Practice Exam – Patient Length of Stay (SOA PA 6/16/20) Solution

Task 1 – Edit the data for missing and invalid data (8 points)

The following changes have been made to the variables to fix missing values and erroneous data types.

Gender

The three missing values are simply removed as this small number is unlikely to significantly affect the outcome.

Admit_type_id

This variable was originally labeled as numeric but is actually a factor so the data type has been corrected. The base level is set to "Emergency" as it has the most observations. There are 2,021 cases where this information was labeled "Not Available", although it is unclear whether this indicates a missing value or an intentional omission. The missing values were not removed as there could be some significant reason for their absence. For example, some patients with omitted values may have deliberately decided not to answer due to medical reasons. I chose to include these values and test if they will be predictive.

Race

The 226 missing values are grouped into an "other" category. There are only few records with certain levels (e.g. race = Asian) and therefore those are combined with the "other" category to enhance credibility.

A pattern in the reasons why these values are missing would cause these results to be biased. For instance, asking patients to identify their race during data collection might have caused certain groups to omit some information.

Weight

As there were 9,688 values missing out of the overall 10,000 records, this variable was removed. It does not matter why this variable was missing so often; there are simply not enough records to be useful.

Num_ip

There are 12 patients who each had nine inpatient visits within the prior 12 months. There were also 15 patients who had eight visits each. Although these numbers seemed suspiciously high, I chose to include them but recommend further investigation to ensure they are actually comparable to other patients in this analysis.

Num_diags

As the following chart demonstrates, many patients suffered from multiple ailments, with most reporting between one and 10 diagnoses. For example, 4,914 patients had a total of nine diagnoses. After nine however, the numbers drop abruptly, showing only two patients with 10 diagnoses. As this factor is described as "Number of diagnoses entered to the system in the twelve months preceding the

encounter" it could be that these two records are errors or undercounts. I chose to leave these as they are.

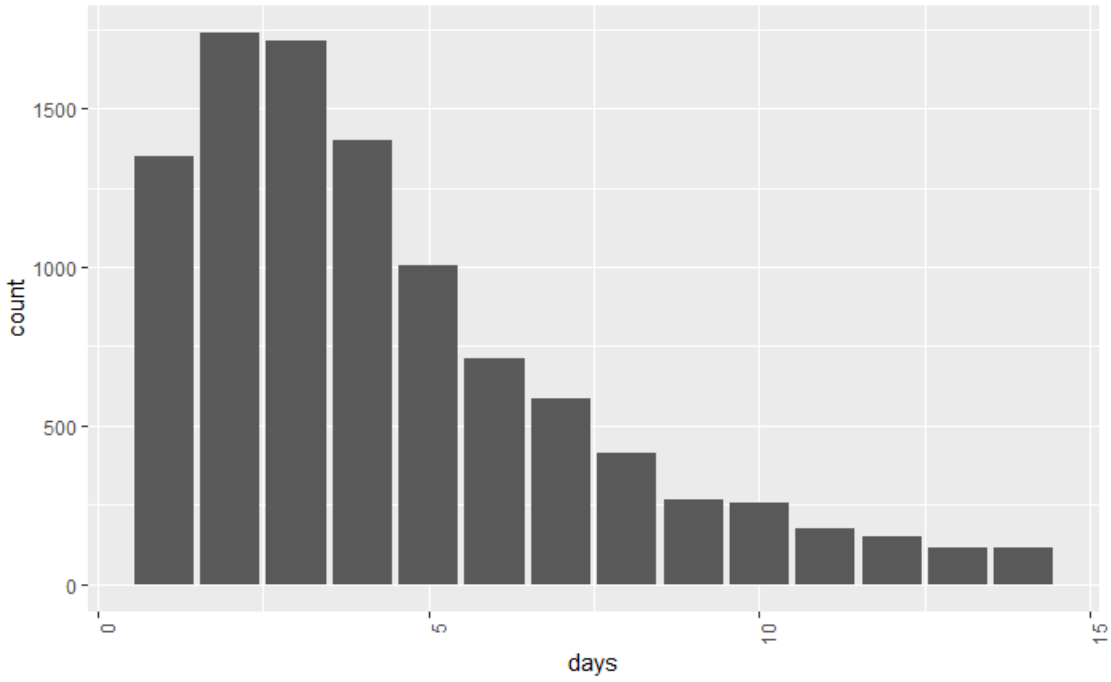
num_diags	n
1	23
2	96
3	267
4	570
5	1063
6	992
7	996
8	1069
9	4914
10	2

Lastly, I converted the factor's base levels to those which had the most observations. These values are the first ones listed in the chart below, including Gender = Female, age = [70-80], etc.

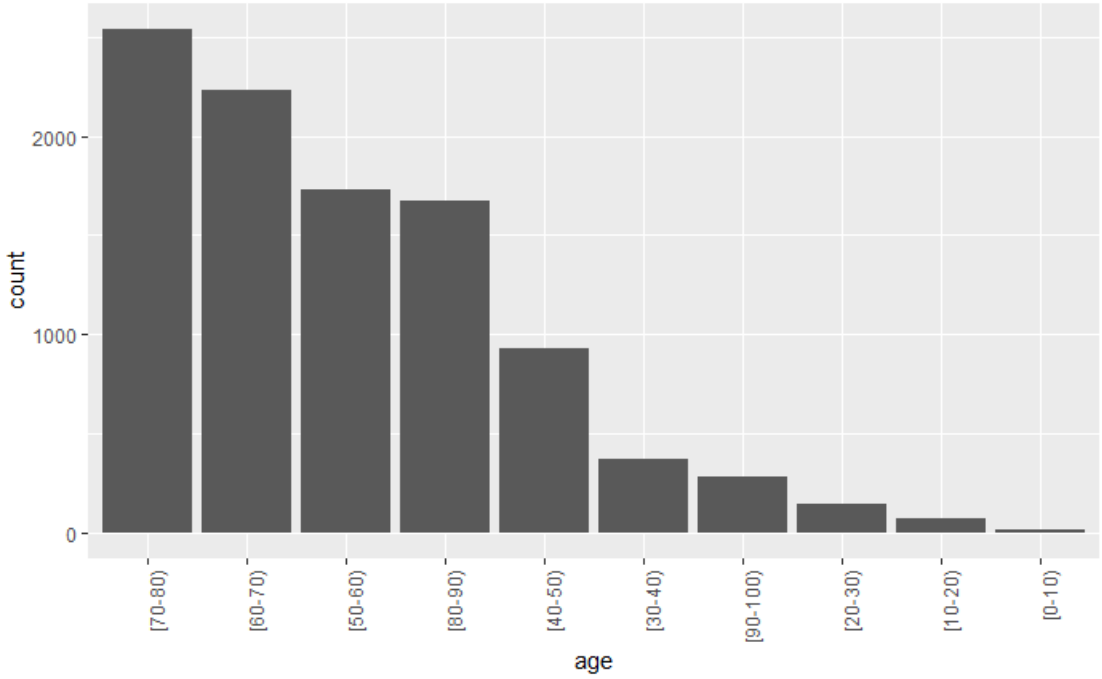
Task 2 – Explore the data (15 points)

