency histograms later on
- The values in the Member ID column were changed from integers to characters
  - This prevents the regression model from viewing member IDs as hierarchical (for example, we do not want a member ID of 102 to be treated as two times as large as a member ID of 51 - they simply represent different people)
- Using the information from the “Primary Condition Group Data,” a new column was created called “Condition Code” which grouped like condition groups together
  - For example, all cancers were grouped together
  - This limits the types of condition groups, and will lead to less independent variables in our regression later (easier interpretation)

Initial Data Manipulation for All Other Tables

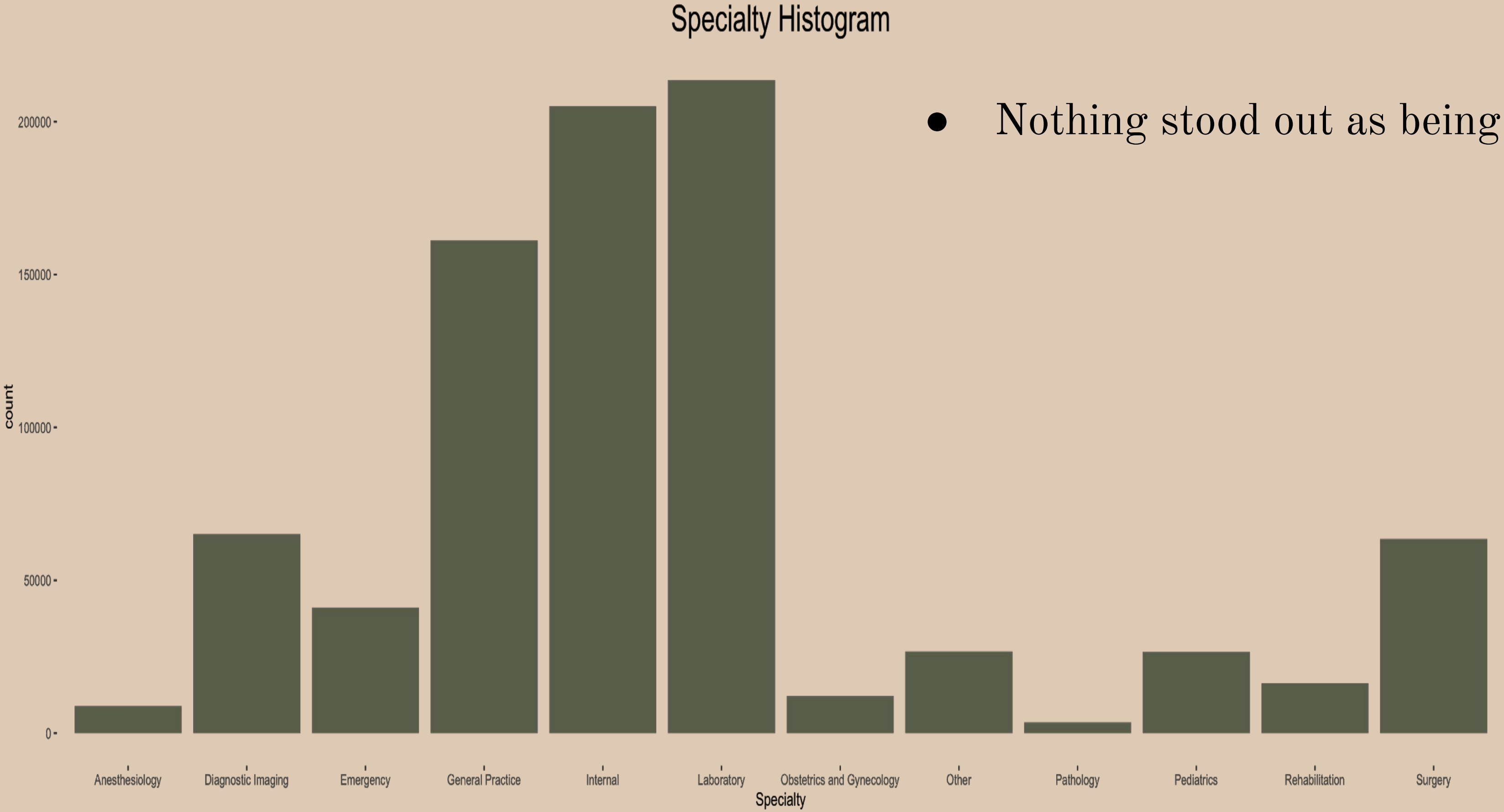
- I dropped any rows that had an NA or blank value in all columns
    - I was able to confirm that no NA or blank values were present using frequency histograms later on
- The values in the Member ID columns were changed from integers to characters

Looking at Histograms of Each Factor

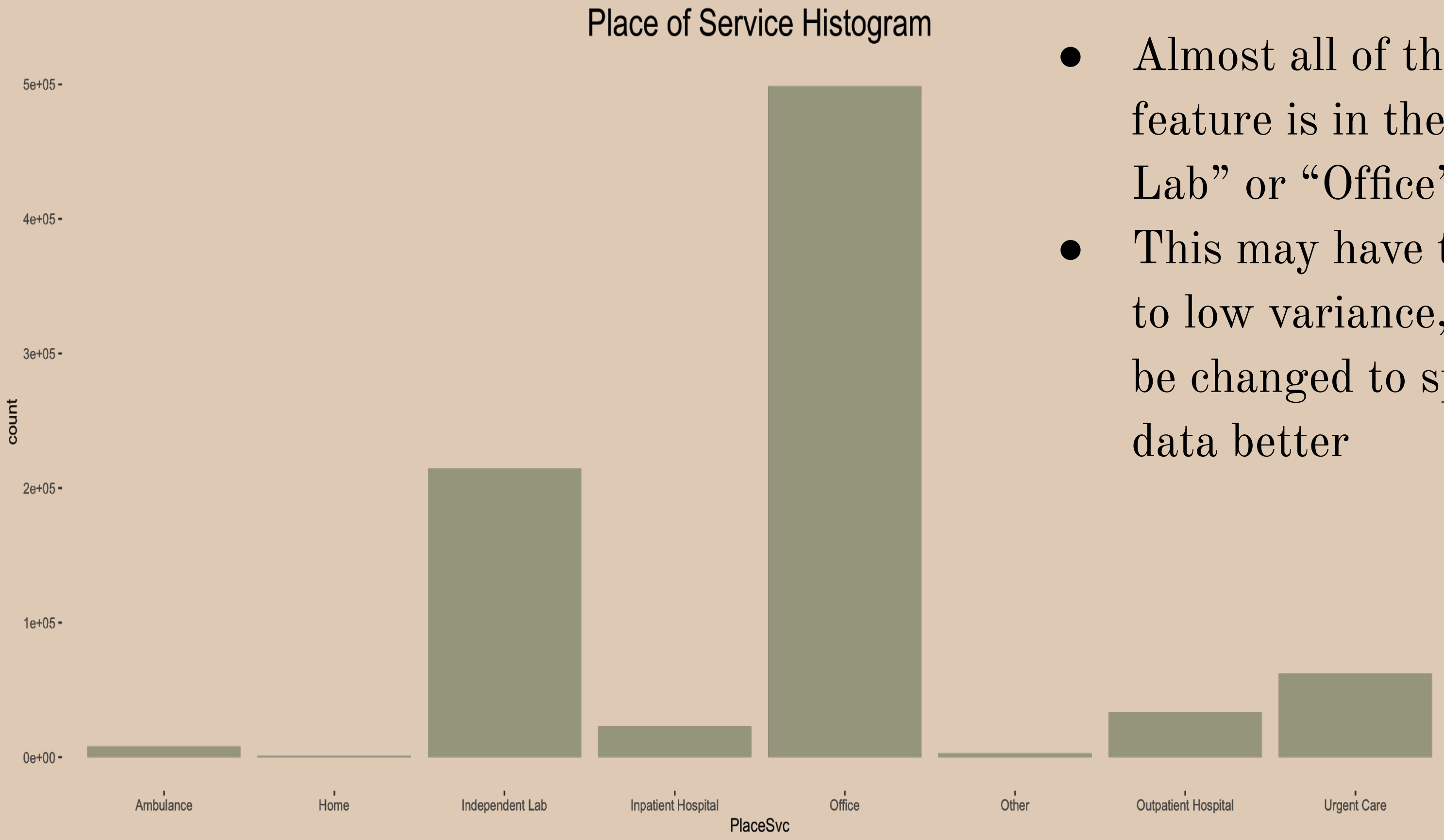
- My next step was to look at histograms of each factor from the members, claims, drug, and labs tables
- Viewing the histograms can give me information about if certain factors should be dropped due to low variability, or if certain levels within factors should be collapsed
- The following slides will show each histogram and the conclusions gained

