Exam PA June 19, 2020 Project Report Template

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

Also be sure all the documents you are working on have June 19 attached.

As indicated in the instructions, work on each task should be presented in the designated section for that task.

> *This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.*

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

> *Most candidates successfully identified and made adjustments to missing and invalid data. To earn full points, candidates had to make appropriate adjustments and provide clear rationale for their decisions.*

- There were 2 unknown/invalid values for the gender variable. Because there were only 2 records out of 10,000 total records, these rows were removed from the data.
- There were 9,714 records with missing values for the weight variable. Because most of the weight data was missing, the variable was removed from the dataset.
- The admin_type_id variable was coded as a numeric variable. Since the numeric values are codes representing categorical data, the variable was changed to a factor variable.
- The race variable contained 222 missing values. We do not know if these are missing because of a data collection error or the race is routinely unknown. Because we do not know whether it is missing at random, we should keep the variable and see whether missing race has predictive power. A new race category was created called “Missing.” After reviewing the relationships, I also combined the “Asian” and “AfricanAmerican” levels because “Asian” had a small number of records and similar mean lab tests performed as “AfricanAmerican.” I also combined the “Hispanic” and “Other” levels because both had small counts and similar mean lab tests.
- The factor variable levels were reordered, so the most frequent level was first.
- The num_meds variable had values ranging from 1 to 68. It seems unlikely that in a large dataset there would be no individuals that took 0 medications in the prior year. This is suspicious and should be investigated because it could be an indicator of invalid data, but the values look reasonable aside from that, so I will use the variable without alterations.

After the changes 9,998 records remained in the dataset.

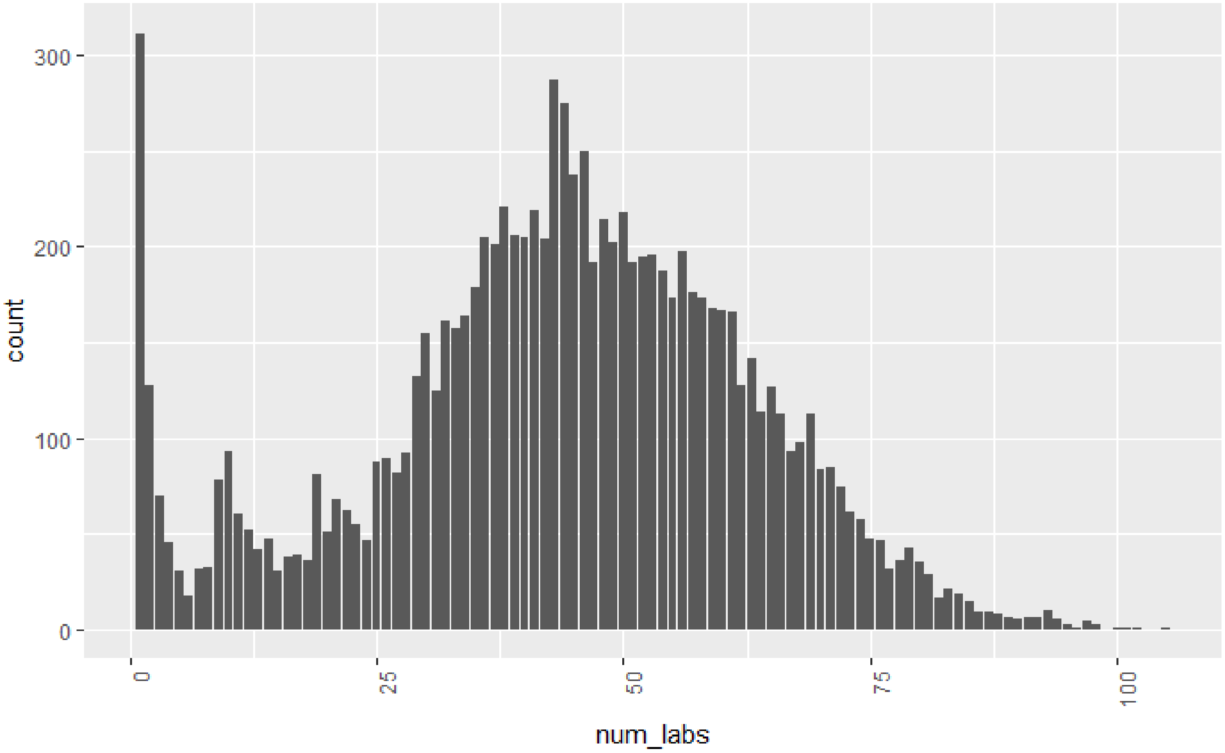
Task 2 – Explore the data (15 points)

> *Candidates were expected to use a combination of summary statistics and visualizations for each variable but limit their tables and charts to those that showed the key relationships discussed in the report. The best candidates made insightful observations relating to the business problem*