The target variable is the number of days that a patient spends in a hospital after admission. This has a mean of 4.4 and a median of 4.0. It is right-skewed, as evident in the histogram below.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   4.000   4.409   6.000  14.000
```



Older people tend to have longer hospital stays than younger people and most patients are over 50 years old. Therefore, it might be expected that the age variable would be predictive of the length of stay. This is in fact reflected in the data, as the median length of stay for people over age 50 is 3 - 4 days but is only 2 - 3 days for younger patients.



age	mean	median	n
[70-80)	4.658	4	2541
[60-70)	4.408	4	2228
[50-60)	4.091	3	1726
[80-90)	4.8174	4	1676
[40-50)	3.9914	3	931
[30-40)	3.7984	3	372
[90-100)	4.7396	4	288
[20-30)	3.5172	3	145
[10-20)	3.125	2	72
[0-10)	3.2222	3	18

The above table shows significant differences in the mean length of stays for different age categories.

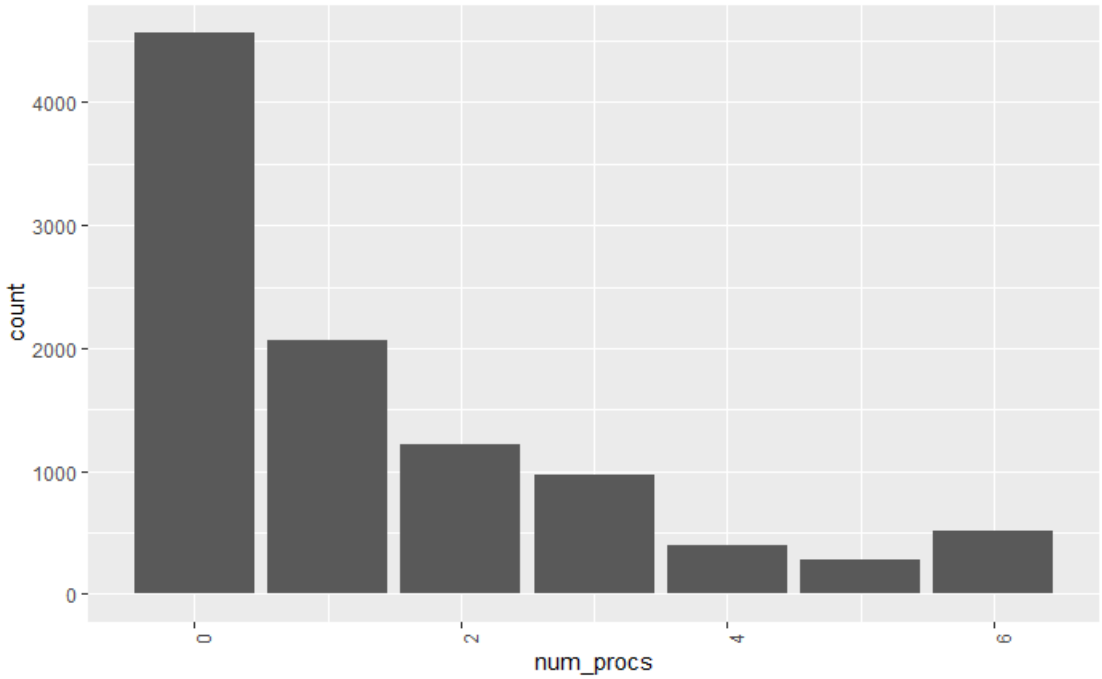
Num_procs

Unhealthy patients may need to undergo more procedures and are therefore likely to have longer recovery times and lengths of stay. Therefore, this category may also be predictive. It only includes the preceding 12 months, and there is an even spread of observations. Most patients have not had any procedures, and have the lowest average length of stay, 3.78 days. Patients who have had five

procedures have an average stay of 5.4 days. This makes sense because patients who have had more procedures probably need more time to heal.

num_procs	mean	median	n
0	3.7786	3	4561
1	4.5586	4	2064
2	4.9984	4	1223
3	5.0402	4	969
4	5.6317	5	391
5	5.2482	4	274
6	5.435	5	515

The correlation between the number of procedures and length of stay was 19%, with hospital stays lengthening as the number of procedures increases.



Readmitted

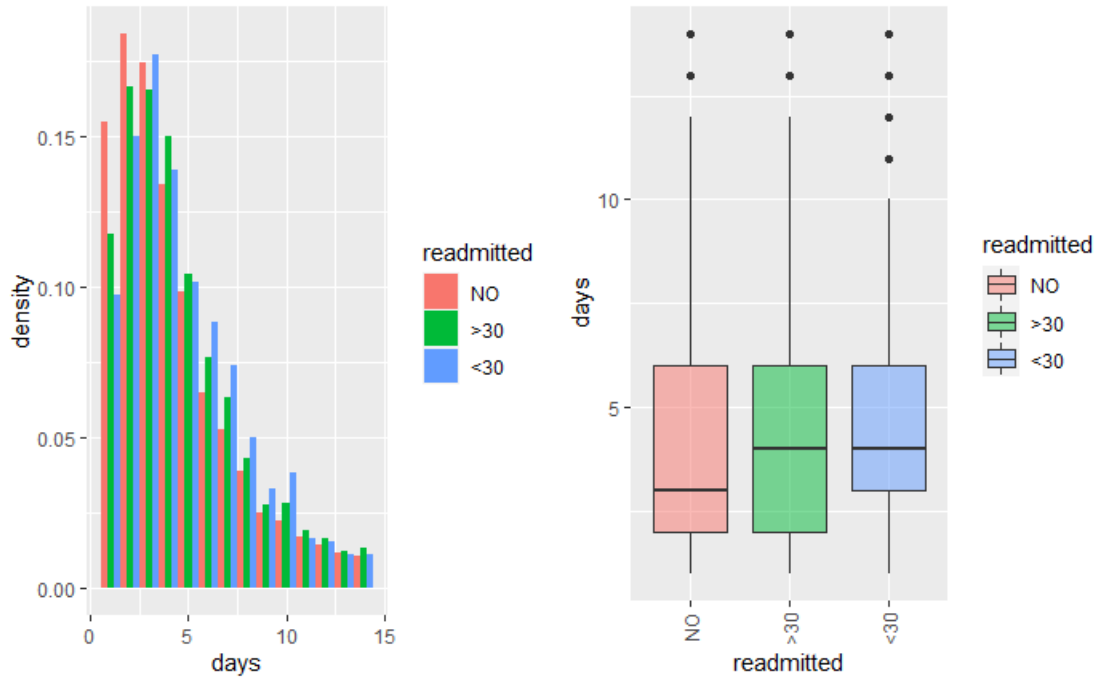
Patients with a history of readmissions tend to have longer stays. Patients admitted for the first time had the shortest length of stay at 4.2 days, whereas patients who have had another admission during the previous 30 days had a mean of 4.77. As there are over 1,000 patients in the latter group, it is clear that this difference is not random or caused by chance.

	mean	median	n
readmitted			
NO	4.2	3	5370
>30	4.6	4	3525
<30	4.8	4	1102

3 rows

The graphs below show that patients who have had previous hospital stays within the prior 30 days stay longer in their current admission. In the box plot on the right, median days are indicated by the center line in each box. The blue box (readmitted < 30) is shows the longest stays and the red box (no readmissions) shows the shortest.

The histogram on the left demonstrates that patients who have been readmitted spend more days in the hospital because the values indicated by the blue lines are always higher than those shown by the orange lines.



Task 3 – Consider two data issues (4 points)

The variable of race has potential ethical implications because racial discrimination has been a problem in hospitals.

The client, MACH, is concerned with the quality of care that patients receive, believing that patients need the best possible treatment to be as healthy as possible. The race variable may help achieve this goal by raising awareness of possible discriminatory practices and including it in the model could help to identify unfair actions. However, there may problems with using this factor. For instance, race data is not audited, and there might be an unequal distribution of missing values, or values might have been recorded incorrectly or unfairly. Prejudices reflected in the collection of this variable would introduce discrimination into the model. However, race data might be useful from a medical standpoint and including it might enable MAAC to better care for their patients.

Even though the number of lab procedures done adds important information to the model, we do not have this data until after a patient's hospital stay, so its usefulness is limited. Using the number of lab procedures conducted during a patient's current stay would duplicate information from the target variable, the number of days, and make the results meaningless. Contrast this variable with num_procs, which records the number of procedures done during the preceding 12 months. Obtaining data about these lab procedures would be useful.

Task 4 – Write a data summary for your actuarial manager (6 points)

This analysis uses historical records from 10,000 diabetic patients who have been readmitted to the hospital. My exploratory study is intended to resolve issues with the data and to help MAAC to improve patient health. This is tracked based on the data shown below, the number of days which patients spend in the hospital after being admitted. This length of stay ranges between zero and 15 days, with an average of about four days.

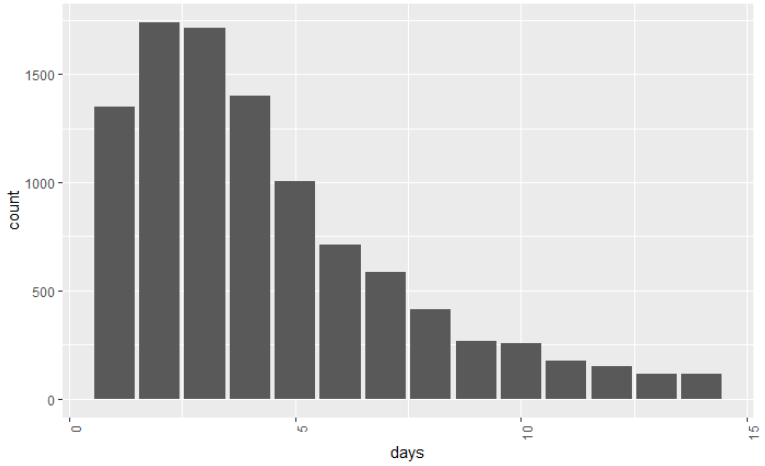


Figure 1 Distribution of Number of Days in Hospital

Age	Average Number of Days Spent in Hospital
[0-10)	3.2
[10-20)	3.1
[20-30)	3.5
[30-40)	3.8
[40-50)	4.0
[50-60)	4.1
[60-70)	4.4
[80-90)	4.8
[70-80)	4.7
[90-100)	4.7

I found three key drivers (or "leading indicators") for patient length of stay. This information may help hospital staff, enabling them to take proactive measures for current inpatients to help them to recover more quickly and return to normal life sooner. These indicators include three important factors: patient age, the number of procedures done during the prior 12 months, and any history of readmission.

Patients who are older generally need to stay longer in a hospital. The table below shows a positive correlation between patient age and the number of days spent in a hospital. Patients who are older take longer to heal after having a medical procedure and so extra care should be given to elderly patients. It is expected that younger patients can often recover on their own at home. Therefore, they are released from the hospital sooner, freeing up time and resources to care for other patients.

Patients who had prior admissions during the preceding 30 days had a longer average length of stay (4.7 days) than patients who did not (4.2 days). The former patients may have underlying issues or more serious cases of diabetes. Therefore, by considering their medical history, staff can be extra careful when helping these patients to recover.

Problems with the provided data included incorrect and blank values, and these were manually corrected using my best judgement. Records were fixed when possible but were omitted if they Could not be repaired.

Two important issues which I considered were whether there are ethical issues with recording a patient's race and whether we can determine the number of laboratory procedures patients undergo. I concluded that considering race variables would be unethical if the data were collected in a discriminatory fashion, but that this knowledge could help doctors to provide better care. For this analysis we chose to include it. However, I advise against including the number of lab procedures because MAAC would not be able to acquire this information in advance of a patient's hospital stay.

Task 5 – Perform a principal components analysis (8 points)

Principal component analysis (PCA) is a dimensionality reduction method which attempts to maintain information in data while using fewer variables. It breaks down linearly related (or correlated) variables into principal components, which are linear combinations of the original uncorrelated variables. It allows us to use only a subset of the PCs based on the percentage of variation explained by each. First, scaling is applied which subtracts the mean and divides by the standard deviation. This helps to ensure that each variable is given the same amount of weight. Otherwise, variables which have the highest numeric value would have too much influence. Then, each variable is rotated, or multiplied by a scalar, known as the loading. We can use this info to create a "recipe" for each PC, which helps us to interpret it.

Advantages

- By reducing the number of variables needed, we can use only two or three of the principal components (instead of using num_procs, num_meds, num_ip, and num_diags) while still capturing most patient hospital stay patterns.
- PCA can help to us to identify groups of patients who have similar characteristics. For instance, num_procs and num_meds may be correlated. Using a single PC would show that patients who have undergone a high number of procedures are also taking a large number of medications. We could find this information by reviewing the PC'S loading factors.

Disadvantages

- Using a principal component will be less interpretable than using the original variables. As the goal is to get insight into the length of patient stays, this is a big disadvantage.

The results of the PCA are shown below. Each PC explains a percentage of the total variation. If these variables were independent with a correlation of 0, then the cumulative proportion would be 25% for each PC. The result showing that the first two PCs only explain 65% of the variation, as compared to the 50% explained if they were independent, means that this PCA analysis is not able to simplify the data very well.