Specialty

- Nothing stood out as being unusual

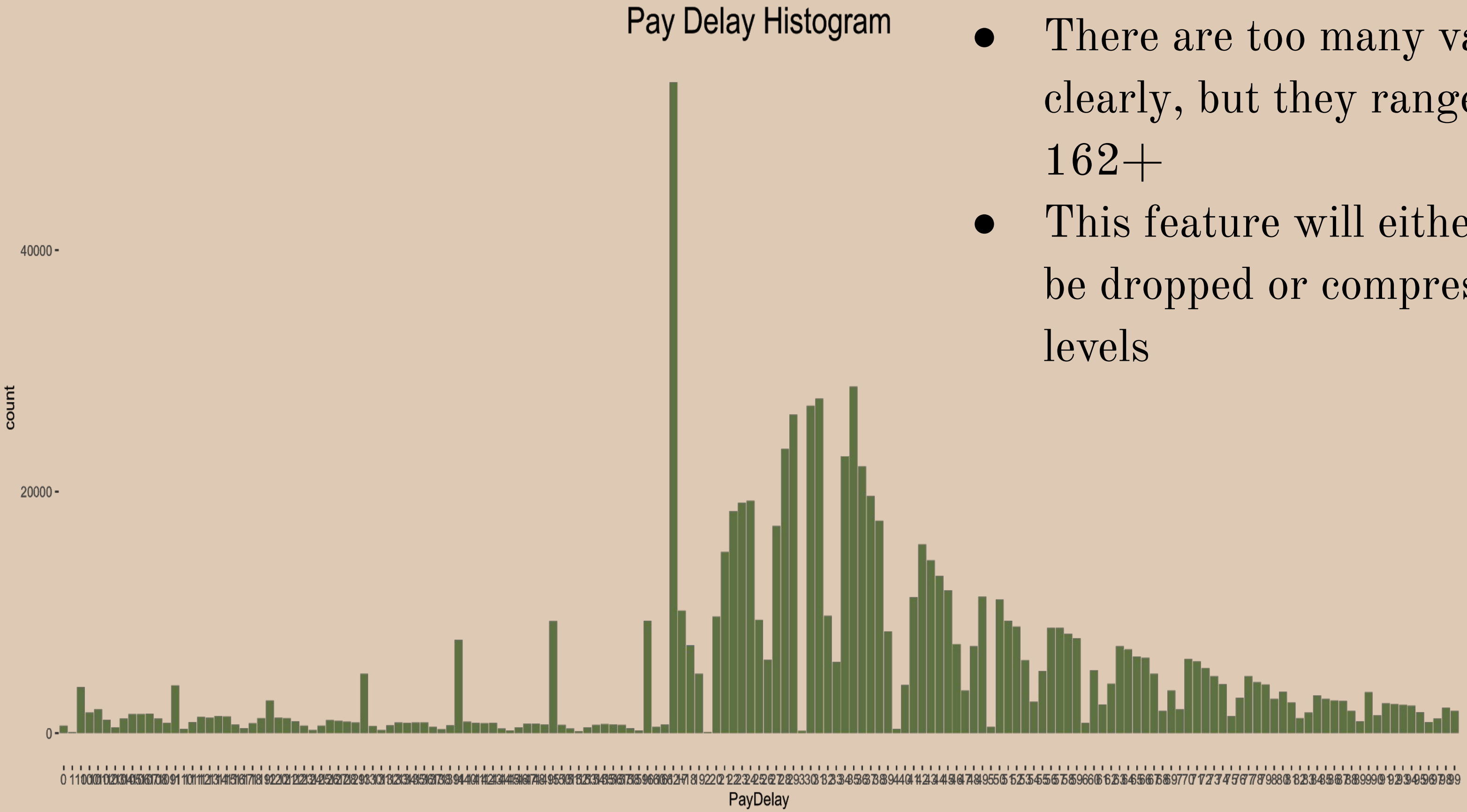


Place of Service



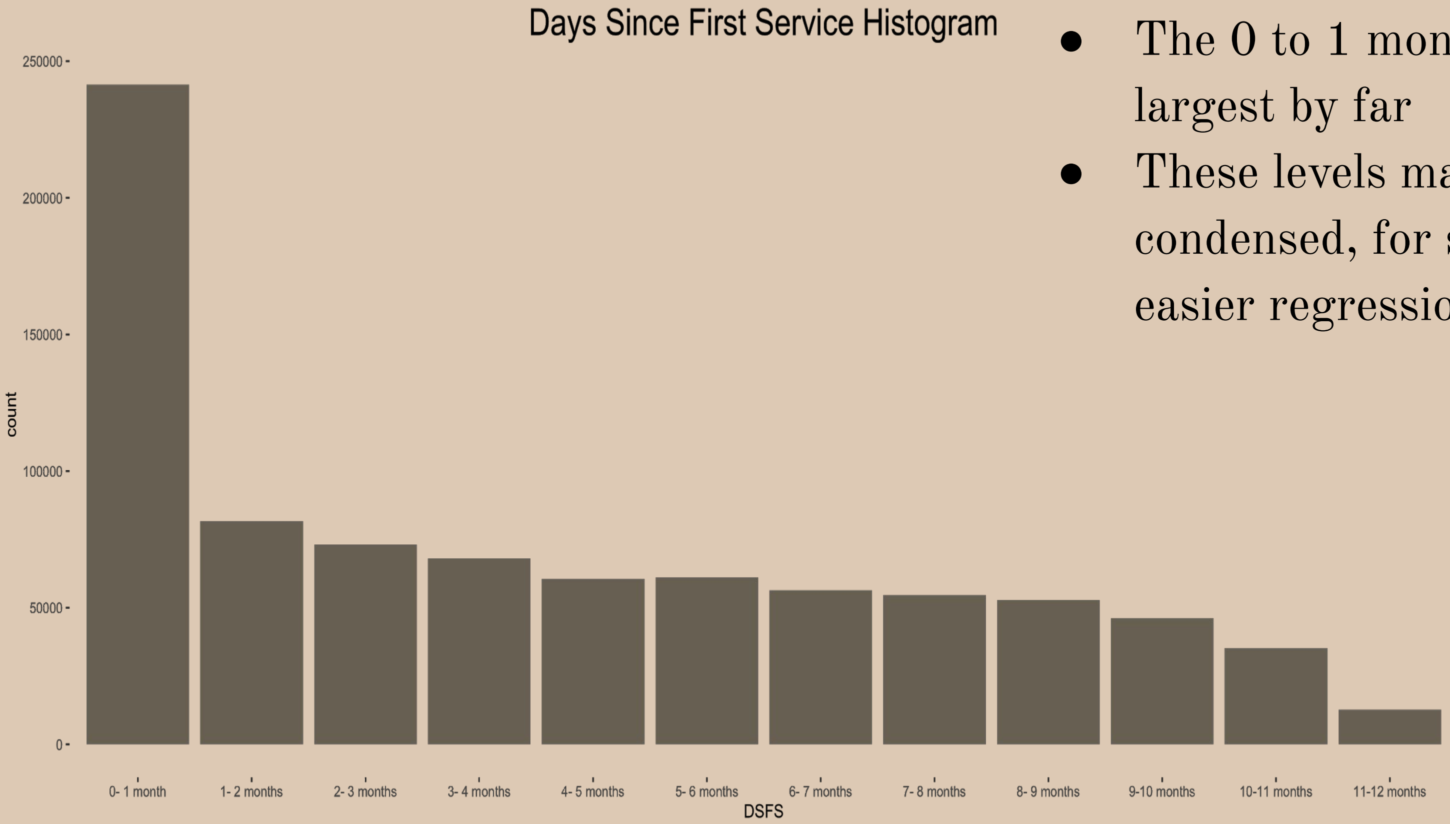
- Almost all of the data in this feature is in the “Independent Lab” or “Office” categories
