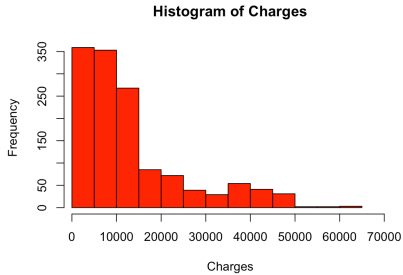


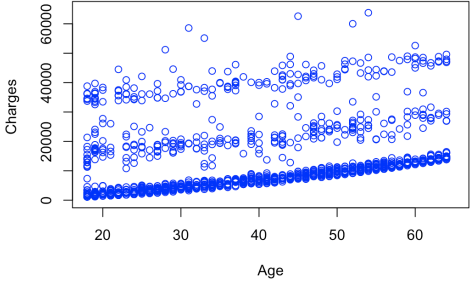
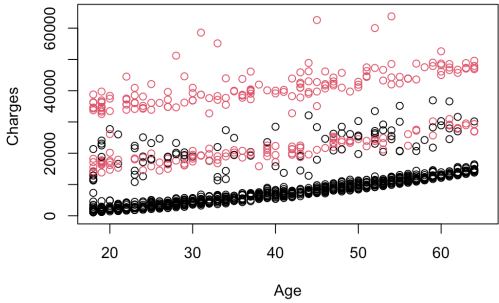
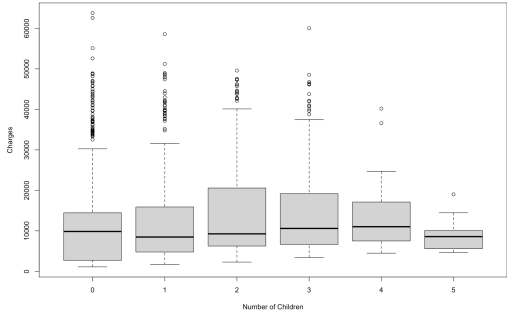
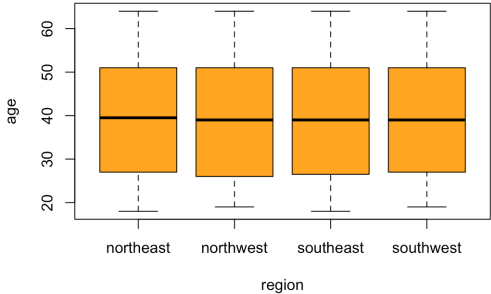
The data set that we planned to analyze was the US Health Insurance Data Set, found on the online data repository Kaggle. This data set contains several attributes that are related to one's insurance premium. This data set includes 1,338 observations, with 7 variables- age, sex, BMI, children, smoker, region, and charges. The description of the variables can be seen below:

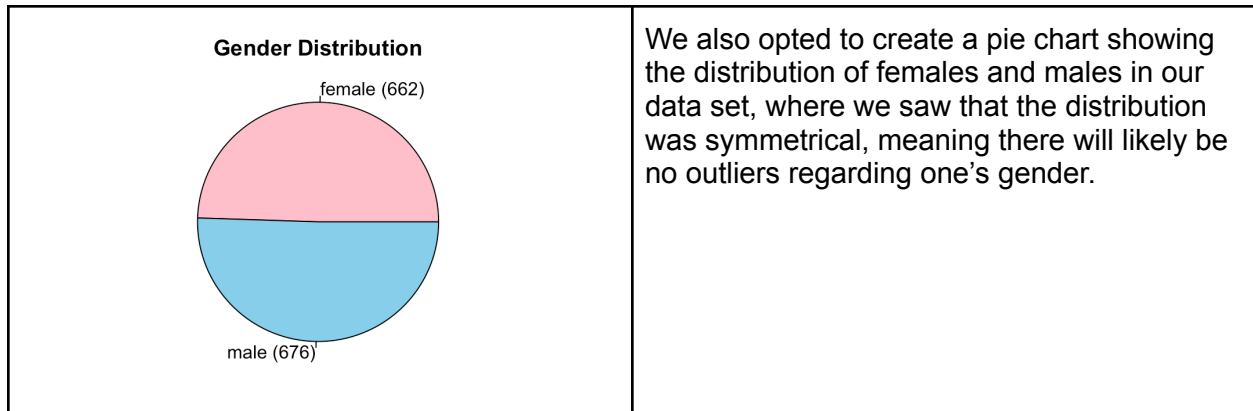
- Age: numerical data, age of primary beneficiary
- Sex: categorical data, gender of insured (male or female)
- BMI: numerical data, body mass index of insured, it is body's mass in relation to the height and weight
- Children: numerical data, number of children that is covered by the insurance
- Smoker: categorical data, whether the primary insured is a smoker or not (smoker vs non-smoker)
- Region: categorical data, insured's residency in the US (Northeast, Southeast, Northwest, Southwest)
- Charges: numerical data, individuals' medical costs billed by the health insurance