```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.2267 1.0426 0.9141 0.7568
Proportion of Variance 0.3762 0.2717 0.2089 0.1432
Cumulative Proportion  0.3762 0.6479 0.8568 1.0000
```

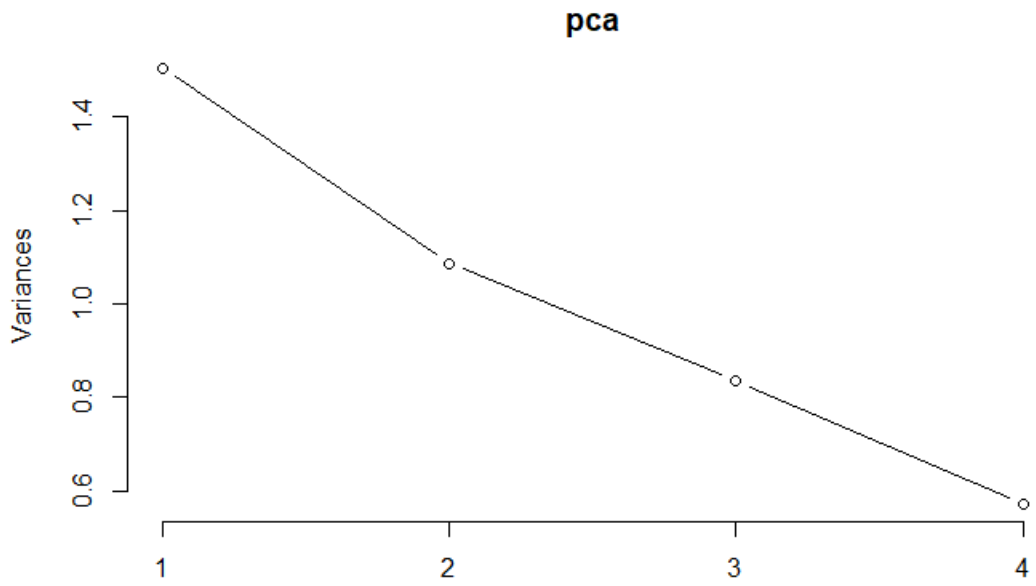
Each PC is a linear combination of the original variables. A recipe for creating the first PC is 0.56*num_procs + 0.68*num_meds + 0.12*num_ip + 0.47*num_diags. This tells us that the largest group of patients have had many diagnoses, medications, inpatient visits, and procedures. We know that this is the largest group because it is the first principal component. You could think of this as a proxy for the patient's health status; those who are healthy have a low score while sicker patients have a higher score.

The second PC has a positive sign on the number of inpatient visits and diagnoses but a negative sign on the number of procedures and medications. This group includes patients who might have had an accident or injury which resulted in a hospital visit to treat their diabetes, but who did not require medications or procedures. We know that this group includes fewer patients because it is the second PC and only explains 27% of the variation.

The third and fourth PC accounts for only 34% of the variation. They represent patients who have had procedures, medications, and inpatient visits, but few diagnoses, and patients with procedures, inpatient visits, and diagnoses, but few medications.

	PC1	PC2	PC3	PC4
num_procs	0.56	-0.45	0.37	0.60
num_meds	0.67	-0.05	0.10	-0.73
num_ip	0.12	0.80	0.58	0.13
num_diags	0.47	0.41	-0.72	0.31

MAAC is more concerned with inference than prediction and so I recommend using the original variables. The first PC explains only 38% of the total variation, not sufficient for these purposes. If using PCA, I would recommend including at least the first two PCs.



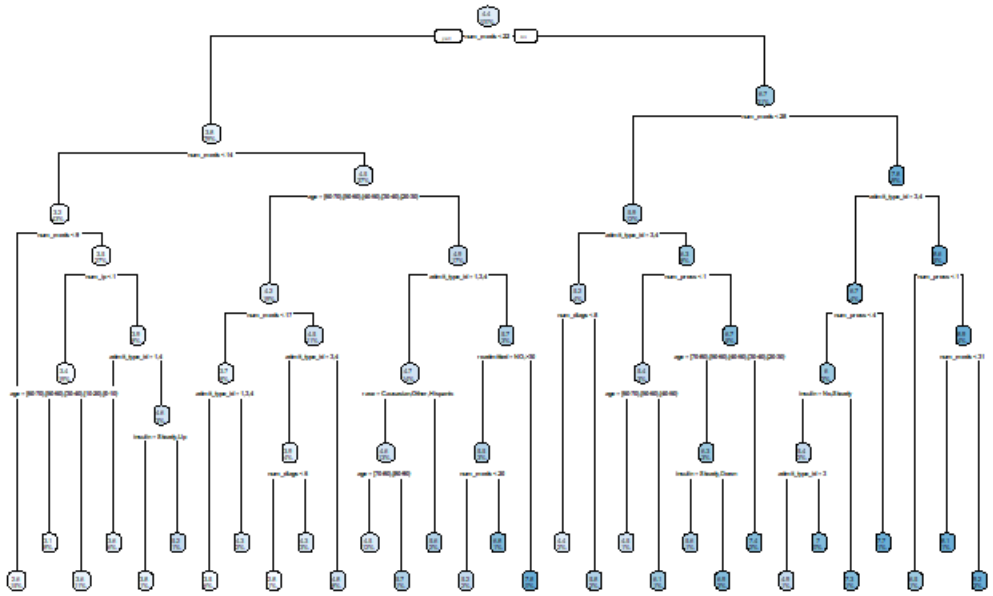
I split the data into training and test sets to expedite evaluation of new test data results and to determine real life implications. I compared the mean number of days in the training and test sets and found that they are roughly equal to the overall data set (about 4.4).