- This may have to be removed due to low variance, or the levels can be changed to spread out the data better

Pay Delay



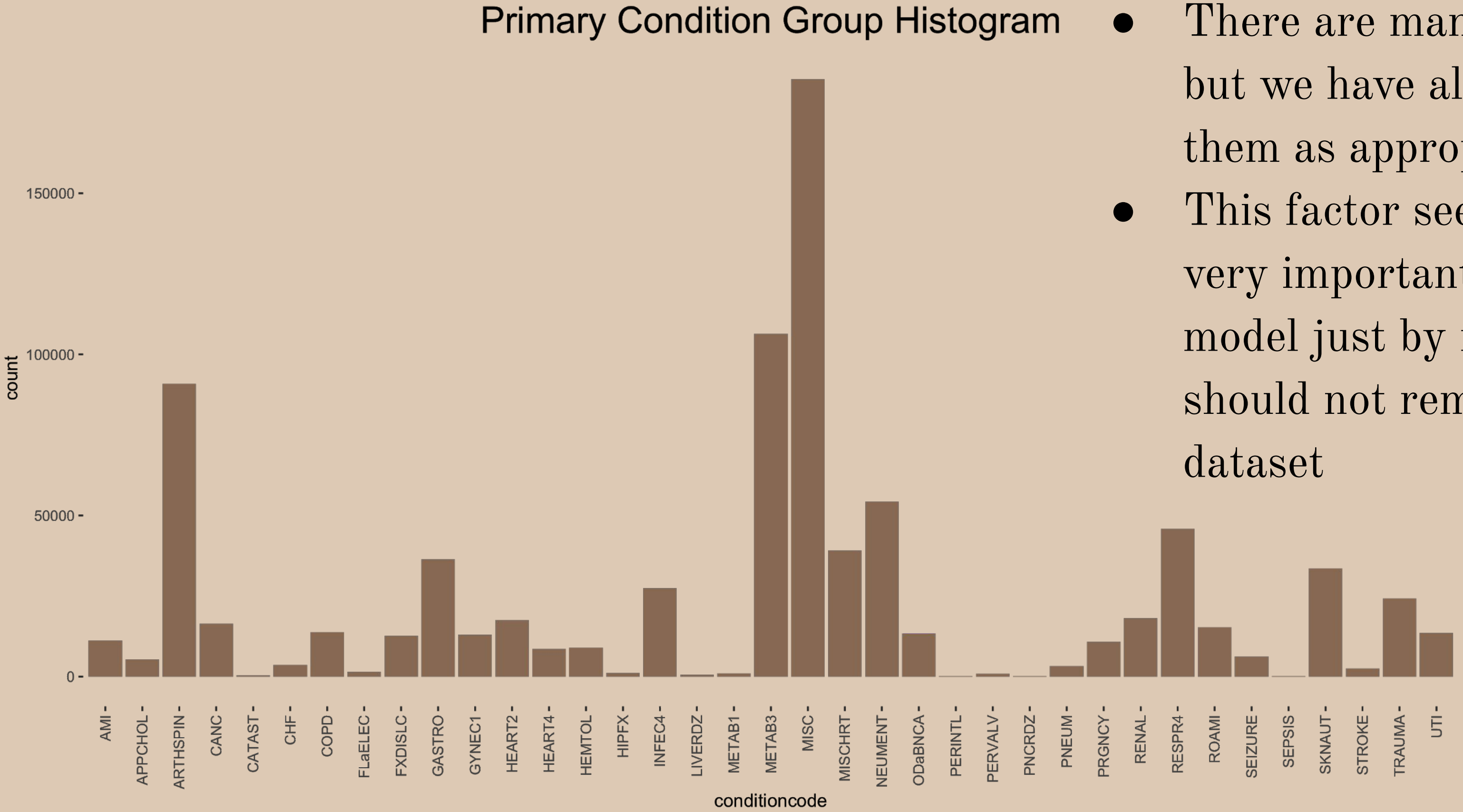
- There are too many values to see clearly, but they range from 0 to 162+
- This feature will either need to be dropped or compressed into levels