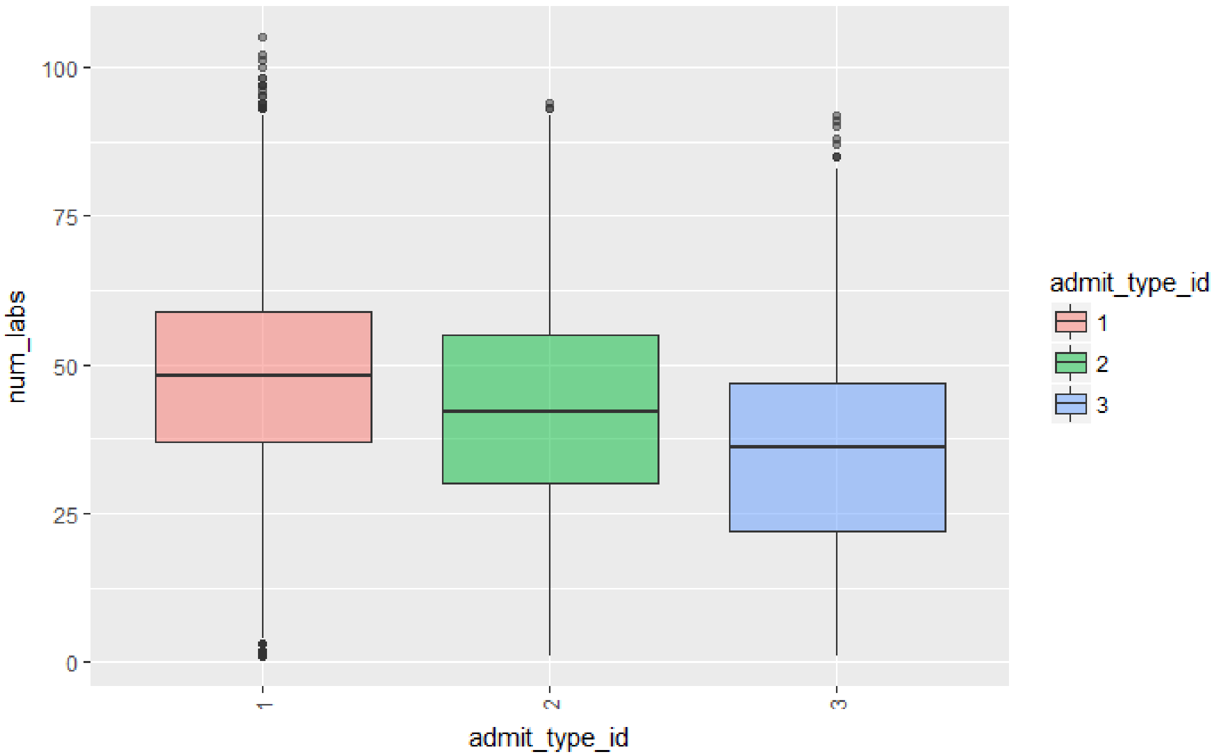
when referring to their summary statistics and visualizations. Many candidates failed to adequately explain their reasons for choosing the three predictors.

The target variable is the number of lab tests administered during an inpatient stay. The variable takes on integer values from 1 to 105. The center of the distribution is around 43 to 45 labs based on the mean/median. From the bar chart below, we can see a large concentration of patients that had only 1 or 2 lab tests. Outside of that, the distribution looks somewhat normal, but with a lighter tail on the right side of the distribution.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	33	45	43.49	57	105

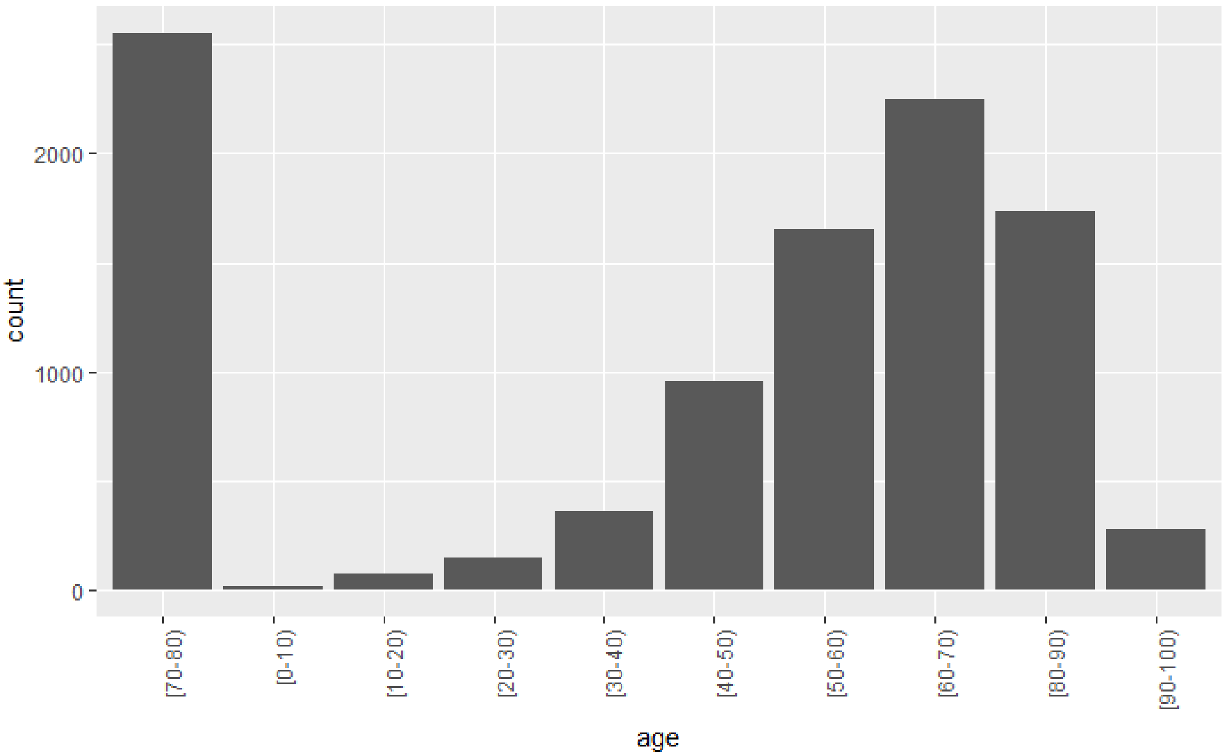


The admit_type_id variable indicates whether the patient was admitted due to 1 - emergency, 2 - urgent care, or 3 - an elective reason. Sizeable differences were observed in the mean number of labs administered and boxplot distributions for each of these groups with emergency admissions requiring the most labs, followed by urgent, and finally elective. I selected this variable because (1) each variable level has over 2000 records and a noticeable difference in the num_labs was seen in the mean days analysis and the boxplots and (2) it makes intuitive sense that a patient being admitted for an emergency or urgent reason may have a more serious condition requiring more tests than a patient being admitted for an elective procedure.