```
[1] "TRAIN"
[1] 4.397857
[1] "TEST"
[1] 4.435769
[1] "ALL"
[1] 4.409223
```

Task 6 – Construct a decision tree (10 points)

Pruning, also known as cost-complexity pruning, is based on a complexity parameter (cp) and creates a nested sequence of subtrees of a decision tree by recursively snipping off the least important splits. First, a decision tree with many leaves is created. It is then simplified using a pruning algorithm. The amount by which a model improves on each split, measured by the number of days on the right and left branches of the node, is taken into consideration. This is appropriate for the current business purpose of predicting a length of stay because MAAC is interested in insights they can share with hospital managers. Having a simpler tree will be more useful for healthcare providers.

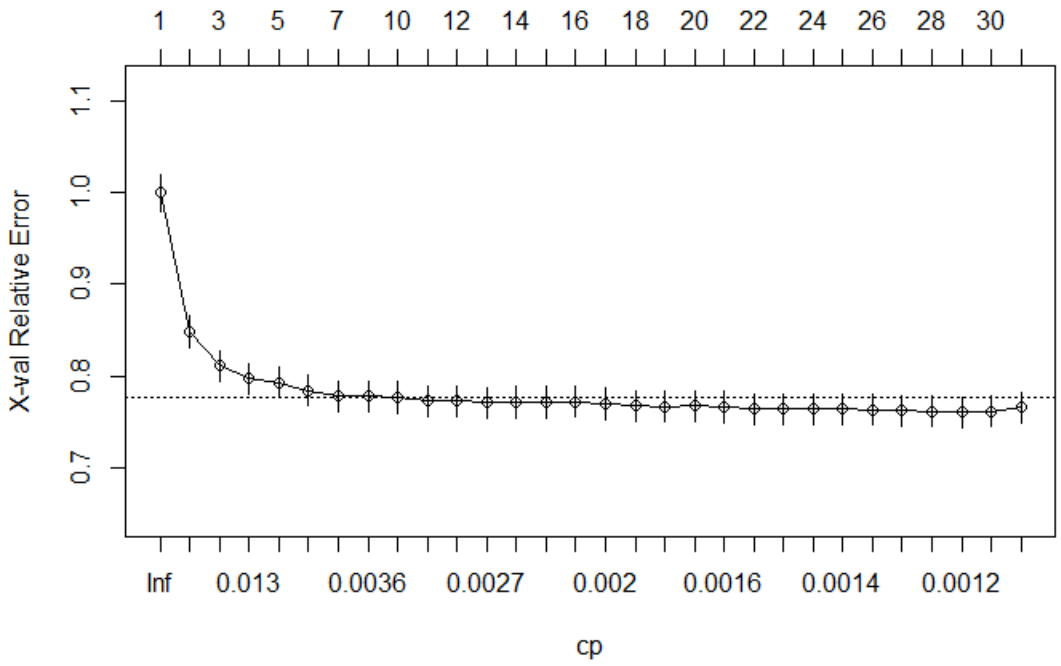
The unpruned tree is shown below.



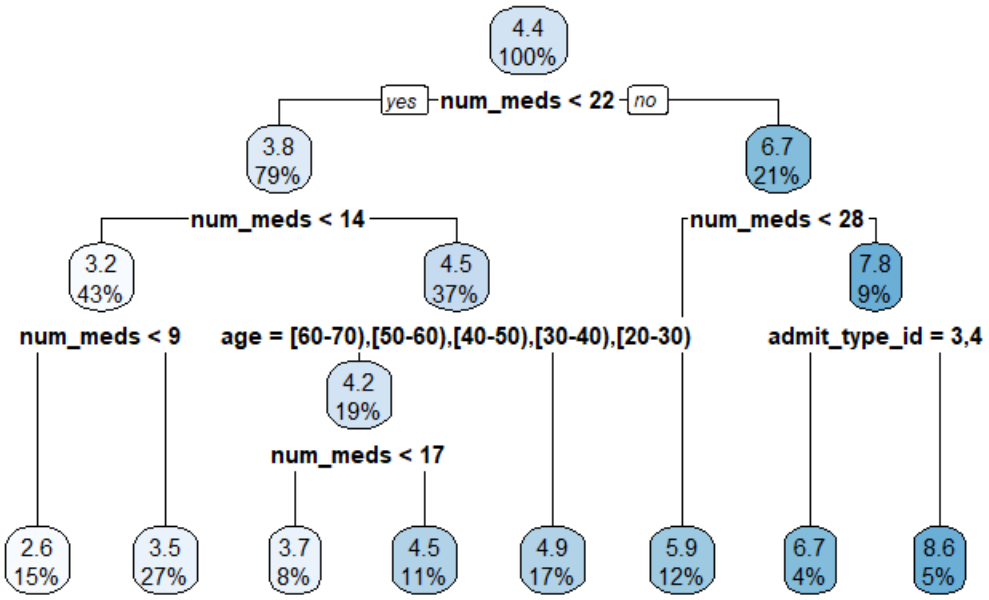
I use cross-validation to fit trees with different values of cp and then choose the one which has the lowest error. However, the lowest error found in this case is for a tree that is too complicated for this problem because it has 28 nodes, for a CP value of **0.001223032.**

