## Health Costs Practice Exam Solution

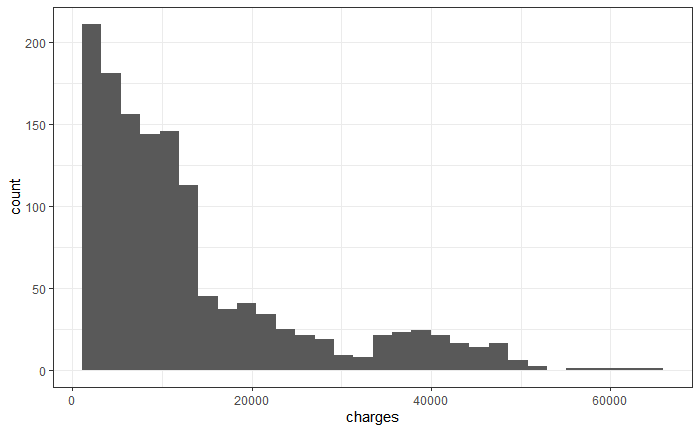
## Task 1 – (5 points) Write a high-level summary of the data

The data consists of 1,338 observations of 6 predictor variables which are related to healthcare costs, or medical claims. The target is the annual medical costs for a policy. Prior to this report, we already cleaned the data and removed missing values.

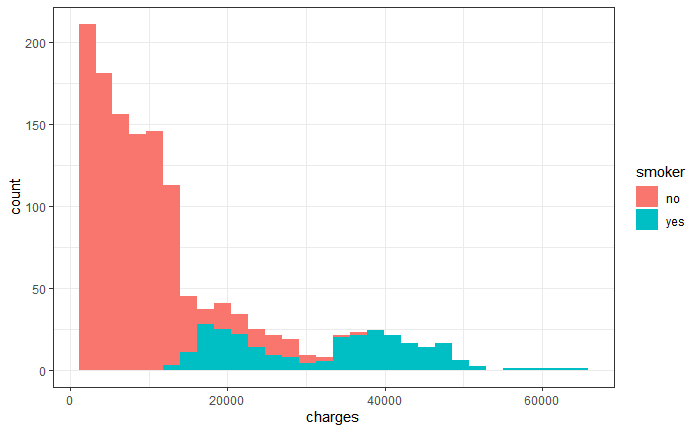
## Task 2 – (5 points) Explore the target variable charges and its relationship to smoker and bmi

The target distribution is positive and right skewed. The mean is 13,270 and the median is only 9,382.

We



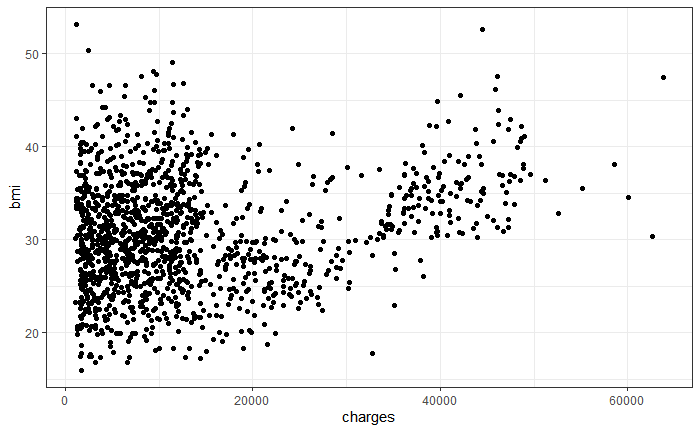
Smokers have higher claims on average than non-smokers. The distribution of smokers is bi-model at about 20,000 and 40,000.