Days Since First Service



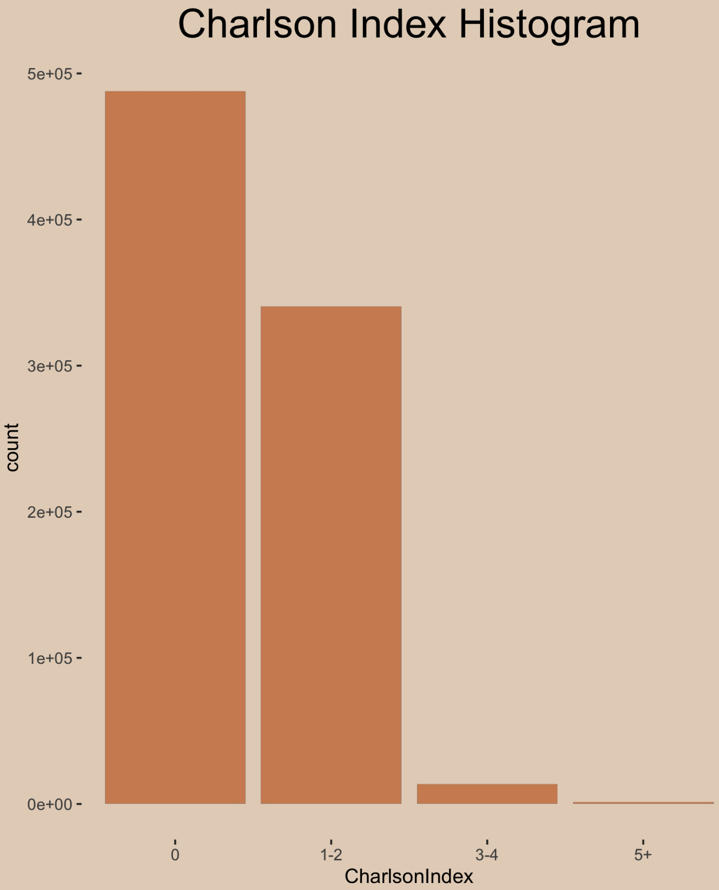
- The 0 to 1 month group is the largest by far
- These levels may have to be condensed, for simplicity and easier regression interpretation

Primary Condition Group



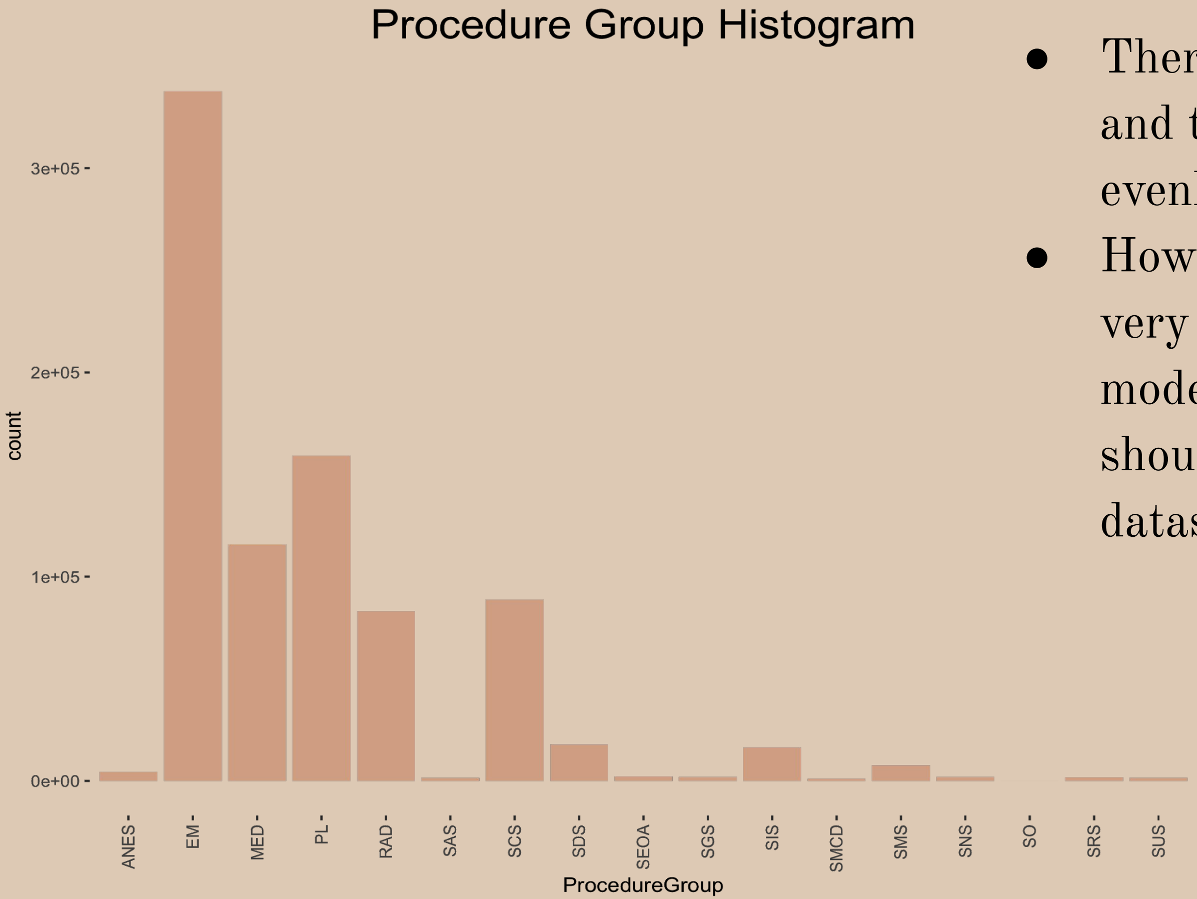
- There are many factors here, but we have already condensed them as appropriate
- This factor seems like it may be very important to our regression model just by its nature, so we should not remove it from the dataset

Charlson Index



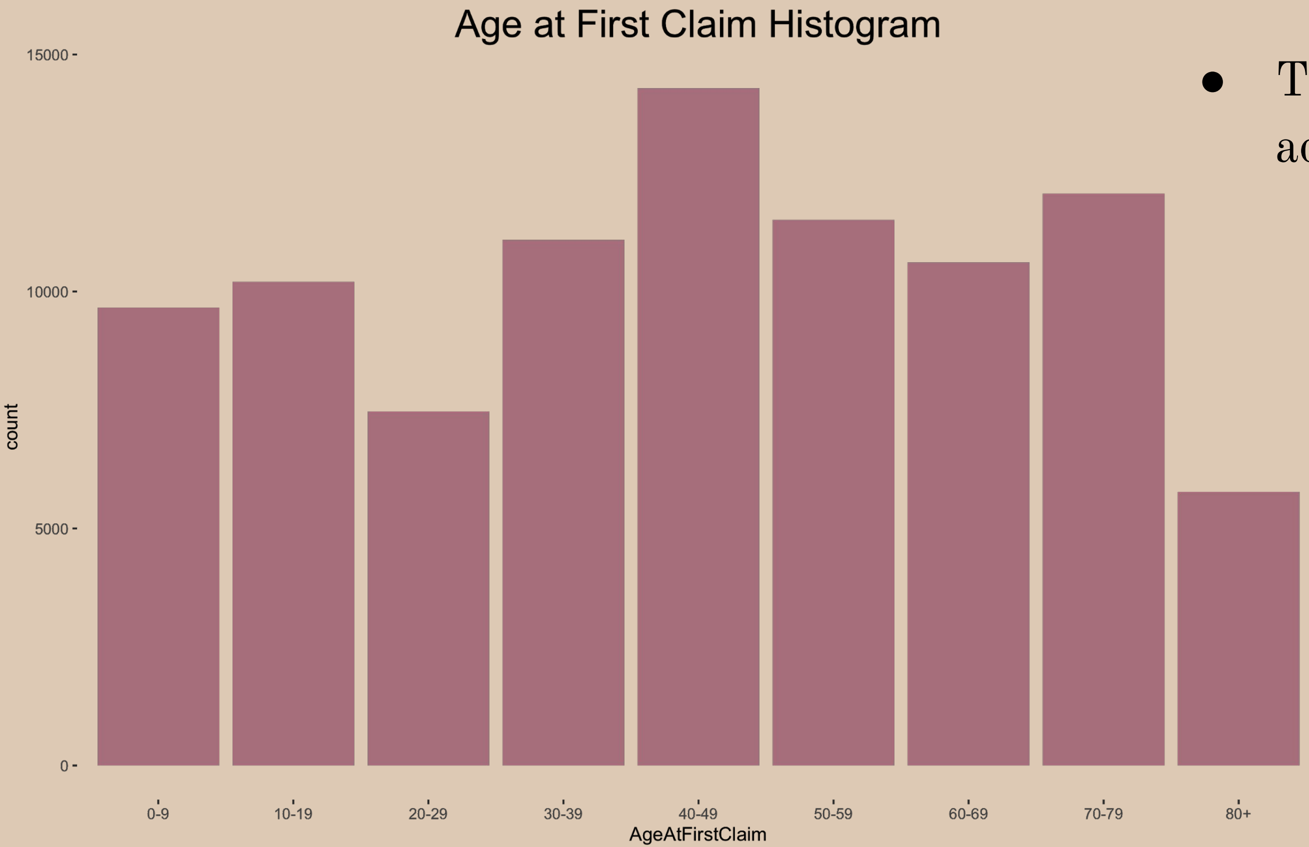
- Most of the data in this factor is within the 0 and 1-2 categories
- There is not much variance across the other categories, so we should change the levels to spread out the data better