```
         CP nsplit rel error    xerror       xstd
1  0.152707521      0 1.0000000 1.0001587 0.02033349
```

2	0.037698542	1	0.8472925	0.8480251	0.01798971
3	0.019130886	2	0.8095939	0.8107041	0.01729121
4	0.009240202	3	0.7904631	0.7966838	0.01725127
5	0.008676511	4	0.7812228	0.7926694	0.01717010
6	0.005343208	5	0.7725463	0.7839649	0.01699532
7	0.003715982	6	0.7672031	0.7777088	0.01689852
8	0.003568539	7	0.7634871	0.7773847	0.01695395
9	0.003008571	9	0.7563501	0.7762636	0.01695438
10	0.002949588	10	0.7533415	0.7720921	0.01690122
11	0.002883700	11	0.7503919	0.7720237	0.01692143
12	0.002491909	12	0.7475082	0.7707511	0.01692526
13	0.002315013	13	0.7450163	0.7713557	0.01695682
14	0.002125084	14	0.7427013	0.7712409	0.01695252
15	0.001972886	15	0.7405762	0.7716506	0.01695304
16	0.001965305	16	0.7386033	0.7692351	0.01690420
17	0.001753610	17	0.7366380	0.7670100	0.01681699
18	0.001692801	18	0.7348844	0.7666489	0.01679296
19	0.001626374	19	0.7331916	0.7673546	0.01680379
20	0.001603454	20	0.7315652	0.7657792	0.01677920
21	0.001468164	21	0.7299618	0.7633766	0.01671210
22	0.001457705	22	0.7284936	0.7633931	0.01671218
23	0.001455661	23	0.7270359	0.7633931	0.01671218
24	0.001407144	24	0.7255802	0.7634303	0.01672284
25	0.001294707	25	0.7241731	0.7629509	0.01670685
26	0.001273377	26	0.7228784	0.7616959	0.01672131
27	0.001242786	27	0.7216050	0.7611078	0.01670085
28	**0.001223032**	**28**	**0.7203622**	**0.7603318**	**0.01669433**
29	0.001002113	29	0.7191392	0.7609538	0.01674067
30	0.001000000	30	0.7181371	0.7651169	0.01687691



Therefore, instead of using this minimum CP value, I chose to use 0.003568539 because this result is from a tree with only 7 nodes.



To interpret the above tree, start at the top. Move left if a patient meets the given criteria, otherwise move right. After proceeding through the choices, a predicted number of days is arrived at. The predicted number of days spent in a hospital are:

- 2.6 days when num_meds < 9
- 3.5 days when 9 >= num_meds < 14
- 3.7 days when 14 <= num_meds < 17, and age is between 20 and 70
- 4.5 days when 17 >= num_meds < 22, and age is between 20 and 70
- 4.9 days when num_meds >= 14, and age is not between 20 and 70
- 5.9 days when num_meds between 23 and 27
- 6.7 days when num_meds > 28 and admit_type_id is either Elective or Not Available
- 8.6 days when num_meds > 28 and admit_type_id is either Urgent or Emergency

