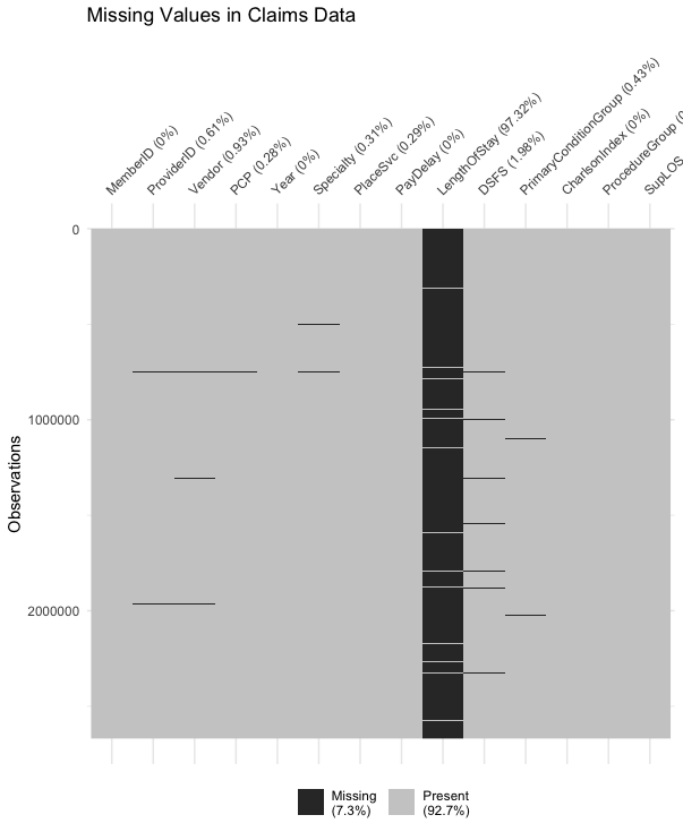
Adit Shetty

Feature Engineering

Feature Engineering is the process of extracting features (variables) through raw data via data mining. It's usually a prerequisite to data analysis, as the data needs to be prepared to suit the appropriate analyses. In this case, linear regression is the model of choice, as it will help us find which features are best in predicting the number of days spent in the hospital.



Plot 1

I checked for missing values across all datasets - my philosophy here was that if NA values > 50% of a feature, the feature would be dropped. Features with less than 50% were generally imputed with the column's mean value (if values were numeric) or replaced by the value with the highest frequency (in the case of categorical variables).

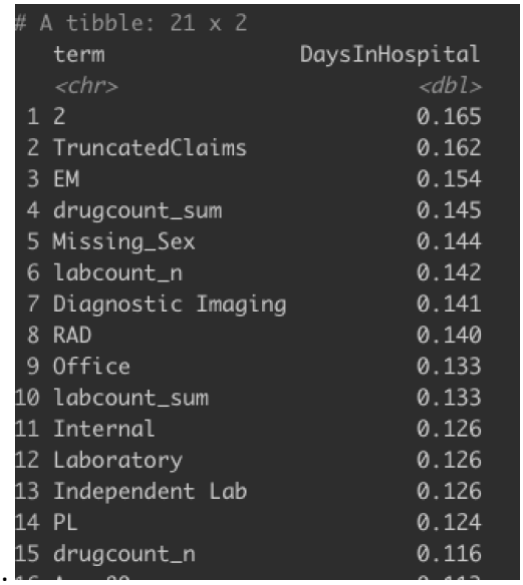
A noticeable characteristic of each dataset was that the datatype of almost every variable was wrong. Each variable was converted to its appropriate datatype, i.e., data stored as character strings were reformatted to fit the feature's purpose and converted to data type: numeric; identifying variables were converted to character strings. Each dataset was converted to a wide

format in order to merge them together, which means that each unique cell value is its own column. This reduces each member to a single row of data, with each column comprising of 1's and 0's representing a positive action (1) or a negative action (0) against that variable.

Feature Selection

Feature Selection is a crucial step in preparing the data to be analyzed. It reduces the number of insignificant features, reduces complexity, computational load, and makes it easier to interpret. If done right, it can increase the accuracy of the model. Identifiers apart from MemberID were left out of the merge as they would add unnecessary complexity to the model and increase computational times without significantly influencing DaysInHospital.