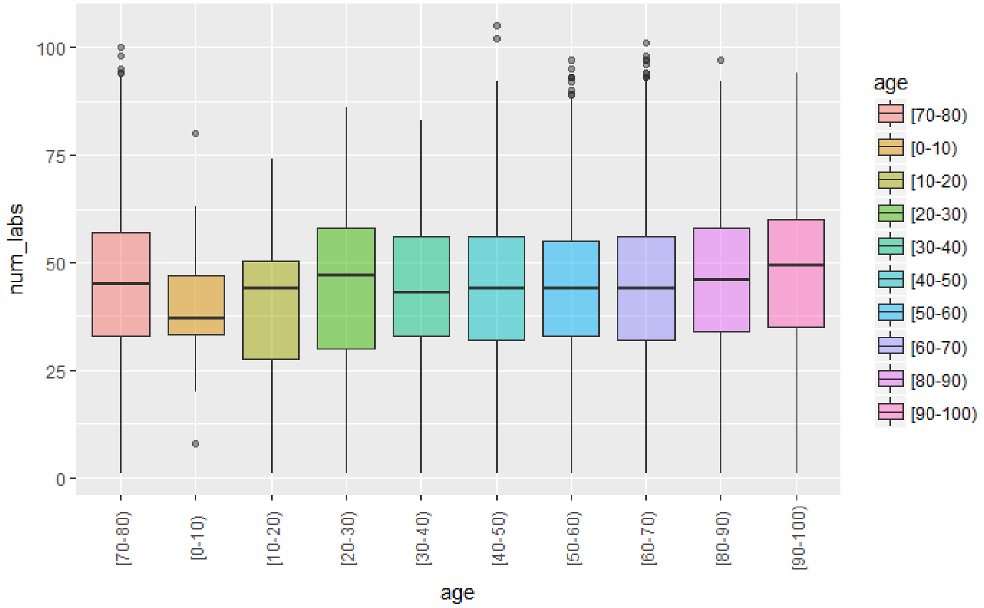


admit_type_id	mean	median	n
1 Emergency	47.19310	48	5945
2 Urgent	41.29324	42	2012
3 Elective	34.85301	36	2041

The age variable is a factor variable, and frequency counts can be seen in the bar chart and table below. Most of our data has ages between 50 and 90, with 70-80 containing the most data. If age were numeric, we would say it was skewed left. I chose this variable because the table and box plots below show that above age 70, higher ages are associated with more lab tests. It makes sense that older patients tend to need more lab tests since they tend to be in poorer health. The distribution of num_labs at ages 0 to 30 appear somewhat erratic, but there are not many records at those ages, so differences in the num_labs there might just be noise.

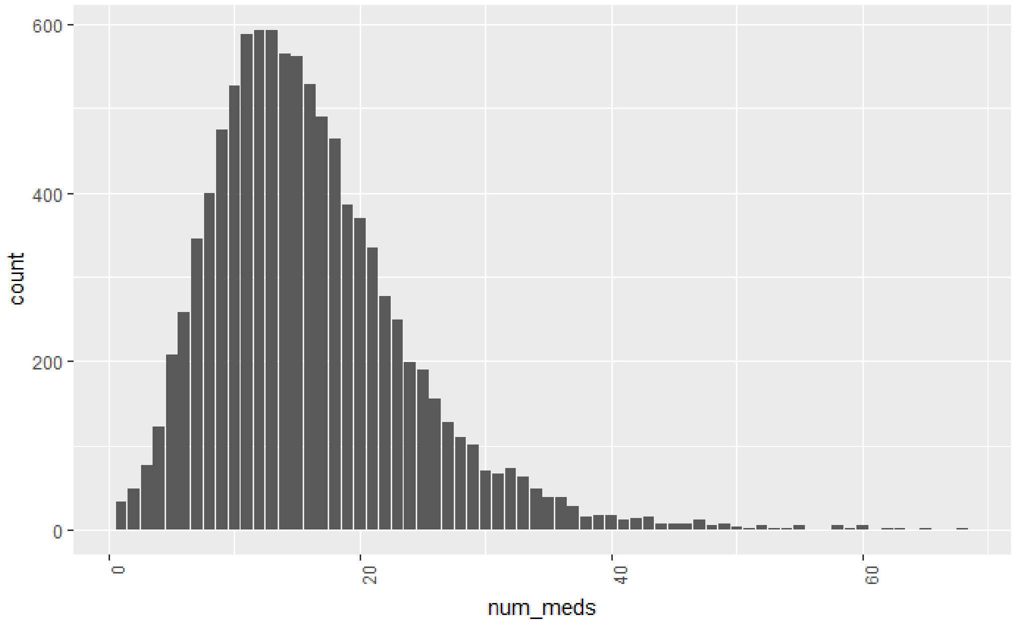


age	mean	median	n
[70-80)	44.11063	45.0	2549
[0-10)	41.00000	37.0	14
[10-20)	39.80282	44.0	71
[20-30)	42.98611	47.0	144
[30-40)	42.50980	43.0	357
[40-50)	42.37251	44.0	953
[50-60)	42.95584	44.0	1653
[60-70)	42.80124	44.0	2249
[80-90)	44.54619	46.0	1732
[90-100)	46.28623	49.5	276



The num_meds variable takes on integer values from 1 to 68. The center of the distribution is around 15 to 16 based on the mean/median. From the bar chart below, we can see that the distribution is skewed right with 13 being the most frequent number of medications. The correlation between num_meds and num_labs was 0.275, which was the strongest correlation among the numeric variables. This suggests that as num_meds increases, num_labs increases. Like the other variables selected, this relationship makes intuitive sense – patients taking many medications likely have more underlying conditions, have poorer overall health, and may require more lab tests to diagnose their conditions and monitor adjustments to their many medicines. I selected this variable because the correlation to the target variable was the strongest of the numeric variables and the narrative made sense.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	11	15	16.08	20	68



Task 3 – Consider two data issues (4 points)

Most candidates successfully identified reasons for considering removing the race variable, but few candidates were able to discuss potential benefits of including the variable. Future candidates should consider the business context more thoroughly when discussing controversial variables. In this case, excluding the race variable could be unethical if it leads to worse care. Lower quality responses mentioned concerns about the race variable due to regulation or privacy laws with little or no explanation.