```
n= 7000

node), split, n, deviance, yval
      * denotes terminal node

 1) root 7000 62304.970 4.397857
   2) num_meds< 21.5 5559 36673.060 3.804281
     4) num_meds< 13.5 2985 14840.800 3.200670
       8) num_meds< 8.5 1069  3351.667 2.612722 *
       9) num_meds>=8.5 1916 10913.420 3.528706 *
     5) num_meds>=13.5 2574 19483.450 4.504274
      10) age=[60-70),[50-60),[40-50),[30-40),[20-30) 1351  9365.500 4.162102
        20) num_meds< 16.5 568  3306.394 3.676056 *
        21) num_meds>=16.5 783  5827.581 4.514687 *
      11) age=[70-80),[80-90),[90-100),[10-20),[0-10) 1223  9785.045 4.882257
*
   3) num_meds>=21.5 1441 16117.470 6.687717
     6) num_meds< 27.5 843  8468.833 5.921708 *
```

```
 7) num_meds>=27.5 598  6456.691 7.767559
  14) admit_type_id=3,4 262  2721.958 6.690840 *
  15) admit_type_id=1,2 336  3194.143 8.607143 *
```

A Pearson Goodness of Fit statistic has been used to evaluate the model. This assumes that the squared error divided by the predicted value has a chi-squared distribution. Higher values are less desired because the residuals are larger; conversely, a lower value is better because it means that the residuals are smaller. I have not used a metric such as R^2 or RMSE or MAE because we are modeling a counting value, the number of days spent in a hospital. This gives the data specific properties. For instance, the likelihood that a person spends an additional day in a hospital decreases as the number of days increases. RMSE or MAE does not take this into consideration.

The Pearson Goodness of Fit statistics for the training and test sets are below. The results are higher (worse) on the test sets than on the training sets, as is always the case. After pruning, the test stats get only slightly worse, from 1.56 to 1.60. This is a good value considering that the tree has 7 nodes instead of 28.

Before pruning:

```
Train 1.452799
Test 1.55681
```

After pruning:

```
Train 1.539727
Test 1.596192
```

Task 7 – Construct a generalized linear model (7 points)

The target variable of the number of days is positive and right skewed. It only takes on discrete values in the data although it is technically possible that a patient could stay for a fractional number of days. Any response distribution which matches these criteria is possible. The binomial is not useful because the data requires more than two values. The Poisson is the best choice because its counting is variable. The Gamma is also possible because it is right skewed and positive. In the context of the current business problem, it is important that the predicted results show positive numbers because hospital staff would not be able to understand a negative number of days spent in a hospital.

The Goodness of fit stats are calculated below. A GLM which has a principal component instead of the original variables resulted in higher (worse) test statistics. The principal component is also more difficult to interpret. This implies that using the first model is better for both performance and for ease of interpretation.

GLM using original variables instead of principal component:

```
Train 1.525184
Test 1.558293
```

GLM using the principal component instead of the original variables:

```
Train 1.587999
Test 1.618514
```

Task 8 – Perform feature selection with lasso regression (4 points)

The features used in the lasso are

```
genderMale
age[50-60)
age[80-90)
age[90-100)
raceAfricanAmerican
admit_type_id2
admit_type_id3
readmitted<30
num_meds
num_ip
num_diags
```

The results from the GLM in Task 7 are below. This has a lot of variables but results in a lower (better) value of the Pearson goodness of fit stat (1.56) when compared to the Lasso, which has a value of 1.57. However, MAAC is more concerned with ease of interpretation than with performance. The second model, the Lasso, removes several variables, making it easier to explain to medical personnel. I recommend that this model be used. One aspect of the Lasso which may be difficult to explain is the fact that only certain age ranges (e.g. 50-60, 80-90, or 90-100) are considered instead of having a different coefficient for each age.

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.7179703  0.0316229  22.704  < 2e-16 ***
genderMale         -0.0348400  0.0116236  -2.997 0.002723 **
age[60-70)         -0.0805106  0.0165719  -4.858 1.18e-06 ***
age[50-60)         -0.1342346  0.0181664  -7.389 1.48e-13 ***
age[80-90)          0.0648094  0.0174044   3.724 0.000196 ***
age[40-50)         -0.1042567  0.0225005  -4.634 3.59e-06 ***
age[30-40)         -0.0929152  0.0342933  -2.709 0.006740 **
age[90-100)         0.1144602  0.0347736   3.292 0.000996 ***
age[20-30)         -0.1015377  0.0580091  -1.750 0.080054 .
age[10-20)          0.0457339  0.0747208   0.612 0.540495
age[0-10)           0.0302069  0.1679155   0.180 0.857236
raceAfricanAmerican 0.1156573  0.0150857   7.667 1.77e-14 ***
raceOther           0.0509221  0.0291368   1.748 0.080518 .
raceHispanic        0.0876614  0.0421711   2.079 0.037644 *
admit_type_id2      0.1254435  0.0152688   8.216  < 2e-16 ***
admit_type_id3     -0.0858340  0.0163343  -5.255 1.48e-07 ***
admit_type_id4     -0.0230973  0.0200092  -1.154 0.248363
metforminSteady    -0.0156711  0.0152461  -1.028 0.304007
metforminUp         0.1526204  0.0472482   3.230 0.001237 **
metforminDown       0.0674198  0.0716793   0.941 0.346922
insulinSteady      -0.0222760  0.0137654  -1.618 0.105608
insulinDown        -0.0109830  0.0186137  -0.590 0.555157
insulinUp           0.0237683  0.0188068   1.264 0.206296
readmitted>30       0.0423494  0.0125898   3.364 0.000769 ***
readmitted<30       0.0793006  0.0189584   4.183 2.88e-05 ***
num_procs           0.0112080  0.0036594   3.063 0.002193 **
num_meds            0.0308537  0.0007074  43.618  < 2e-16 ***
num_ip              0.0140475  0.0044002   3.192 0.001411 **
num_diags           0.0273086  0.0035108   7.779 7.34e-15 ***
```

GLM from Task 7:

```
Train 1.525184
Test 1.558293
```

```
(Intercept)          .
genderMale          -0.0005247084
age[60-70)           .
age[50-60)          -0.0318520898
age[80-90)           0.0609863438
age[40-50)           .
age[30-40)           .
age[90-100)          0.0271444979
age[20-30)           .
age[10-20)           .
age[0-10)            .
raceAfricanAmerican  0.0408150599
raceOther            .
raceHispanic         .
admit_type_id2       0.0879293247
admit_type_id3      -0.0365864924
admit_type_id4       .
metforminSteady      .
metforminUp          .
metforminDown        .
insulinSteady        .
insulinDown          .
insulinUp            .
readmitted>30        .
readmitted<30        0.0014827045
num_procs            .
num_meds             0.0296963221
num_ip               0.0025137279
num_diags            0.0223963381
```

Lasso from Task 8

```
Train 1.541731
Test 1.572154
```

Task 9 – Discuss the bias-variance tradeoff (7 points)

Bias is the difference between expected values of the model and expected values of the target. It can be thought of as the "difference from the center of the target".

Variance is the amount by which the predicted values change when the input data changes. This is just the statistical variance of the predicted values.