We opted to keep all of the variables in the data set, as there were only 6 predictor variables to choose from. This gives us the chance to see which variables were significant, and it also allows us the flexibility of removing some variables when creating and testing new models. We also chose to convert the "children" variable, which was originally a numerical variable, into a categorical variable, given that the number of children a person can have isn't continuous, rather, it is an integer value.

We chose to do many exploratory data analysis models in order to better understand the data:

Model	Description
 <p>A histogram titled "Histogram of Charges" showing the frequency distribution of medical charges. The x-axis is labeled "Charges" and ranges from 0 to 70,000 with major ticks every 10,000. The y-axis is labeled "Frequency" and ranges from 0 to 350 with major ticks every 50. The histogram consists of red bars. The first bar (0-5,000) has the highest frequency, exceeding 350. The frequency decreases sharply for subsequent bins, with a few smaller peaks around 20,000 and 40,000, and a long tail extending towards 70,000, indicating a right-skewed distribution.</p>	<p>Our first plot was a histogram, which shows the distribution of charges from our data set. From this histogram, we were able to determine that the data was heavily skewed to the right, which may hint that there could be possible outliers in our data set.</p>

<p style="text-align: center;">Age vs. Charges</p> 	<p>We also created a plot that compares Age vs. Charges, in which we saw that there was an overall positive trend between the two variables, however there were many data points that were above this cluster (indicating higher charges), which could also indicate possible outliers.</p>
<p style="text-align: center;">Age vs. Charges</p> 	<p>Through further exploratory data analysis, we were able to see that many of these potential outliers were explained by the fact that they were smokers, indicating that being a smoker would likely increase the insurance charges that one may face.</p>
<p style="text-align: center;">Comparison of Number of Children to Charges</p> 	<p>We had also done a model that shows the relationship between the number of children in an insured household and their charges, where we found that those who had the most kids (5) were, on average, charged the least. Also, the majority of the outliers were located in insured individuals without children or only one child.</p>
<p style="text-align: center;">Ages per Region</p> 	<p>We also chose to do more basic analysis, where we saw that the distribution of ages among the 4 regions was roughly symmetrical, indicating that region and age would likely not be an outlier for our data.</p>



- Lastly, we created a correlation matrix to see potential interaction effects, as well as to see which variables most impacted our target variable. With this correlation matrix, we were able to see that the “smoker” variable had the highest correlation to charges, which validates the scatter plot that we had previously created. The second highest correlation was the “bmi” variable. On the other hand, the lowest correlation to charges was the “region” variable, with the “sex” variable following shortly after. The correlation matrix can be seen below:

	AGE	SEX	BMI	CHILDREN	SMOKER	REGION	CHARGES	OBESE
AGE	1.0000	-0.0209	0.1093	0.0425	-0.0250	0.0021	0.2990	0.0856
SEX	-0.0209	1.0000	0.0464	0.0172	0.0762	0.0046	0.0573	0.0473
BMI	0.1093	0.0464	1.0000	0.0128	0.0038	0.1576	0.1983	0.7992
CHILDREN	0.0425	0.0172	0.0128	1.0000	0.0077	0.0166	0.0679	0.0098
SMOKER	-0.0250	0.0762	0.0038	0.0077	1.0000	-0.0022	0.7873	0.0000
REGION	0.0021	0.0046	0.1576	0.0166	-0.0022	1.0000	-0.0062	0.1103
CHARGES	0.2990	0.0573	0.1983	0.0679	0.7873	-0.0062	1.0000	0.1995
OBESE	0.0856	0.0473	0.7992	0.0098	0.0000	0.1103	0.1995	1.0000