The race variable presents ethical concerns that the analysts should weigh before using the variable. Historically, racial groups have been mistreated, and efforts to create racial equality continue today.

Hospital administrators intend to use information about important model factors to better manage patient needs. If race turns out to be an important factor, would they apply different treatment plans to different races? This could be seen as discriminatory, but it could also be the ethical choice if it leads to improved care for all races. Failing to take measures to close the gap in the number of lab tests administered between races could also be seen as discriminatory – since it could be seen as disproportionate care or because lab tests are expensive, and one race might generally be charged more than another.

Whether or not the race variable is included in the model, users and other stakeholders should make sure races aren't unfairly impacted by the model as it is applied.

Many candidates were unable to describe precisely the problems associated with including the number of days variable.

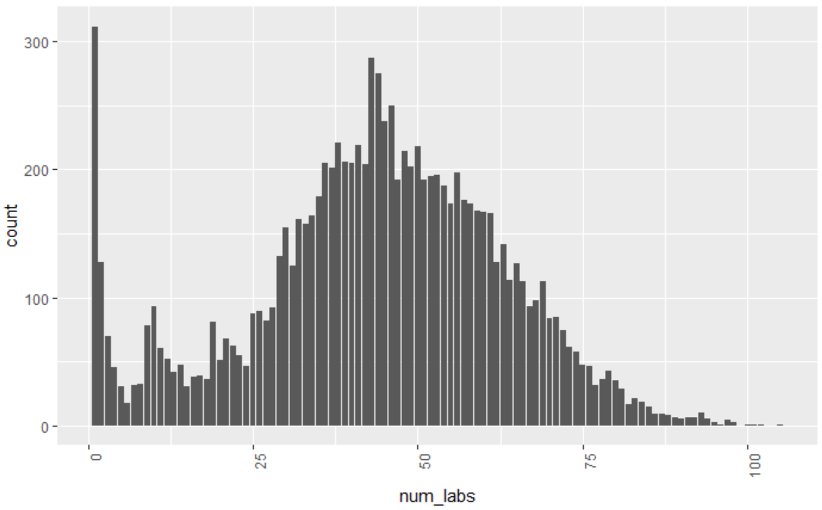
The additional variable that indicates the number of days spent in the hospital should not be included in the model. Typically, variables collected after the time of model application should not be included in a model. Here are two reasons it shouldn't be used:

1. The variable would likely leak information about our target variable, leading to artificially high model performance that could not be realized when the model was used. For example, the number of laboratory procedures might be impacted by the number of days a person is admitted because the hospital might periodically perform labs tests to monitor changes in the patient's health during the hospital stay.
2. Unless hospital administrators know how many days a patient is going to have in the future, they would not be able to use the information in any way. The model cannot be applied at the time of admission if all of the inputs are not known.

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

Most candidates performed well on this task. Lower performing responses often excluded a high-level description of the data source and did not include any visualizations.

The initial data contained 10,000 records based on historical inpatient encounters for patients with diabetes from U.S. hospitals between 1999 and 2008. The dataset contained the following variables about the hospital stay, the patient, their recent treatments, and their treatment upon admission: num_labs (the target variable), gender, age, race, weight, admit_type_id, metformin, insulin, readmitted, num_procs, num_meds, num_ip, num_diags. The distribution of the target variable, with high concentrations at very small values and then a symmetric distribution thereafter, is shown below.



The following are specific issues I explored and adjustments I made.

Completeness and Reasonableness

When reviewing the completeness of the data, I considered the percentage of missing records for each variable and whether having missing values impacted the target variable. Two records were missing gender data, and were removed. About 97% of the records were missing weight information, so that variable was removed. The race variable had 222 missing values, but the records did not appear to be missing at random, so I created a new race category called “Missing.”

Ethical Concerns

Including the race variable in the model could lead to discriminatory model applications, so we should consider whether to remove the variable from the final model. Before making decisions based on the final model application, we could use the race data to understand whether or not there are any unfair impacts created. We may also want to discuss the issue with legal experts and MACH.

Relevance

Note that the data is limited to diabetes patients, which limits the applicability of this work. I explored the data to see if the variables were appropriate for the problem we are addressing. Descriptive statistics and visualizations were used to analyze the univariate distributions and bivariate (between the variable and target variable) distributions for each variable. Three variables appeared particularly relevant – likely to predict our target variable, the number of lab tests administered during the inpatient stay. Based on my analysis, the type of admission (e.g. emergency vs. elective), the number of medications administered in the prior year, and the patient’s age are expected to lead to more lab tests.

Additional data preparation steps included recoding variables as factors, combining factor levels, and reordering the factor levels.

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

Most candidates were able to describe principal components analysis, but many were unable to describe advantages and disadvantages of using PCA for this problem. The best candidates discussed how correlated variables, centering, and scaling impact PCA. Many candidates thought that PCA dealt with relating the variables to the target variable somehow, leading to poor scores.

Principal component analysis is a method to summarize high dimensional numeric data with fewer dimensions while preserving the spread of the data. It can be particularly helpful when variables are highly correlated. PCA finds perpendicular linear combinations of the input variables (which are typically centered and scaled) called principal components that maximize variance to retain as much information as possible. The principal components are ordered according to their variance. The sum of their variances is the total variance explained. It is then common to look at the proportion of variance explained by each principal component to decide how many PCs to use.

<u>Advantages of PCA</u>

PCA could allow us to build a simpler model with fewer features. When exploring data, PCA can help visualize high-dimensional data to explore relationships between variables. PCA can help identify latent variables; in our case a variable named “overall health” could be based on combinations of our input variables.

<u>Disadvantages of PCA</u>

Using a subset of the principal components results in some information loss. The principal components will be less interpretable than the original variable inputs. Because the hospital administrators want to understand the factors that impact the number of labs, using PCA may not be appropriate for this problem.