Procedure Group



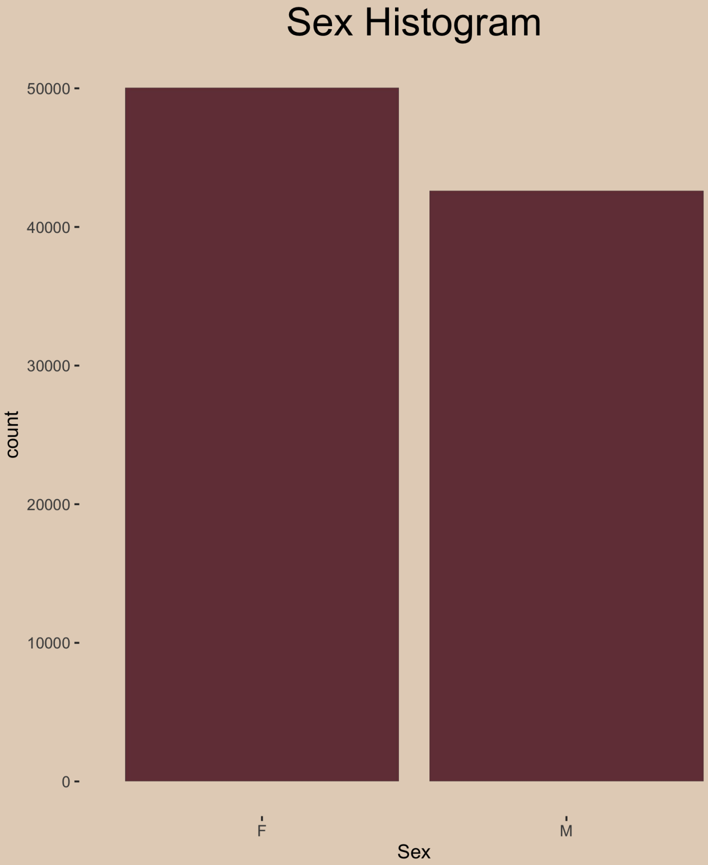
- There are many factors here, and the data is not spread out evenly between all categories
- However, this feature seems very important to our regression model just by its nature, so we should not remove it from the dataset

Age at First Claim



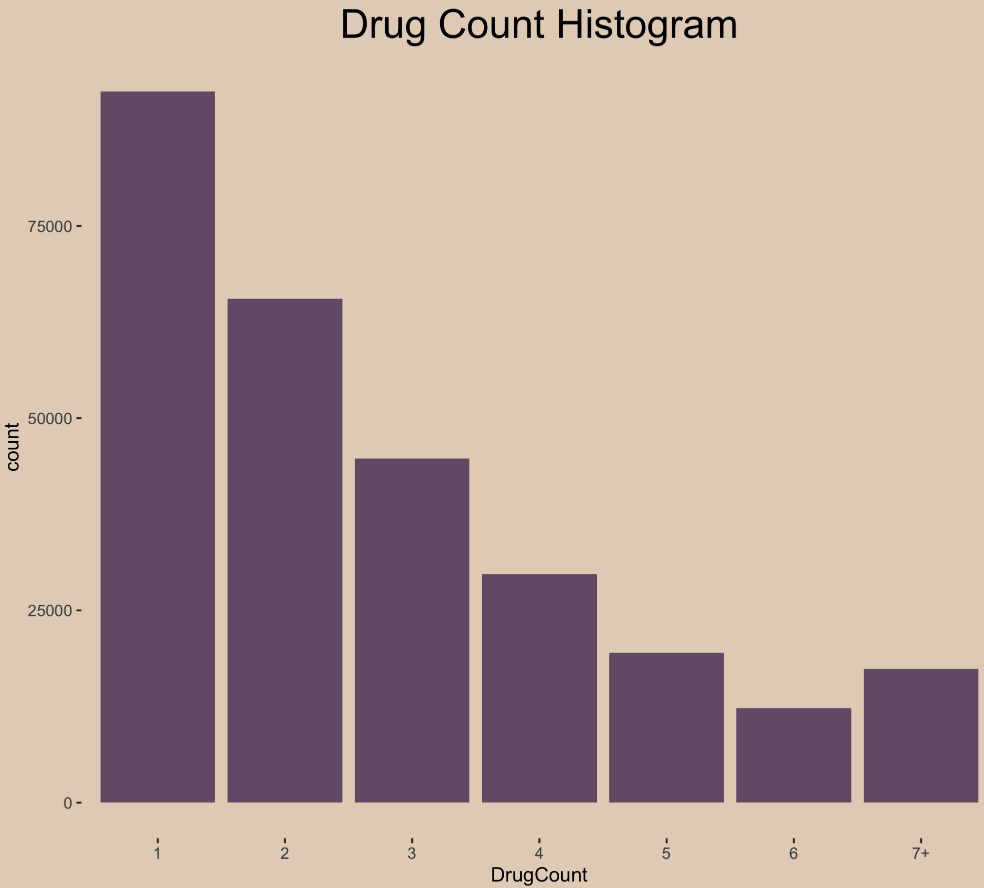
- This data is well spread out across all levels