Upon examining the new dataset's summary statistics, almost every feature seemed to be skewed, with max values amounting to several multiples of the mean. All numeric features were transformed using the log1p function to fix this as a logarithmic transformation changes the scale to 0-1, allowing variables with different scales to be compared. This transformation is beneficial while running a regression model as it reduces the impact a skewed variable would have on a relatively normal variable. A correlation analysis was performed against the target variable, with the top 15 features chosen as independent variables in the linear model.



A tibble: 21 x 2

	term	DaysInHospital
	<chr>	<dbl>
1	2	0.165
2	TruncatedClaims	0.162
3	EM	0.154
4	drugcount_sum	0.145
5	Missing_Sex	0.144
6	labcount_n	0.142
7	Diagnostic Imaging	0.141
8	RAD	0.140
9	Office	0.133
10	labcount_sum	0.133
11	Internal	0.126
12	Laboratory	0.126
13	Independent Lab	0.126
14	PL	0.124
15	drugcount_n	0.116

Table 1

Regression Model

The correlation coefficients are not very high, which could be due to the modifications done to the data to prepare it for analysis. The model summary suggests that out of the 15 features

chosen for this model, 8 are statistically significant (P value < 0.05), which means that they have an impact on the number of days spent in the hospital. It is evident that there was a slight increase in DaysInHospital for a percentage increase if the Charlson Index was 2, the patient underwent the EM procedure, were serviced at an Independent lab, or if the patient had a higher percentage of prescription drugs. ("Missing" Sex was not included as it could be attributed to either gender). There was a slightly larger increase in days spent at the hospital if the patient had truncated claims from the previous year. On the other hand, there was a decrease in days spent at the hospital if the patient was serviced at the Office, and if they had a lower percentage of drug prescriptions. This model explains approximately 6.3% of the variation in DaysInHospital.

```
Call:
lm(formula = DaysInHospital ~ two_CI + TruncatedClaims + EM +
    drugcount_sum + Missing_Sex + labcount_n + Diagnostic_Imaging +
    RAD + Office + labcount_sum + Internal + Laboratory + Independent_Lab +
    PL + drugcount_n, data = dfinal)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71123 -0.20618 -0.12591 -0.06856  2.72320

Coefficients:
                   Estimate Std. Error t value             Pr(>|t|)
(Intercept)         0.035528   0.012465   2.850              0.00437 **
two_CI              0.018605   0.002659   6.998     0.00000000000265 ***
TruncatedClaims     0.297761   0.018989  15.681 < 0.0000000000000002 ***
EM                  0.061208   0.009051   6.762     0.00000000001385 ***
drugcount_sum       0.063636   0.007410   8.588 < 0.0000000000000002 ***
Missing_Sex         0.186253   0.010453  17.818 < 0.0000000000000002 ***
labcount_n         -0.004453   0.016837  -0.264              0.79140
Diagnostic_Imaging -0.073681   0.044244  -1.665              0.09586 .
RAD                 0.012269   0.011002   1.115              0.26476
Office             -0.039318   0.007897  -4.979     0.00000064361996 ***
labcount_sum        0.010909   0.006934   1.573              0.11568
Internal           -0.003441   0.003892  -0.884              0.37668
Laboratory          0.083775   0.043638   1.920              0.05490 .
Independent_Lab     0.026975   0.011923   2.262              0.02368 *
PL                 -0.002333   0.014481  -0.161              0.87200
drugcount_n        -0.077269   0.011323  -6.824     0.00000000000902 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.478 on 28417 degrees of freedom
  (21808 observations deleted due to missingness)
Multiple R-squared:  0.06338,   Adjusted R-squared:  0.06289
F-statistic: 128.2 on 15 and 28417 DF,  p-value: < 0.00000000000000022
```

Table 2

Conclusion

I made several decisions concerning the data that may have significantly impacted the model's predictions. The Lookup workbooks were not utilized as the data required was already contained in the claims dataset. Given the model's constraints, and the resulting correlations produced by the features, an R^2 of 0.63 is acceptable. For further improvement of data, it would be helpful to clean the data and run several different feature selection methods to

choose the best subset of features. Additionally, a tedious but useful approach could be running the regression model with several combinations of features could be to finding the optimal model.