The PCA Analysis yielded the following output:

```
Importance of components:
                         PC1    PC2    PC3    PC4
Standard deviation     1.238 1.0520 0.9066 0.7343
Proportion of Variance 0.383 0.2767 0.2055 0.1348
Cumulative Proportion  0.383 0.6597 0.8652 1.0000
                 PC1         PC2         PC3        PC4
num_procs -0.5580785  0.45338079 -0.35130693 -0.5996480
num_meds  -0.6781220  0.06304633 -0.08316373  0.7275022
num_ip    -0.1221477 -0.77549429 -0.60842859 -0.1162033
num_diags -0.4623623 -0.43483287  0.70674033 -0.3125053
```

The first table shows the proportion of variance explained by each component. PC1 has about 38% of the total variance. The bottom row of that same table shows the cumulative variance explained. If we used 3 PCs in our model, we would retain about 87% of the information.

The second table shows the coefficients applied to the input variables to create the principal components. The size and sign of the coefficients indicates the relative influence each input variable has on the PC. For PC1, the number of medications has the most influence, followed by the number of procedures, the number of diagnoses, and the number of inpatient visits.

Using only the first principal component in our model would result in significant information loss since it only explains 38% of the variance. For this reason, additional PCs should be included or the original input variables should be used instead. If we need to include 3 PCs to retain avoid losing a lot of information, it might be better to keep 4 input variables that are more easily interpreted.

Task 6 – Construct a decision tree (10 points)

> *This task was made up of several very specific subtasks, and it is important that candidates make sure they are performing the exact task requested. Many candidates seemed to ignore detailed requests in the problem statement, resulting in poorer scores. Some candidates did not explain why pruning was important in the context of the business problem. When choosing the optimal CP parameter, alternative approaches were able to earn full points as long as the approach was justified. Where specifically requested or necessary for facilitating discussion (e.g. discussing the CP table or plot), candidates should include R output in their report. Candidates were expected to interpret all leaves of the pruned tree.*

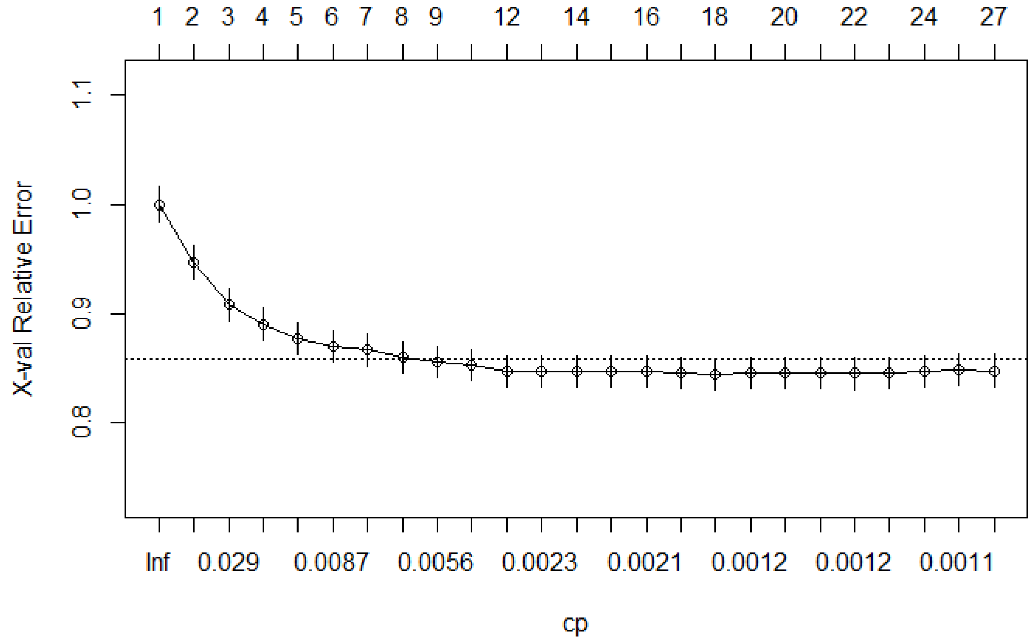
When a decision tree is trained, it can become very large and include splits that are not particularly valuable for predictions on new data. When examining the fit of a tree, it is a good idea to try to prune a tree when it has many splits that do not improve the performance. Pruning reduces the size of the tree, hopefully removing less valuable splits from the tree. This process reduces overfitting the tree on the training data, can lead to better predictions, and results in a simpler, more interpretable tree.

To help the hospital administrators understand the factors that lead to more lab tests, it is important that our tree can be understood and that we ignore any relationships that are based on noise in the training data. For these reasons, pruning should be used.