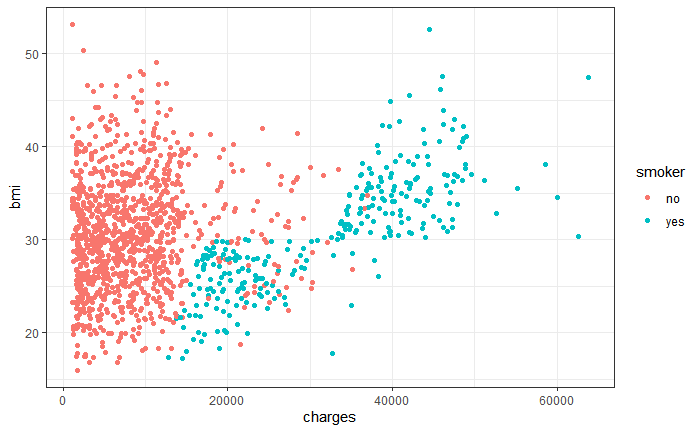


The median is much higher ($32,050) for smokers than non-smokers ($8,434). The mean is similar.

| **smoker**  <fctr> | **mean**  <dbl> | **median**  <dbl> |
| --- | --- | --- |
| no | 8434.268 | 7345.405 |
| yes | 32050.232 | 34456.348 |

BMI is the body-mass-index (weight divided by the height) of a person. We see that as BMI increase, the annual claims tend to increase as well.





For smokers, BMI has a strong correlation with charges. For non-smokers, this correlation is much weaker.

## Task 3 – (9 points) Engineer additional features in order to improve the predictive power of the models

I created three new features: age\_bucket, age\_household ratio, and log\_household size. This will help to improve the predictive power of the model.

Here is a summary of the values:

age\_bucket age\_household\_ratio log\_household\_size

3:383 Min. :0.01562 Min. :0.0000

4:356 1st Qu.:0.03279 1st Qu.:0.0000

2:349 Median :0.05263 Median :0.6931

1:250 Mean :0.06014 Mean :0.5818

3rd Qu.:0.07692 3rd Qu.:1.0986

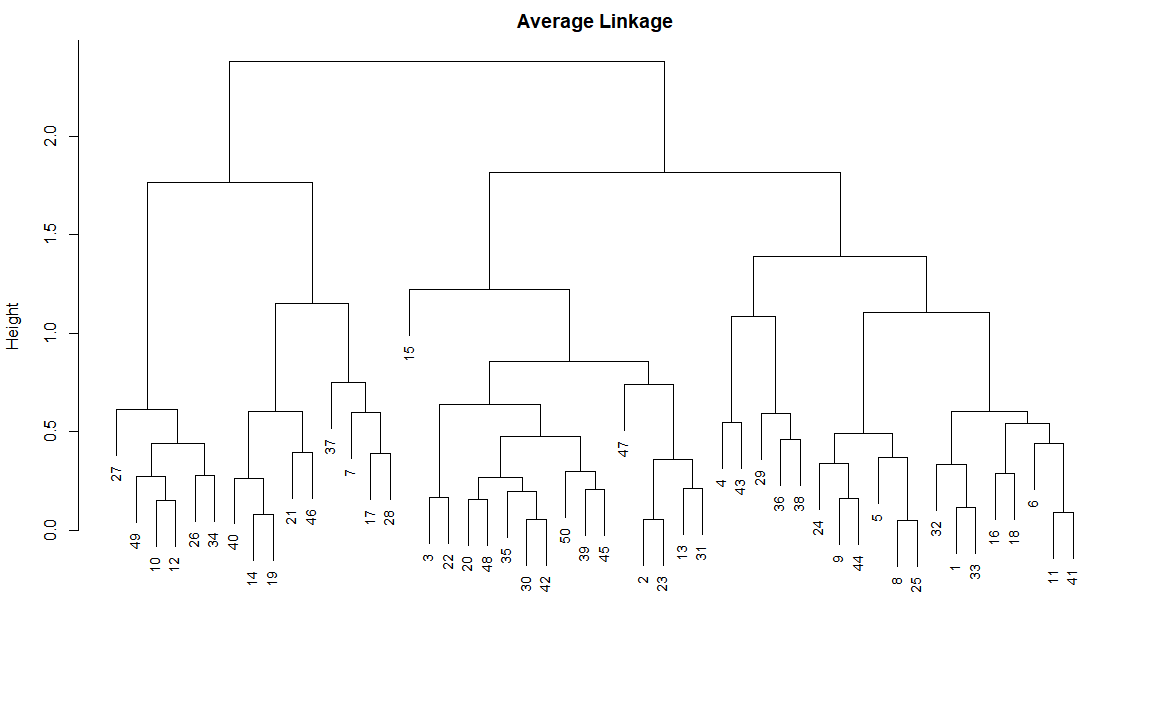
Max. :0.31579 Max. :1.7918

## Task 4 – (8 points) Examine hierarchical clustering on age and bmi

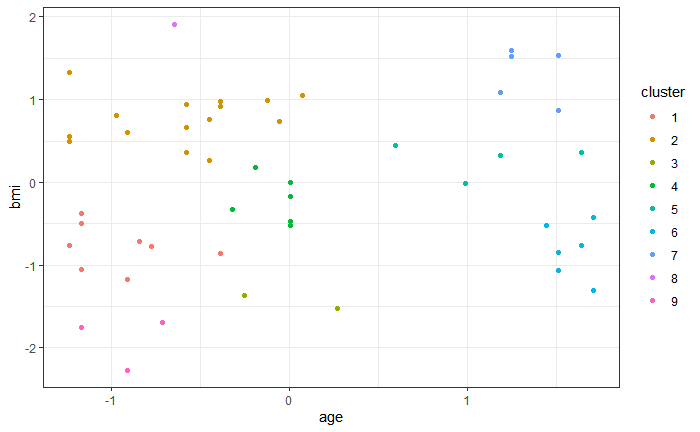
Clustering is an unsupervised learning method that seeks to group together observations into clusters which share similar characteristics. Hierarchical clustering starts with each observation in its own cluster and then groups together those that are similar. Then it iteratively groups these subgroups together in a sequential manner.

Euclidean distance means that the units of age and bmi need to be on the same scale prior to clustering. Otherwise, more weight will be given to age because it has larger values. I rescaled by subtracting the mean and dividing by the standard deviation.

The cutoff controls how many clusters are chosen. For the height equal to 2, there would be 2 clusters. For the height equal to 1.5, there would be three. Each of the final end points (on the bottom of this graph) represent the case when there are 50 clusters.



I chose to set the cutoff at 1 which resulted in 9 clusters



## Task 5 – (5 points) Select an Interaction

An interaction effect is when a predictor variable impacts the response differently depending on the value of another predictor variable. In Task 2, I found that the impact of BMI on Charges was different for smokers vs. non-smokers, and so this means that there is an interaction between smoker and BMI. I added this to the model formula.

## Task 6 – (15 points) Fit two GLMs

I split the data into training (80%) and testing (20%). I train the model on the training data and then evaluate it on the test data.

A glm is a way of relating the target variable to multiple input variables at once. The GLM assumes that the target distribution follows a specific family and that the mean of this distribution is related to the data through a link function.

The response variable claims is continuous and so we can rule out the binomial or Poisson distributions. The only link functions that were appropriate where the identity, log, sqrt, inverse squared, and inverse. The others were not used because this is a regression problem and not a classification problem. The Gamma and Inverse Gaussian are both positive which matches the range of the charges distribution.

Out of all of the Gaussian models, the identity link provided the lowest error.

[1] "family:gaussian, link:identity, test mae:3010"

[1] "family:gaussian, link:log, test mae:3958"

[1] "family:gaussian, link:sqrt, test mae:3304"

[1] "family:gaussian, link:1/mu^2, test mae:5134"

[1] "family:gaussian, link:inverse, test mae:4824"

Out of the Gamma models, the link with the lowest error was again the identity.