The bias-variance tradeoff states that root mean squared errors can be divided into three parts:

1) The bias squared – models which have high bias (underfitting) tend to have low variance;
2) The variance – models which have high variance (overfitting) often have low bias; and,
3) Irreducible error – random noise which no model can completely remove.

Lasso controls model bias and variance using a lambda parameter. This imposes a penalty on log likelihood so coefficients are either removed or changed in size. Models which have fewer variables have a lower variance but higher bias because they are less flexible; conversely, models which have low bias but higher variance have more variables. By looking at many values of lambda and choosing the one which has the lowest error, the GLM can be optimized.

Without splitting the training and test sets, we would not be able to accurately measure bias because although a model would fit the data that it was trained on well, it would have far worse results if new data is used. By splitting into training and test sets, model error (Pearson goodness of fit statistic in this case) can be accurately estimated. It would then be possible to decide on a model which accurately reflects reality.

Task 10 – Consider the final model (4 points)

One advantage to using a GLM instead of a tree for this problem is that the predictions will change gradually as the input variables change. This is possible because a GLM uses coefficients for each variable instead of yes/no questions in a decision tree that could result in difficult to explain stepwise predictions. Hospital staff may be more comfortable seeing a patient's projected length of stay change gradually rather than suddenly.

One disadvantage is that decision trees automatically handle missing values, while they had to be manually fixed in our data. The GLM does not handle missing values automatically and so requires additional time. MAAC should consider the cost of paying an actuary for the time required to produce this model when making its decision.

One advantage to using a Lasso is that it is easier to interpret because it removes variables using penalty terms by setting coefficients to zero. This can make the results easier for doctors to understand.

One disadvantage to using the Lasso in R is that the log link function is not supported and therefore only the identity link function can be used. The actuaries who would be hired to train this model would need to take this into consideration.

Task 11 – Interpret the model for the client (7 points)

I reran the GLM from Task 7 on the entire data set.

This model can be interpreted using a simple formula. We start with the most common patient, predicting the length of hospital stay (number of days). Then, depending on patient characteristics, this number may either increase or decrease.

The most common patient is female, age 70-80, Caucasian, has no metformin, insulin, or history of being readmitted, and has been admitted for an emergency. This patient has a predicted length of stay of 2.2 days. Then, for men, we decrease this prediction by multiplying by 0.97. The other pieces of information about the patient produce similar changes. The coefficients below which have a negative sign on the estimate decrease this prediction and those with a positive sign increase it. Then we multiply by 1.012 to the nth power, where n is the number of procedures which a patient has had in the past 12 months. There are similar relationships for the number of procedures, number of inpatient visits, and number of diagnoses.

```
Call:
glm(formula = days ~ . - PC1, family = poisson(link = "log"),
    data = data.all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8113  -0.9898  -0.2784   0.5766   4.8714

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.7978752  0.1383465   5.767 8.06e-09 ***
genderMale         -0.0280296  0.0103461  -2.709 0.006744 **
age[10-20)         -0.0983412  0.1542494  -0.638 0.523769
age[20-30)         -0.2302232  0.1454871  -1.582 0.113551
age[30-40)         -0.2458890  0.1407981  -1.746 0.080743 .
age[40-50)         -0.2329239  0.1390500  -1.675 0.093913 .
age[50-60)         -0.2670869  0.1386212  -1.927 0.054012 .
age[60-70)         -0.2123259  0.1384994  -1.533 0.125265
```

```
age[70-80)          -0.1467226  0.1384676  -1.060 0.289319
age[80-90)          -0.0794199  0.1385948  -0.573 0.566620
age[90-100)         -0.0146799  0.1410411  -0.104 0.917104
raceAfricanAmerican  0.0901276  0.0130461   6.908 4.90e-12 ***
raceHispanic         0.0710824  0.0387262   1.836 0.066430 .
raceOther            0.0283174  0.0449429   0.630 0.528646
raceAsian            0.0877105  0.0637325   1.376 0.168751
admit_type_id2       0.1034508  0.0128909   8.025 1.01e-15 ***
admit_type_id3      -0.0962379  0.0139251  -6.911 4.81e-12 ***
metforminSteady     -0.0099422  0.0135395  -0.734 0.462760
metforminUp          0.1625055  0.0432158   3.760 0.000170 ***
metforminDown        0.1319025  0.0651327   2.025 0.042854 *
insulinSteady       -0.0280576  0.0123218  -2.277 0.022782 *
insulinDown         -0.0065038  0.0163592  -0.398 0.690952
insulinUp            0.0266025  0.0165736   1.605 0.108469
readmitted>30        0.0336618  0.0112498   2.992 0.002770 **
readmitted<30        0.0636384  0.0166012   3.833 0.000126 ***
num_procs            0.0120259  0.0032266   3.727 0.000194 ***
num_meds             0.0310800  0.0006329  49.109  < 2e-16 ***
num_ip               0.0134967  0.0039234   3.440 0.000582 ***
num_diags            0.0361976  0.0031564  11.468  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16548  on 8770  degrees of freedom
Residual deviance: 12335  on 8742  degrees of freedom
AIC: 40052

Number of Fisher Scoring iterations: 5
```

Task 12 – Executive summary (20 points)

Our client, Merged and Acquired Clinics and Hospitals (MACH), has hired us to help their hospital executives gain a better understanding of the factors that drive inpatient lengths of stay. We used predictive analytics to identify the reasons why some patients are sent back home quickly, while others need to spend several days in the hospital. We used historical data about patients with diabetes admitted to U.S. hospitals between 1999 and 2008. Any findings from this report are limited to the population of this specific data set and may change if applied to a different population.

Each encounter includes the length of hospital stay measured in days as well as gender, age, race, weight, reason for admission (emergency, elective, or non-elective), and any changes to metformin or insulin prescriptions. It also includes their medical history for the preceding 12 months such as the number of readmissions within the prior 30 days, number of procedures performed, medications prescribed, inpatient visits, and diagnoses.

The following formula can be used by your healthcare staff to predict how long a patient will probably spend in the hospital. It uses a simple spreadsheet and adjusts the predicted length of stay based on patient characteristics,. It can be used to predict which patients will likely have a long stay, giving staff a chance to intervene, or can just be used for informational purposes. The calculations listed below show the different factors that may either increase or decrease a length of stay. The initial prediction of a 2.2 day admission represents the most common patient. This estimate is modified based on specific characteristics. For example, if the patient is male, the stay is decreased by multiplying 2.2 by 0.97. For a patient between the ages of 10 and 20, you can decrease the estimate by multiplying by 0.91. Each of the following variables is handled and interpreted in the same way.