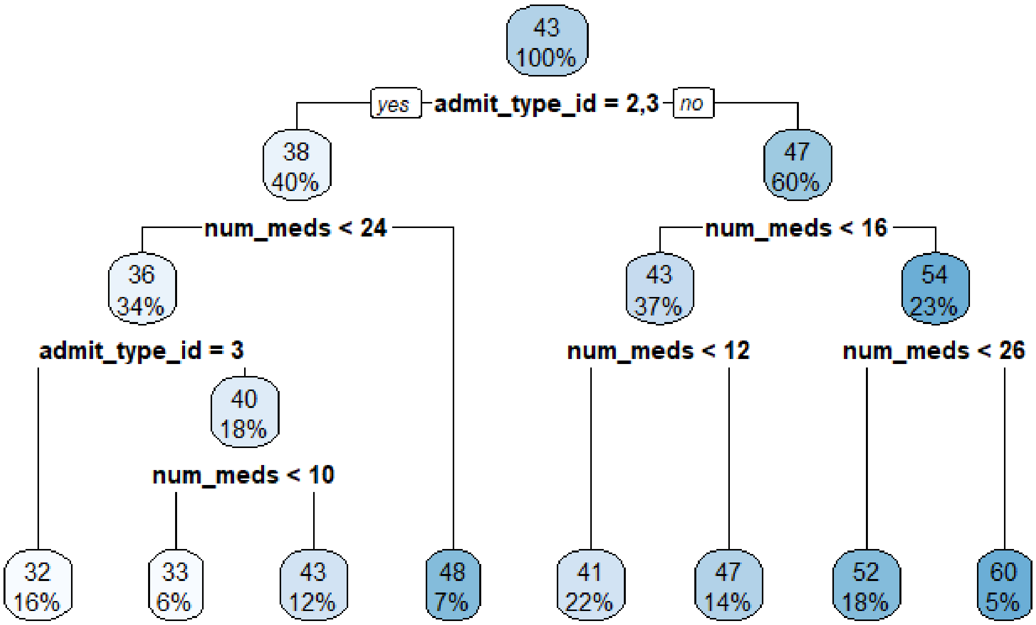
The following output is from the initial unpruned tree. The optimal CP parameter is the one that minimizes the cross validation error (in the xerror column). Row 17 accomplishes that, with CP = 0.001242402. Pruning with this CP value will result in a tree with 17 splits and 18 leaves.

```
            CP nsplit rel error    xerror       xstd
1  0.052549375      0 1.0000000 1.0002234 0.01623384
2  0.040657044      1 0.9474506 0.9470172 0.01562055
3  0.020630619      2 0.9067936 0.9080491 0.01511151
4  0.011963871      3 0.8861630 0.8905970 0.01477732
5  0.009222034      4 0.8741991 0.8772946 0.01468477
6  0.008263540      5 0.8649771 0.8706159 0.01456317
7  0.006652854      6 0.8567135 0.8667220 0.01451012
8  0.006596162      7 0.8500607 0.8606864 0.01441693
9  0.004711295      8 0.8434645 0.8559989 0.01435923
10 0.003280983      9 0.8387532 0.8527496 0.01433076
11 0.002363181     11 0.8321912 0.8479815 0.01425771
12 0.002329369     12 0.8298281 0.8470266 0.01427858
13 0.002276467     13 0.8274987 0.8470266 0.01427858
14 0.002251511     14 0.8252222 0.8470266 0.01427858
15 0.001949715     15 0.8229707 0.8469483 0.01431242
16 0.001276336     16 0.8210210 0.8459324 0.01435956
17 0.001242402     17 0.8197447 0.8444454 0.01438970
18 0.001239697     18 0.8185023 0.8456433 0.01438731
```

```
19 0.001235297     19 0.8172626 0.8456433 0.01438731
20 0.001227634     20 0.8160273 0.8457108 0.01439279
21 0.001143473     21 0.8147996 0.8455746 0.01441509
22 0.001114931     22 0.8136562 0.8462970 0.01441423
23 0.001069155     23 0.8125412 0.8477991 0.01444465
24 0.001042787     24 0.8114721 0.8486127 0.01447410
25 0.001000000     26 0.8093865 0.8480624 0.01446675
```



Using CP = 0.0066 to prune the tree results in a tree with 8 leaves.



The table below shows the mean squared error for the original (unpruned) tree and the pruned tree with 8 leaves on the training and test data. The statistic is a way to measure the error between the predicted values and the actual values, so smaller values are better. Based on the table below, the original tree performed better than the pruned tree, but the performance difference is small compared to the added simplicity.

Tree	**Mean squared error on Train Data**	**Mean squared error on Test Data**
Original	306.9206	311.8836
Pruned	322.3443	321.2924

The pruned tree allows for 8 possible predicted values, 1 for each of the leaves. The tree can be interpreted as 8 series of if statements. The possibilities are summarized below for the 8 leaves pictured above from left to right. Note that predicted number of labs is rounded based on the decision tree image.

1. If admit_type_id = 3 and num_meds < 24, predict 32 labs
2. If admit_type_id = 2 and num_meds < 10, predict 33 labs
3. If admit_type_id = 2 and 10 <= num_meds < 24, predict 43 labs
4. If admit_type_id in [2,3] and num_meds >= 24, predict 48 labs
5. If admit_type_id = 1 and num_meds < 12, predict 41 labs
6. If admit_type_id = 1 and 12 <= num_meds < 16, predict 47 labs
7. If admit_type_id = 1 and 16 <= num_meds < 26, predict 52 labs
8. If admit_type_id = 1 and num_meds >= 26, predict 60 labs

According to the pruned tree, the distinguishing factors for predicting the number of labs administered during the inpatient visit are the admit_type_id and the number of medications.

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

> *The most successful candidates were able to discuss their distribution choices in light of the data structure and the business problem. Some candidates failed to point out that using the PCs may make it more difficult for the hospital administrators to understand the drivers of longer hospital stays.*

Binomial distributions are typically used when there are only two outcomes, so it would not be a good fit. Gamma is used for non-negative continuous variables. Although the target variable we have is discrete because a patient is unable to have a fraction of a lab test, it could also be modeled as continuous, so the gamma distribution could be a viable alternative.

GLM	**Mean squared error on Train Data**	**Mean squared error on Test Data**
All variables except PC	314.4913	309.4742
With PC and without original numeric variables	321.1392	318.0764

The GLM with the PC in place of the numeric variables performed slightly worse, and using the PC instead would make interpreting the factors that lead to more labs more difficult. I will use the model without the PC since it will be easier to gain a better understanding of the factors driving number of labs and has better performance.

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

> *Some candidates did not know whether a higher or lower Pearson goodness-of-fit statistic was good or bad. Candidates were expected to make a clear model recommendation and justify their choice. Better candidates went beyond comparing the performance and variables of the GLM and LASSO models and discussed how both were good or bad for the specific business problem. Either method could be justified and receive full credit.*

The features used by the model are:

- admit_type_id = 2
- admit_type_id = 3
- num_meds
- num_diags
- metformin = Steady

GLM	**Mean squared error on Train Data**	**Mean squared error on Test Data**
GLM selected in Task 7	314.4913	309.4742
LASSO Model	320.06	312.839

The output from the GLM selected in task 7 is below.