[1] "family:Gamma, link:identity, test mae:3020"

[1] "family:Gamma, link:log, test mae:4238"

[1] "family:Gamma, link:sqrt, test mae:3328"

[1] "family:Gamma, link:1/mu^2, test mae:5086"

[1] "family:Gamma, link:inverse, test mae:5142"

I also tried the inverse Gaussian family. The inverse square link did not converge. The inverse link retuned NA values.

[1] "family:inverse.gaussian, link:identity, test mae:3063"

[1] "family:inverse.gaussian, link:log, test mae:7298"

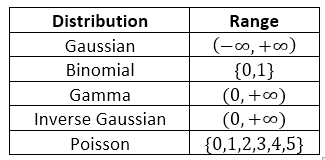
[1] "family:inverse.gaussian, link:sqrt, test mae:3764"

Out of all of these combinations, the two models which had the lowest MAE on the test set where the Gaussian with identity link and the Gamma with identity link.

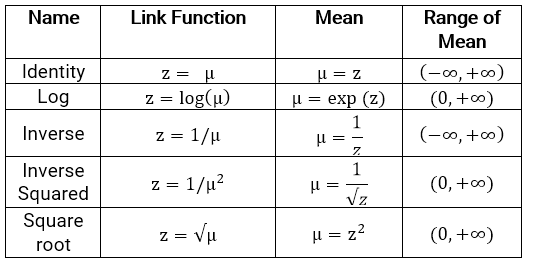
## Task 7 – (10 points) Compare the two models

The target distribution is positive, and so only response families which are positive make sense. Otherwise it would be possible for the model to make negative predictions, even if these are unlikely. The Gamma distribution is always positive but the Gaussian allows for negative values.

The link function relates the mean of the response distribution to the linear predictor. In the case of the Gaussian family, the mean can be any positive or negative number. In the case of the Gamma, the mean must be positive.



The identity link means that the mean of the response distribution can be positive or negative. This makes sense in the Gaussian case but not in the Gamma case. This is shown in the upper right cell of this table, where the range of the mean with the identity link can be positive or negative. If we wanted to use a link function that forced the mean to be positive, we could use the log, inverse squared, or square root links.



The Gamma model with identity link has a problem because the Gamma mean must always be positive, but the identity link technically allows for negative values. The Gaussian model also has a problem because it allows for negative predicted values.

Based on this information, I chose to use the Gamma model with identity link.

[1] "family:gaussian, link:identity, test mae:3010"

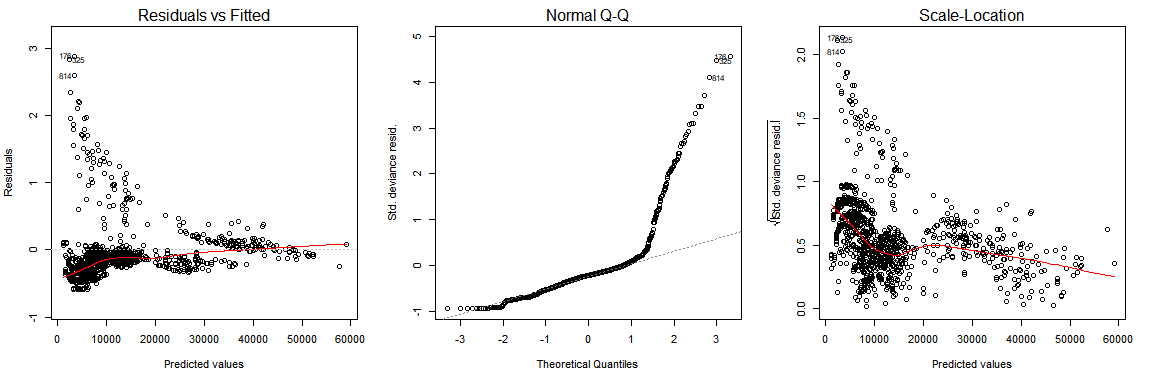
[1] "family:Gamma, link:identity, test mae:3020"

Note: In this problem, both models are a poor fit (as can be seen from the residual plots) and the MAE is almost the same, and so either choice of link and response family is acceptable, provided that the pros and cons of each were mentioned.