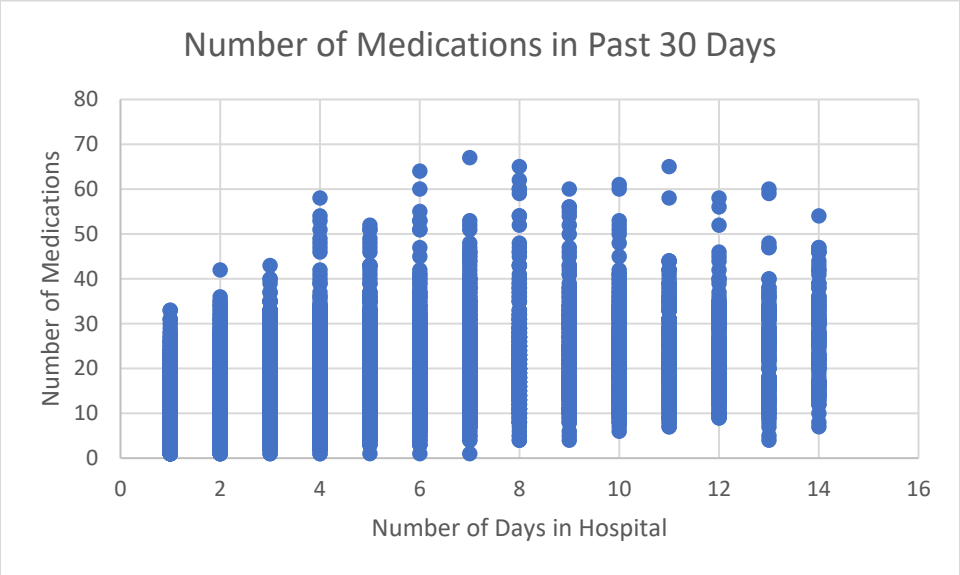
Interpretation
Start with a prediction of 2.2 days
If the patient is Male, multiply by 0.97
If the patient is age is between 10 and 20, multiply by 0.91
If the patient is age is between 20 and 30, multiply by 0.79
If the patient is age is between 30 and 40, multiply by 0.78
If the patient is age is between 40 and 50, multiply by 0.79
If the patient is age is between 5 and 60, multiply by 0.77
If the patient is age is between 60 and 70, multiply by 0.81
If the patient is age is between 70 and 80, multiply by 0.86
If the patient is age is between 80 and 90, multiply by 0.92
If the patient is age is between 90 and 100, multiply by 0.99
If the patient's race is African American, multiply by 1.09
If the patient's race is Hispanic, multiply by 1.07
If the patient's race is Other, multiply by 1.03
If the patient's race is Asian, multiply by 1.09
If the patient's readmission was Urgent, multiply by 1.11
If the patient's readmission was Elective, multiply by 0.91
If the patient's Metformin was Steady, multiply by 0.99
If the patient's Metformin was Up, multiply by 1.18
If the patient's Metformin was Down, multiply by 1.14
If the patient's Insulin was Steady, multiply by 0.97
If the patient's Insulin was Down, multiply by 0.99
If the patient's Insulin was Up, multiply by 1.03
If the patient had not been readmitted in the last 30 days, multiply by 1.03
If the patient had been readmitted in the last 30 days, multiply by 1.07
For each procedure that the patient has had, multiply by 1.01
For each medication that the patient has had, multiply by 1.03
For each inpatient visit that the patient has had, multiply by 1.01
For each diagnosis that the patient has had, multiply by 1.04

To ensure optimal quality, we performed integrity checks of the data prior to beginning the analysis. This involved correcting errors in the data and removing incomplete values. There were a few patients who had missing gender records, and these were removed. The patient's race was not included in 226 cases, but due to the large number of records we included them as a separate group. We were supplied with the patient's weight but did not use this factor because it was missing from most patient records. This is a serious problem with the data which MAAC should examine.

We used visual and statistical methods to look for patterns. We found that most patients who are readmitted spend about four days in the hospital. We looked at other patient information to determine if there were clear differences between the types of patient who have long vs. short hospital stays.

We found that patients who have had another hospital stay in the recent past (within the last 30 days) had a longer stay on average. There were 1,102 patients in this group, a significant number. It may be worth considering if these patients have something different about them, such as other chronic illnesses and including this information within future analyses. Perhaps your medical team would have additional helpful insight to this issue.

	Average Days in Hospital	Number of Patients
No readmission history	4.2	5,370
History of readmission	4.6	3,525
Recent readmission	4.8	1,102



There are ethical concerns with using information about patient race in the model. It may be the case that this information helps healthcare providers prevent discrimination, or it may create a discriminatory bias in the model, giving people different levels of care depending on their race. There may also be risk of potential lawsuits.

We conclude that the other information was complete and affected length of stay. We recommend against using the number of laboratory procedures, which my assistant mentioned, because this information cannot be collected in advance of admission.

We sent our actuarial manager a summary of all data steps. You can review this information with them to clarify any additional steps needed. This summary explains how we cleaned the data and would be useful if you wish to repeat this analysis. We found that three factors helped predict a patient's length of stay, including age, the number of procedures undergone in the prior 12 months, and their admission history.

One way to determine which factors are related to the length of a hospital stay is by looking at the correlations. When two things are correlated, it means that they increase and/or decrease together. When the number of medications given to a patient increases, their length of stay did as well. It is important to remember that correlation does not imply causation. It could be, for instance, that patients who have more medications have other underlying health issues which cause an increased length of stay. We recommend using the results from our predictive model, which includes these other factors and adjusts for them.

You can simplify information collected about patient history down to a single number. This would give you the option of using a single score to reflect a patient's history instead of using separate columns for each of these values. The advantage would be that the set of rules would be simpler to implement but it may be more difficult for your hospital staff to understand. The recipe for this is:

(0.56)(Number of Procedures) + (0.67)(Number of Medications) + (0.12)(Number of Inpatient Visits) + (0.47)(Number of Diagnosis)

We investigated several alternate model approaches to ensure that our results are as reliable as possible. We experimented with several methods and chose the one which most closely matched actual patient readmission patterns. We also looked at decision trees and penalized regression models. In other words, we are only showing you the best results from our modeling.

You can be confident that these results will work in real life because they have been tested using a scientific approach known as training-testing validation. This used 70% of the patients as a training set and the remaining 30% were held out as a blind test set. We evaluated each of the models based on this test set and ended up selecting a generalized linear model (GLM) because it had the best result.

We included interpretations of this model into the Table of Rules discussed earlier.

We considered that MAAC may be interested in building a model with better predictive power. Considerations should include the tradeoff between interpretability, how explainable the results are to your hospital staff, and predictive power. We chose a model which is easy to interpret; however, you could consider more powerful models as well. There are advantages and disadvantages to using a different model. For example, some of the data cleaning work that we needed to do could be automated if a tree-based model is used.

You should consider several factors before your next steps. You may wish to follow up with your hospital staff regarding these issues:

- How may these results change for non-diabetic patients?
- What additional data could be collected on these patients? Would the type of prescription, for example, be useful to know so that doctors could see if certain medications are causing longer hospital stays?
- Could more recent data be collected? This study is based on data from 1999-2008, but medical records have changed in the last 12 years and so these findings may be out of date.

In conclusion, we identified the factors which determine how long a patient will spend in the hospital after being readmitted. We present these results to you so that you can take proactive action in caring for your patients in the best way possible.