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.3988   -1.6230    0.1911    1.7584    8.5019

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.5193058  0.0096933 363.068  < 2e-16 ***
raceMissing                0.0746373  0.0119219   6.261 3.84e-10 ***
raceAfricanAmerican_Asian  0.0199431  0.0047495   4.199 2.68e-05 ***
raceOther                  0.0106988  0.0103475   1.034 0.301160
genderMale                -0.0127053  0.0036832  -3.450 0.000562 ***
age[0-10)                  0.2072197  0.0512644   4.042 5.30e-05 ***
age[10-20)                -0.0267957  0.0242081  -1.107 0.268343
age[20-30)                 0.0167697  0.0166043   1.010 0.312514
age[30-40)                -0.0266132  0.0105678  -2.518 0.011791 *
age[40-50)                -0.0360901  0.0070121  -5.147 2.65e-07 ***
age[50-60)                -0.0207741  0.0057796  -3.594 0.000325 ***
age[60-70)                -0.0202862  0.0052737  -3.847 0.000120 ***
age[80-90)                -0.0051694  0.0056541  -0.914 0.360574
age[90-100)                0.0573919  0.0111195   5.161 2.45e-07 ***
admit_type_id2            -0.1223637  0.0047650 -25.680  < 2e-16 ***
admit_type_id3            -0.3447725  0.0053249 -64.747  < 2e-16 ***
```

```
num_procs                 -0.0022143  0.0011993  -1.846 0.064844 .  
num_meds                   0.0152674  0.0002501  61.056  < 2e-16 ***
num_ip                    -0.0041152  0.0014653  -2.808 0.004978 ** 
num_diags                  0.0136203  0.0010786  12.627  < 2e-16 ***
metforminDown              0.0414281  0.0230066   1.801 0.071750 .  
metforminSteady           -0.0698289  0.0049327 -14.156  < 2e-16 ***
metforminUp                0.0077556  0.0184574   0.420 0.674346    
insulinDown                0.0329844  0.0058583   5.630 1.80e-08 ***
insulinSteady              0.0143135  0.0043266   3.308 0.000939 ***
insulinUp                  0.0417504  0.0062187   6.714 1.90e-11 ***
readmitted<30              0.0068774  0.0059666   1.153 0.249051    
readmitted>30             -0.0081643  0.0040457  -2.018 0.043589 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75405  on 7000  degrees of freedom
Residual deviance: 64726  on 6973  degrees of freedom
AIC: 102566

Number of Fisher Scoring iterations: 5
```

The LASSO coefficients are shown below:

```
raceMissing                .          
raceAfricanAmerican_Asian  .          
raceOther                  .          
genderMale                 .          
age[0-10)                  .          
age[10-20)                 .          
age[20-30)                 .          
age[30-40)                 .          
age[40-50)                 .          
age[50-60)                 .          
age[60-70)                 .          
age[80-90)                 .          
age[90-100)                .          
admit_type_id2            -0.049612646
admit_type_id3            -0.263107268
num_procs                  .          
num_meds                   0.012893391
num_ip                     .          
num_diags                  0.006866476
metforminDown              .          
metforminSteady           -0.018840449
metforminUp                .          
insulinDown                .          
insulinSteady              .          
insulinUp                  .          
readmitted<30              .          
readmitted>30              .          
```

The GLM from task 7 slightly outperformed the LASSO model, but there is more to consider for our business problem which focused on interpretability. Both the LASSO model and the GLM from task 7 are easy to interpret. Clearly, the LASSO model is much simpler since it removed several features that were used in the GLM from task 7. The simplification realized by the LASSO model will help narrow the list of important factors for hospital administrators with only a slight decrease in performance, so I recommend the LASSO model for this business problem.

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

> *Many candidates mixed up bias and variance, or were unable to relate variance to overfitting and bias to underfitting. Better candidates explained how model complexity could refer to both the model type and the features included.*

Bias is the expected loss caused by the model not being complex enough to capture the signal in the data. Variance is the expected loss from the model being too complex and overfitting to the training data.

We typically think of the expected loss as Bias + Variance + Unavoidable error. When building models we are trying to minimize this expected loss, but to do so we often need to find a balance between Bias and Variance. Models with low bias tend to have higher variance and vice versa.

Without regularization, coefficients are found that maximize the likelihood function. This results in models that may not be optimal because coefficients are found even for features that may not be important. This process results in models that are tend to be overfit to the training data; they have high variance. LASSO penalizes models that have large coefficients to the extent that it can shrink coefficients of unhelpful predictors to zero. This is essentially trading some of the high variance from our non-regularized model for a little bit of bias, which can potentially reduce the overall error.

With high variance (overfitting), the model will perform better on the training set than on a test set. With high bias (underfitting), the model will perform poorly on both the training set and the test set. When evaluating a single model, using a test set will help detect whether we have high variance because we can see a difference between the training and test set performance. When comparing models with different levels of complexity, comparing the test set performance and selecting the best performing model can also help us select the model design with the least total error.

Task 10 – Consider the final model (4 points)

> *Most candidates were able to identify advantages and disadvantages of GLMs vs decision trees, but many struggled to do the same with GLMs vs LASSO.*

Advantage of a GLM vs Decision Tree

Consider the case where an increase in the number of medications produces an increase in the expected number of labs. A decision tree separates a numeric variable like the number of medications into buckets. In order for the tree to capture the true relationship, it would need to split on the same variable many times, creating a very complex tree that would be difficult to interpret. On the other hand, a GLM can fit to this relationship with a single coefficient that summarizes how the expected number of labs increases with each unit increase of number of medications (provided there is a simple functional form that describes the relationship). Because our data includes numeric variables, a GLM might capture the nature of the true relationship while being more interpretable.

Disadvantage of a GLM vs Decision Tree

GLMs do not capture the effects of variable interactions automatically. If the right interactions aren't explicitly coded in the GLM, the model may be unable to fit the data well. A decision tree will automatically create variable interactions as it is trained. For example, the pruned decision tree built earlier found an interaction between the type of admission and number of medications. No such interaction was even tried with the GLM.

Advantage of a GLM without removing features vs LASSO regularization

The GLM can retain insignificant factor levels that might be dropped by the LASSO model. This can lead to improved interpretability via comparison of factor levels. The LASSO model has to binarize the factor variables and can shrink individual factor level coefficients to zero. In the LASSO model created earlier, nonzero coefficients were not applied to decreases in metformin, but increases and steady doses had nonzero coefficients. The LASSO model will give the same prediction for a patient that does not take metformin as one who has a decreased metformin dosage because both levels were dropped by the model. In this particular case, the result is not that unreasonable, but if it had occurred with the age variable, the results may have been nonsensical. For example, ages 20-40, and 50-90 might be important, while ages 40-50 are totally ignored by the model.

Disadvantage of a GLM without removing features vs LASSO regularization

Building a GLM without removing features can lead to a model that is overfit to the training data because coefficients will be found even for features that are not important.

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

Candidates were expected to know the application of the coefficients would be multiplicative. Many candidates struggled to explain how to use the model using language appropriate for the client.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)      3.523599   0.006813 517.168   <2e-16 ***
num_diags        0.012169   0.000861  14.134   <2e-16 ***
num_meds         0.015512   0.000183  84.768   <2e-16 ***
admit_type_id2  -0.123865   0.003954 -31.326   <2e-16 ***
admit_type_id3  -0.343226   0.004308 -79.672   <2e-16 ***
metforminDown   -0.004445   0.020546  -0.216    0.829    
metforminSteady -0.073874   0.004045 -18.263   <2e-16 ***
metforminUp      0.014587   0.015303   0.953    0.340    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Positive coefficients for an input variable increase the predicted number of labs, while negative coefficients decrease it, with numbers further from zero having larger effects. These model coefficients can be translated into factors that can be multiplied together to determine a patient's predicted number of labs, where 33.9 labs (interpreted from the intercept) are predicted before applying any factors. The table below illustrates how the coefficients are interpreted for the more impactful items. If a patient is admitted for an elective reason, their number of labs on average will be 71% of the number of labs for a patient admitted for an emergency. For the number of diagnoses and number of medications, the multiplier is applied per diagnosis or medication. A patient with 8 medications in the prior year would have, on average, a number of labs that is 101.5% the number of labs for a patient with only 7 medications in the prior year.

Input	Input Type	Coefficient	Interpreted Coefficient
If the patient is admitted for an elective reason	Categorical	-0.343226	0.709478
For each additional medication in the prior year	Numeric	0.015512	1.015633

Task 12 – Executive summary (20 points)

Rather than restating information from prior tasks, candidates were expected to alter their messaging for the intended audience. Often this includes avoiding overly technical language, discussing topics at a different level of detail, and translating performance metrics to be more meaningful to the reader. Brief discussions about approaches attempted are acceptable, but candidates should avoid lengthy discussion about models or techniques that were not ultimately selected. The best candidates were able to incorporate the business context of the problem throughout their summary.

To: Merged and Acquired Clinics and Hospitals Executives

From: Actuarial Analyst

You have asked us to build a model that yields insights about the factors driving the number of lab tests administered during an inpatient hospital visit, so the hospital administrators can better understand and manage patient needs. We were supplied with 10,000 observations based on historical inpatient encounters for patients with diabetes from U.S. hospitals between 1999 and 2008. Each observation contained information about the number of lab tests administered, the patient, their recent treatments, and their treatment upon admission. The model we constructed identifies information that can be used to predict the number of labs administered during inpatient visits for diabetes patients. The model will not be relevant for patients that do not have diabetes. To build a model that generalizes well for all patients, data about other types of patients should be obtained.

Prior to building the model, the data were reviewed for completeness, reasonableness, ethical concerns, and relevance. The variables included the number of lab tests administered, demographic information, the type of hospital admission, history of medical activity in the prior 12 months, and information about diabetes medication changes upon admission (metformin and insulin). The weight variable was discarded because it contained mostly missing values. Observations missing gender information were removed. A separate category was created to address any missing race information.

The final model did not make use of the race variable, but MACH should weigh the risks of using the model without evaluating the impact to different races. The hospital system could be inviting legal action if decisions based on a model, with or without race included, are viewed as discriminatory. To mitigate the risk, additional work could be performed to make sure the races are not disparately impacted by decisions based on the model.

After modifying a few features to prepare them for modeling, I tried a variety of models to see which would best explain the factors affecting the number of labs. Each model was calibrated using 70% of the data and then its performance was measured using the other 30%. This process helps identify models that adequately capture the patterns in the data and generalize well to new data. Many of the models had similar performance, but they had different levels of interpretability.

The selected generalized linear model used input variables identified using regularization. This model had reasonable prediction performance while offering insights into the factors affecting the number of lab tests administered. Model performance was measured using mean squared error, and a smaller value indicates a better model. The selected model had a mean squared error of 309.77, which was nearly as good as the best model performance, 309.47. The best performing model used 7 additional input variables, though, so it was seen as less useful for identifying the key factors leading to more lab tests. The selected model is easy to interpret, making it a good choice for applications where the factors affecting the predictions need to be understood.

The model coefficients can be used to gain insights about the factors affecting the number of lab tests administered. The model starts with a baseline predicted number of labs for each patient of 33.9 labs. Then, it applies the factors below based on the patient data. Note that values have been rounded. Multiplying by factors greater than 1 increase the predicted number of lab tests, while multiplying by factors less than 1 decrease the predicted number of lab tests.

If admit type_id is urgent	Multiply by 0.883
If admit type_id is elective	Multiply by 0.709
If metformin dosage decreased upon admission	Multiply by 0.996
If metformin prescription exists but dosage unchanged upon admission	Multiply by 0.929
If metformin dosage increased upon admission	Multiply by 1.015
If the patient has taken n medications in the prior 12 months	Multiply by 1.016^n
If the patient has had n diagnoses in the prior 12 months	Multiply by 1.012^n

Many of the factors affecting the number of labs make intuitive sense. Emergency and urgent visits are likely associated with more serious conditions than elective visits, so it isn't surprising that more labs are needed on average. It also makes sense that patients that take more medications or have more conditions (diagnoses) require more labs on average, since they likely are in poorer health.

As a next step, I recommend discussing impacts to protected groups to ensure the model is fairly applied.