## Task 8 – (5 points) Examine the residual plots and determine if the model is a good fit

The assumptions of a GLM are that the deviance residuals are normally distributed and the residuals are scattered around zero. The residual plots provide diagnostics to tell us if the model is a good fit. If this was a perfect model, then the residuals vs. fitted (left) would have no pattern, and normal QQ plot (center) would have all points along the 45-degree line, and the scale-location plot (right) would also be random.

The residuals vs. fitted have a clear pattern. The QQ plot shows very good quantiles up to about 1, and then a sharp deviation. This indicates that the model is a good fit for low-cost claims but a poor fit for high-cost claims. The scale-location plot shows that the residuals change depending on the predicted value.



## Task 9 – (10 points) Fit a GBM using the given parameters

A gbm is an ensemble method that uses boosted decision trees. Each tree considers the residuals from all of the prior trees and the model iteratively improves by adding more trees.

* Ntrees is the number of decision trees that will be used in the ensemble. More trees leads to a more complex model and increases the likelihood of overfitting.
* Interaction depth is an integer that specifies the maximum depth of a tree. Higher values lead to deeper trees, which means that more variables are used and overfitting becomes easier.
* Shrinkage is the parameter applied to each tree in the expansion, also known as the learning rate. This is the rate at which each tree’s contribution adds to the predicted value. A higher shrinkage leads to the model’s predictions getting updated faster (aka, faster learning) which makes it easier to overfit.
* N.minobsinode is an integer specifying the min number of observations in the terminal nodes. A lower value means that more granular trees (aka trees with more branches) are allowed, which increases the likelihood of overfitting.

(Additional points / alternative answers)

* The number of trees and the learning rate have an inverse relationship.  When the learning rate is lower, the model needs to "learn more" and so more trees are needed
* Increasing the number of trees and then using early stopping is the ideal method of tuning GBMs.  This means that the model keeps adding more trees until the error stops decreasing.
* Using more trees with a lower learning rate almost always leads to higher performance but there is a law of diminishing returns. For example, adding 1,000 more trees and decreasing the learning rate by 0.01 might reduce the error more than adding an additional 1,000 trees again and decreasing the learning rate to 0.005.

## Task 10 – (5 points) Compare the GLM to the GBM based on the MAE on the test set

The mean absolute error is the average amount by which the model is above or below the annual medical claims. The MAE of the GLM was about 4000. The MAE for the GBM was about 2500, which is better.

GLM MEA:

3020

GBM MAE:

2562

Advantages:

* Automatically detects outliers, fills in missing values
* Detects non-linearities
* Detects interaction effects
* Has higher predictive power
* There are fewer statistical assumptions needed whereas the GLM made assumptions about the distribution family and link which were not met as we can tell by looking at the residual plots.

Disadvantages:

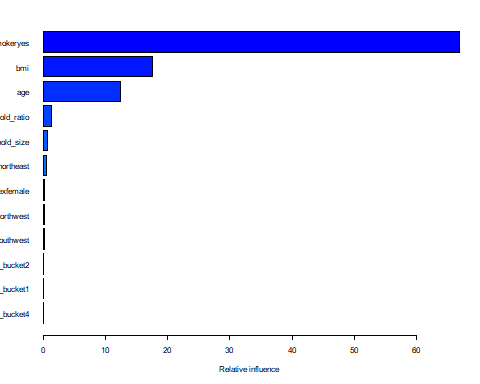
* More difficult to interpret than a GLM
* Takes longer to train
* Is more challenging to deploy than a GLM, which can be converted into a spreadsheet

Based on these, I would choose to use the GBM of this problem. The GLM may work well for low-cost claims, as the QQ-plot was approximately normal for low-values, but there is significant deviation from the normal distribution for higher values. We could truncate the claims at a certain value and use the GLM for these policies.