Sex



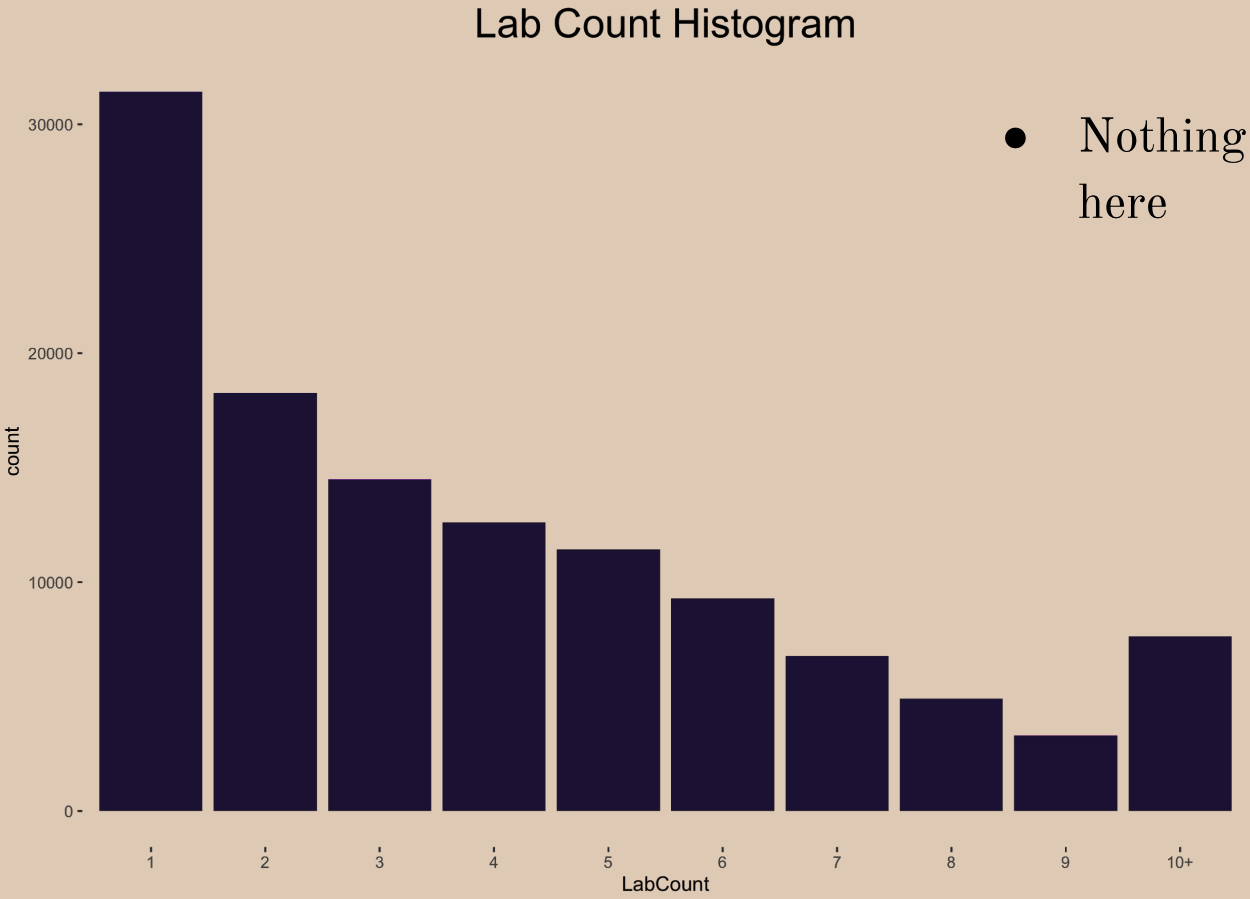
- Nothing stood out as unusual here

Drug Count



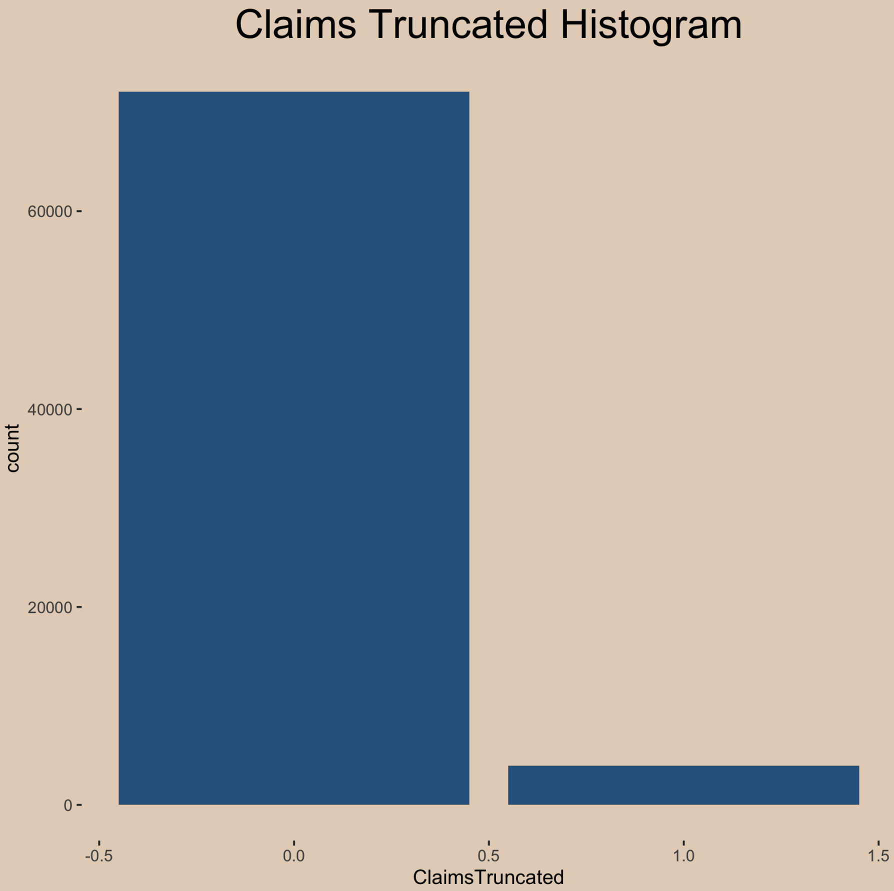
- Nothing stood out as unusual here

Lab Count



- Nothing stood out as unusual here

Claims Truncated



- This feature has extremely low variance, and should be dropped

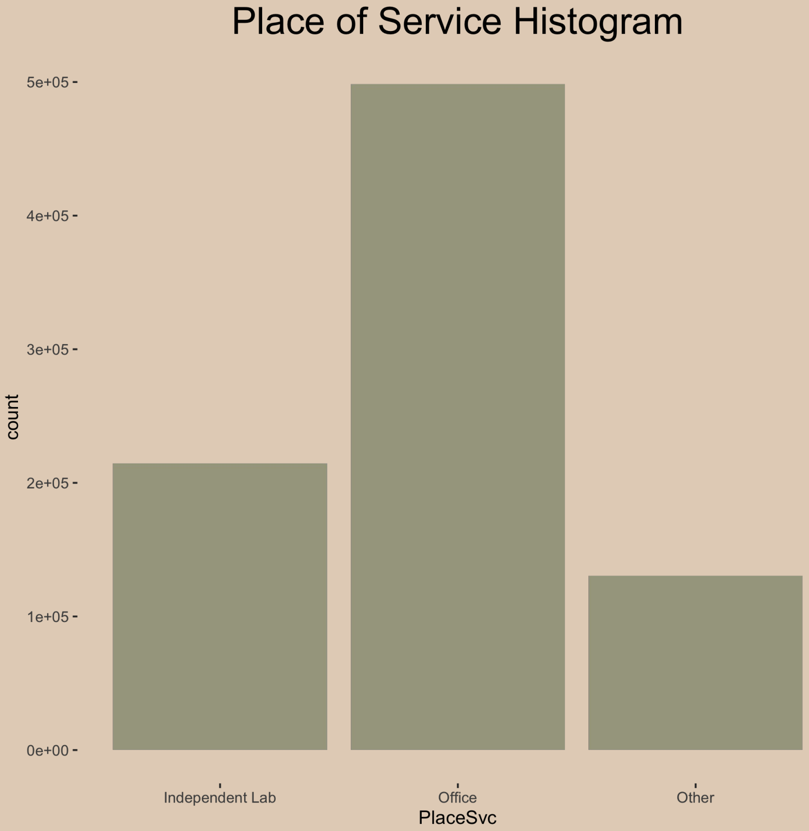
Dropping Factors

- After looking at the histograms, I made the decision to not include the following factors in my regression model:
    - Pay Delay
        - Too many factors, could be limited, but after more thought, it seems unlikely that pay delay would effect days in hospital
    - Claims Truncated
        - Variance is too low

Factor Manipulation

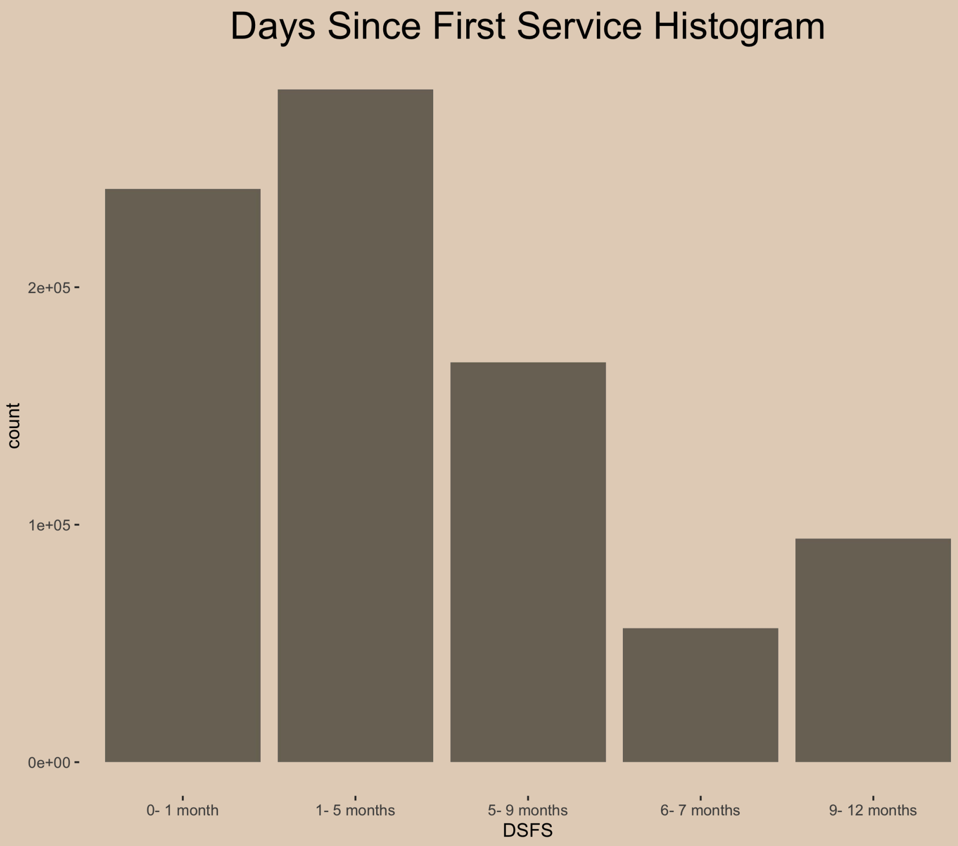
- After looking at the histograms, I decided to manipulate the following factors so they would be better suited for a regression model:
  - Place of Service
  - Days Since First Service
  - Charlson Index
- For all of these factors, the levels were changed to either spread the data across the category, or to limit the number of levels
- The changes will be shown on the following slides

Place of Service Levels Change



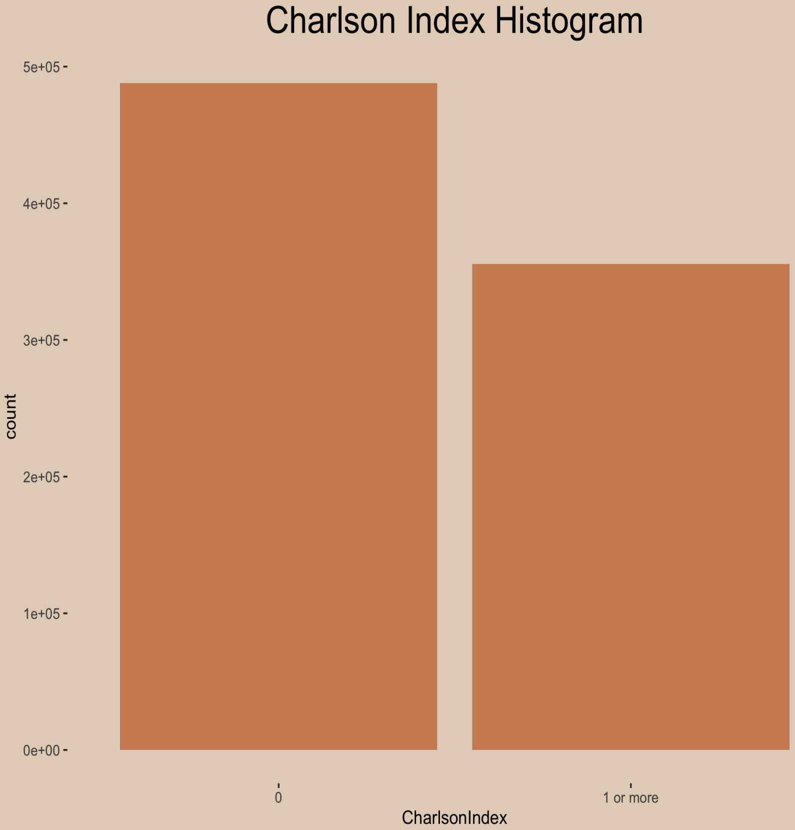
- The “Independent Lab” and “Office” levels were left alone
- All other levels were collapsed into “Other”
- Now, we no longer have many different levels with a very small proportion of the data

Days Since First Service Levels Change



- The 0-1 month group was left alone, as it held the most data
- The rest of the 1 month levels were collapsed into multi-month levels, which forces the data to be spread out more evenly