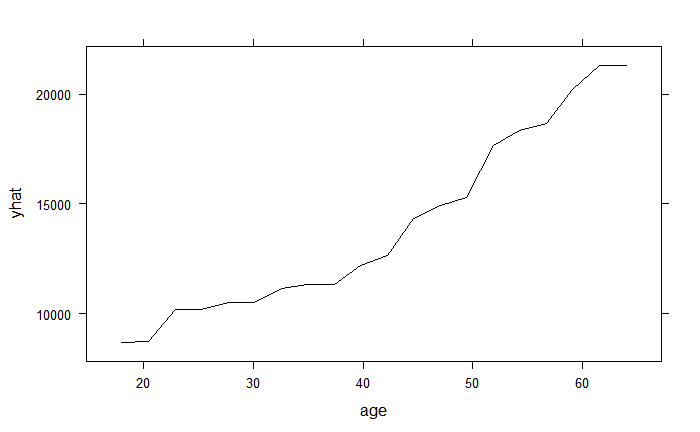
## Task 11 – (5 points) Interpret the partial dependence plots of age, bmi, and smoker status

Variable importance for GBMs is a way of measuring how much each variable contributes to the overall predictions based on the importance of the variable across all of the decision trees that were used in the ensemble. Variables that are more important in trees are high in the tree. Information gain and GINI score are both ways of measuring importance.

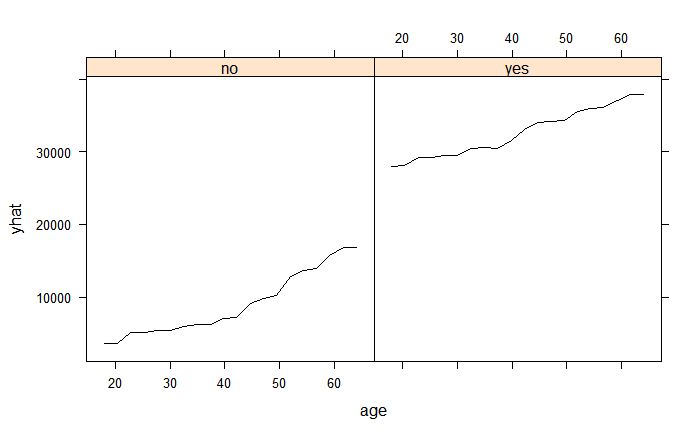
The top three most predictive variables where smoker, bmi, and age, and these explained most of the claims. The other variables were less predictive.



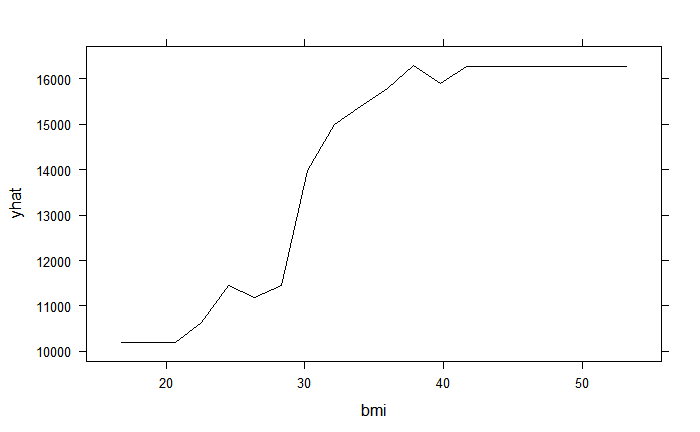
The partial dependence is a way of showing the impact that a variable has on the prediction after adjusted for all other variables. For example, for the variable age, the model looks at records where age = 20, and all of the other combinations of the other variables that are in the training data. Then it takes the average predicted value. This is the point on the graph where age = 20 and yhat = 10,000. It then repeats this process for all other values of age.

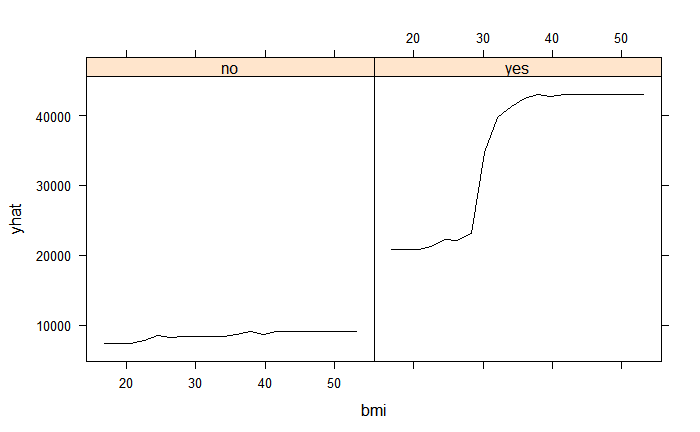


Because this is taking an average, smokers get grouped together with non-smokers. When I split by smoker status, I see that the impact of age on claim costs is more severe (steeper) for smokers than non-smokers.



For BMI, it looks at first like BMI increase medical costs directly.

But when splitting by smoker status, I see that this is only being caused by the smokers.



## Task 12 – (18 points) Executive summary

We can help Freedom Health to build a predictive model that will accelerate their medical underwriting process. This analysis compares two different modeling approaches for predicting future medical costs. Freedom can use this model to assign risk to policies.

We found a new model approach that has much higher accuracy (by a factor of almost 50%), as the average error is about $2500 annually rather than $3000 annually with the current method.

The three factors which contributed most to a person’s medical costs were smoker status, body-mass-index, and age. Costs tend to increase as bmi increases, although this effect is different for smokers vs. non-smokers. The same is true for a person’s age: costs increase as age increases, but smokers have higher costs than non-smokers.

Freedom should have a way of telling if a policyholder is a smoker or not prior to making an underwriting decision, as this has a significant impact on their future costs. This might be difficult info to get because people may not be honest. Age is publically available and so this data should be accessible.

We began with an exploratory analysis of the data. We looked at 1,338 policies from the prior year and had 6 different variables for each related to the policyholder’s demographics, body-mass-index, gender, age, and number of children. We found that the claims distribution was heavily skewed. I found that those with a higher BMI tend to be more expensive on average for smokers than for non-smokers.

We created additional features in order to improve the predictive power. We only began with 6 variables, and so we added another three based on our understanding of health costs, family size, and age categories.

We then considered unsupervised learning in order to find patterns in the policies which we did not observe from looking at the graphs and summary statistics. We used Hierarchical clustering which allows us to set the cutoff threshold to choose the number of groups that we cluster into. We did not find any clear patterns and so this info was not used further.

Freedom has been using GLMs for predicting future claims, and so we first wanted to replicate the type of model that Freedom would have used. A GLM is a commonly used modeling approach that relates the target variable to values of several inputs variables at once. This is commonly used in insurance rating because it is easy to interpret and can adjust for different claims distributions. We tested twelve different GLMs and chose the best one based on a blind test on holdout data. We found that the best GLM had an average error (MAE) of $3,020 per year.

We found problems with this model. The assumptions regarding the residuals were not met, indicating that the model is not a good fit. Specifically, the deviance residuals were not normal for large claim amounts, and there were clear patterns in the residuals vs. fitted values. This implies that a GLM is not a good model for this problem if the goal is predictive power.

Then we looked at GBM. This is powerful ensemble method which combines the results of hundreds of individual decision trees into a single model. When we tested this model on the same holdout data as the GLM had been tested on, the error was much lower ($2,561).

This report showed how Freedom Health can improve their underwriting process, as well as detailing steps for building a better model. Future considerations would be to get additional data, as we only had 6 predictor variables. If Freedom can gather more information about its policyholders then the performance of the models will be better.