Charlson Index Levels Change

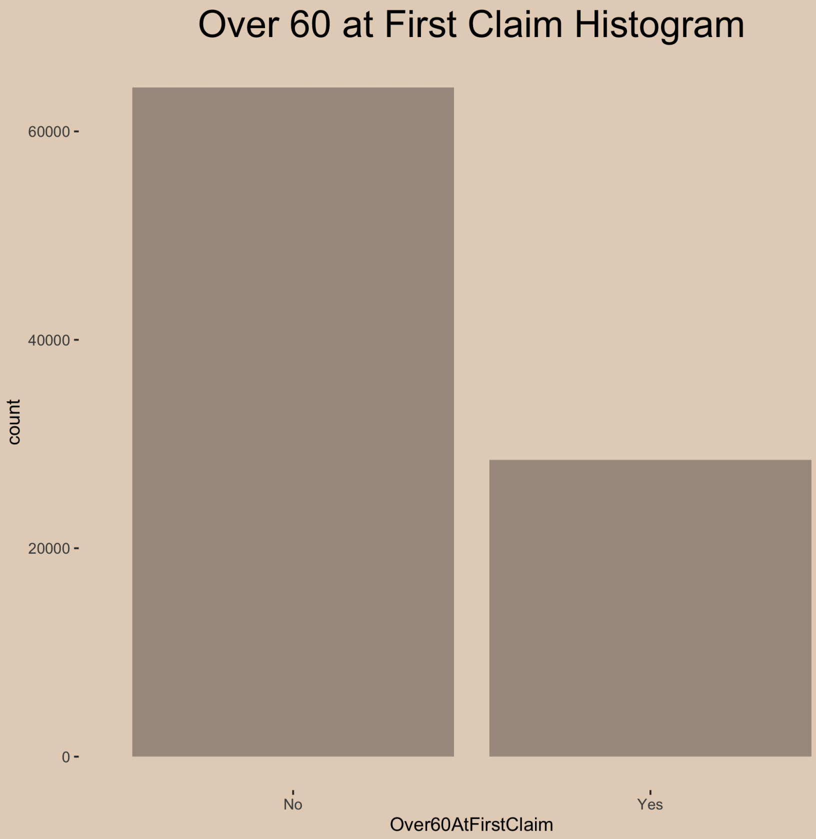


- The 0 group was left alone, as it help the majority of the data
- The rest of the levels were collapsed into a "1 or more" category
- This forces the data to be spread out more evenly

Adding Additional Features

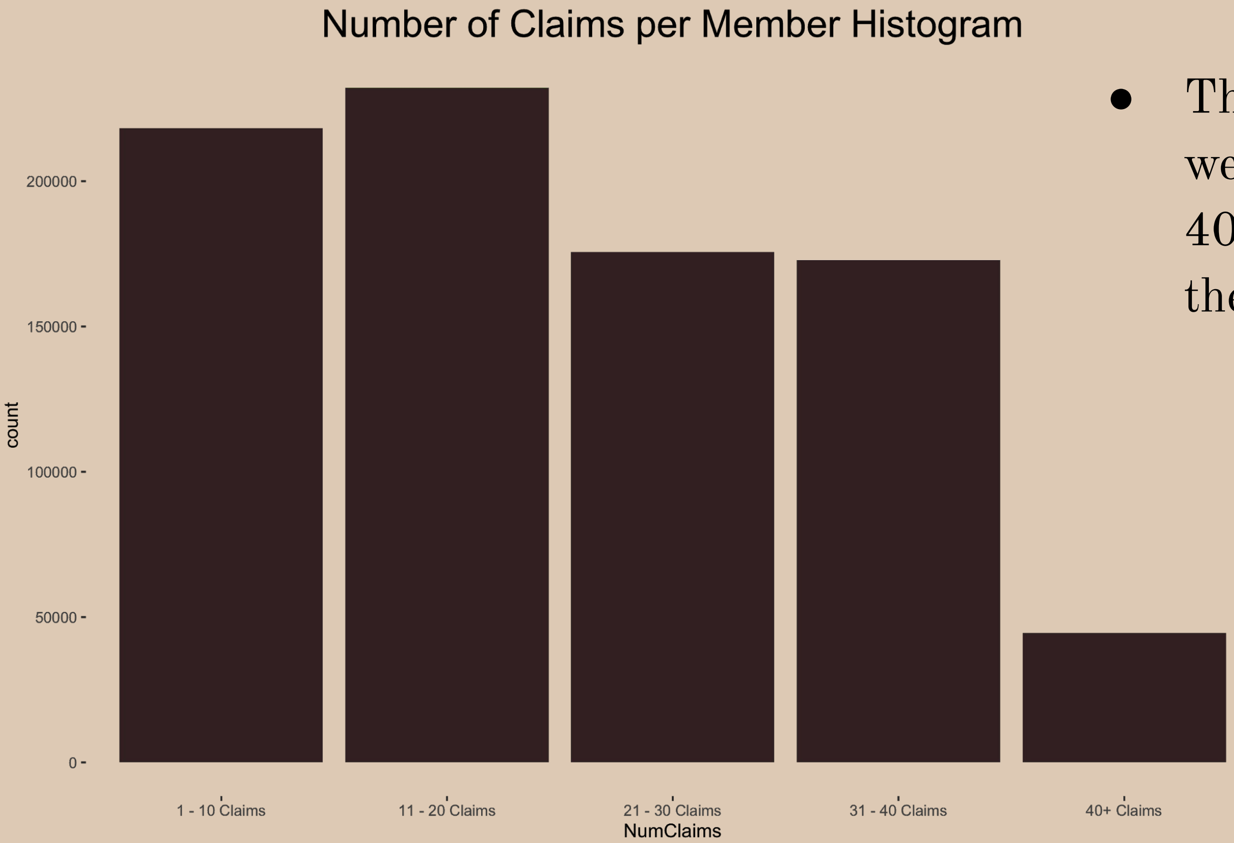
- I added 2 additional features to the dataset:
  - Age at First Claim > 60, yes or no
  - Number of Claims per Member (in categories)
- I felt that knowing if the member was over 60 would be helpful, as older people have a harder time recovering from health issues, and may be more likely to stay longer at a hospital
- I felt that knowing the number of claims per member would be helpful, because I suspect that people with more claims would have more serious health conditions, and would have to stay longer at a hospital
- The histograms of these features will be shown on the following slides

Age at First Claim > 60



- This histogram shows us that the majority of the members in the dataset are under 60 years old

Number of Claims per Member



- The data in this histogram is well spread out, with the 40+ claims category having the lowest frequency

Joining the Tables

- The members and outcome tables had one unique member per row
- However, the claims, labs, and drug tables were set up so one member's information could appear in multiple rows
  - These had to be transformed to be the same format as the member and outcome tables
  - For each selected feature, I grouped the data by member ID, and pivoted the table so the column values became the column headers
  - That way, all information about the member was held in one row
- I then joined all of the tables together into one, so it could be used in a regression analysis
  - The table had 81 columns, which is a lot of features

Running the Regression Model

- I ran a linear regression model predicting days in hospital, using all independent variables except for Member ID
- The adjusted R^2 value was .0746, which is low
    - This means that 7.46 percent of the variability in days in hospital is due to the features I selected
- It is likely that the R^2 value could be raised with further exploration and feature exploration, as the data given was very powerful and comprehensive