When analyzing this data set, we chose the “charges” variable as our target variable, hoping to address the question of which variables would impact the insurance charges a person may have. We chose to treat this data set as a linear model, implementing multiple linear regression models and regression trees until we decided on a “best” model. Our model development was heavily reliant on the correlation matrix, as understanding how the variables impacted each other had allowed us to create interaction effects and also allowed us to take away variables that weren’t significant, all in an effort to create our “best” model. The description of our models can be seen below:

- Model1 = `lm(charges~age+sex+bmi+children+smoker+region)`
 - This model contains all of the variables that were found in the original data set, and this would be our “base model.”

- The statistically significant variables in this model were “age,” “bmi,” “children,” and “smoker.”
- Model2 = $\text{lm}(\text{charges} \sim \text{age} + \text{sex} + \text{bmi} + \text{children} + \text{smoker})$
 - With this model, we opted to remove the “region” variable, as it was the variable that had the weakest correlation to the target variable in our correlation matrix.
 - The statistically significant variables in this model were “age,” “bmi,” “children,” and “smoker.”
- Model3 = $\text{lm}(\text{charges} \sim \text{age} + \text{sex} + \text{children} + \text{region} + (\text{smoker} * \text{bmi}))$
 - In this model, we chose all of the variables in the original data set, but this time, added an interaction effect between the “smoker” and “bmi” variables, as they had the highest correlation to charges in the correlation matrix.
 - The statistically significant variables in this model were “age,” “children,” “regionsoutheast,” “regionsouthwest,” “smoker,” and the interaction effect.
- Model4 = $\text{lm}(\text{charges} \sim \text{age} + \text{sex} + \text{children} + (\text{smoker} * \text{bmi}))$
 - For this model, we kept the same interaction effect that was in Model3, but instead, removed the “region” variable, as it was the variable with the weakest correlation to our target variable.
 - The statistically significant variables in this model were “age,” “children,” “smoker,” and the interaction effect.
- Model5 = $\text{lm}(\text{charges} \sim \text{age} + \text{children} + (\text{smoker} * \text{bmi}))$
 - Our fifth model had the same interaction effect between “smoker” and “bmi,” but we also removed the two weakest variables in relation to our target variable, which were the “sex” and “region” variables.
 - The statistically significant variables in this model were “age,” “children,” “smoker,” and the interaction effect.
- Regression Tree using 10 fold cross validation
 - We also opted to create a regression tree, which gave us a best model where the only variable was “smoker.”

When training and testing our data, we chose to do an 80-20 split, and made sure to use the same training and testing data across our models to ensure that our validation was accurate so that we can find our true best model. We chose to use multiple measurements for our model evaluations, including the MSE, MAE, RMSE, Multiple R^2 , and Adjusted R^2 . The results of our models can be found below:

	Model 1	Model 2	Model 3	Model 4	Model 5	Regression Tree
MSE	32,907,620	32,576,935	22,025,788	21,640,672	21,585,210	239,952,976
MAE	3,940	3,963	2,687	2,706	2,688	
RMSE	5,763	5,785	4,587	4,616	4,624	
Multiple R^2	0.7533	0.7524	0.8424	0.8409	0.8405	

Adjusted R ²	0.7514	0.7512	0.8411	0.8400	0.8398	
-------------------------	--------	--------	--------	--------	--------	--

From this table, we are able to see a whole picture of the performance of our models. The ranking of our models can be seen below:

1. Model 3- Although Model 3 didn't have the lowest MSE of the 6 models we have, it had the lowest MAE and RMSE, and also had the highest Multiple R² and Adjusted R², making it the model with the best performance. This shows the value of the interaction effect between "bmi" and "smoker."
2. Model 4- Model 4 had a higher MSE and MAE than Model 4, but it also had a lower RMSE, and a higher Multiple R² and Adjusted R². The difference between Models 3 and 4 is quite apparent, and shows that removing the "region" variable actually weakens the performance of our model, which can be explained by the "region" variable being statistically significant in Model 3.
3. Model 5- Model 5 was very similar to Model 4, with the changes in model performances varying slightly. While the MSE and MAE were lower in Model 5, it ultimately performed weaker in the other evaluation measurements. While taking away the "sex" variable didn't make too much of a difference, it is clear that keeping all of the variables in the model shows the strongest performance.
4. Model 1- Model 1 had one of the weaker overall performances, with high MSE, MAE, and RMSE values, and somewhat low Multiple R² and Adjusted R² values, proving that this isn't a strong model, especially in comparison to the higher ranking models.
5. Model 2- Model 2 had the weakest performance in regard to linear regression models, with the highest MSE, MAE, and RMSE values and lowest Multiple R² and Adjusted R² values. Model 2, similar to Models 4 and 5, shows that taking away these other variables actually creates a weaker model.
6. Our model that was created through a regression tree was actually the weakest of all of the models. The unusually high MSE, which is ten times higher than the average MSE of the linear regression models, caused us to dismiss this model entirely.

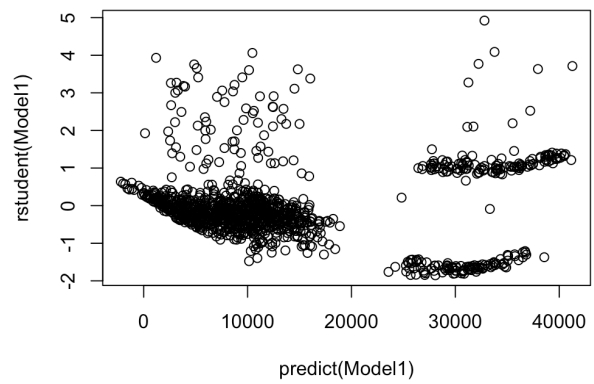
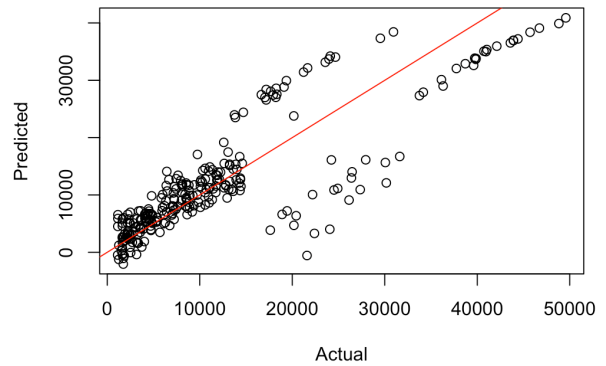
Based on our model evaluations, we ultimately decided that Model 3 was our best model to represent our data. It had the best performance overall, and this was proven when we decided to test our models after the training aspect. We plotted the Actual vs. Predicted scatterplot, and used a red abline as a base to show if our model had accurately predicted the correct "charges" amount. Even by the visual of this scatterplot, it is clear that Model 3 is the best model, as most of the points are on, or close to, the abline. We also used studentized residuals to see how many outliers were present in our models. Though Model 3 had outliers that reached up to 6 standard deviations, we still believe it is the best model for this data set. The plots of our testing results and studentized residuals can be seen below:

Model	Plot
-------	------

Model 1

`lm(charges~age+sex+bmi+children+smoker+region)`

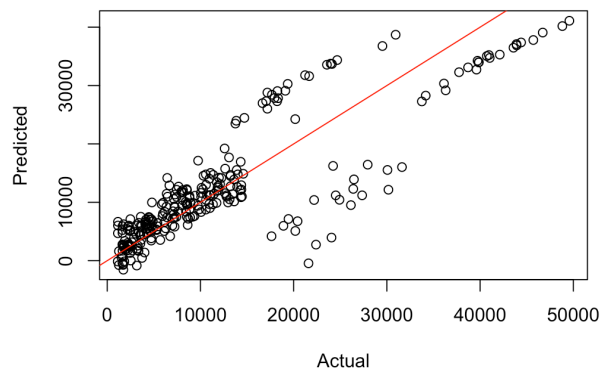
Model 1 Actual vs Predicted

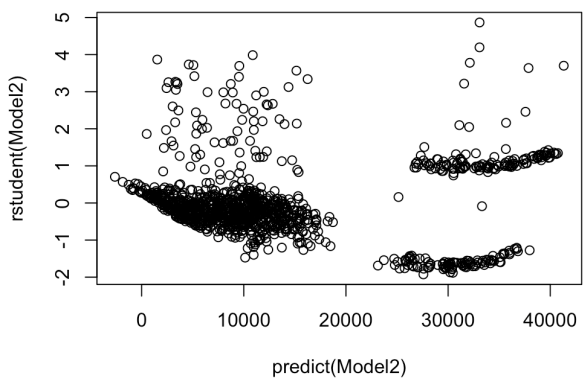
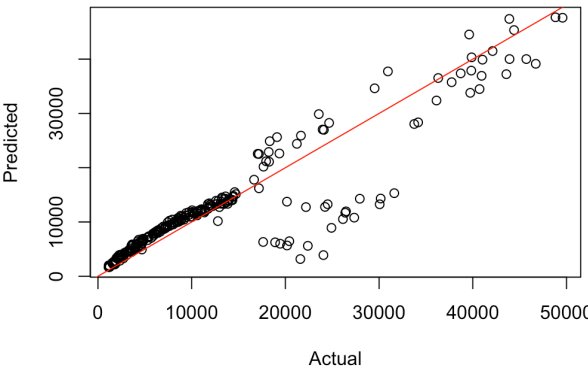
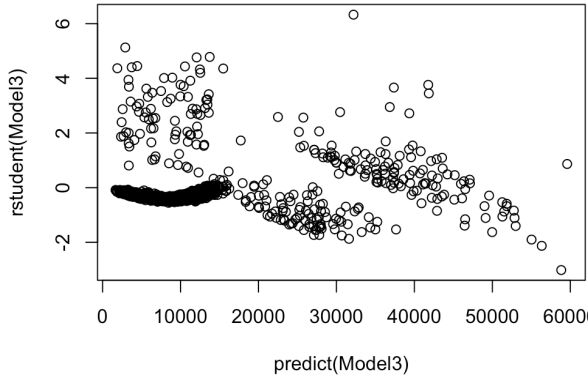


Model 2

`lm(charges~age+sex+bmi+children+smoker)`

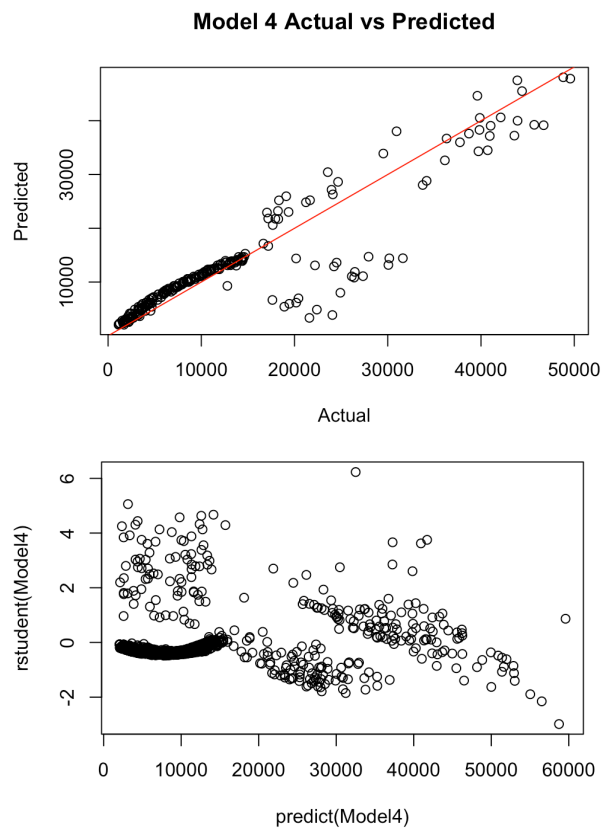
Model 2 Actual vs Predicted



	 <p>A residual plot for Model 2. The x-axis is labeled 'predict(Model2)' and ranges from 0 to 40,000. The y-axis is labeled 'rstudent(Model2)' and ranges from -2 to 5. The data points are scattered around zero, with a slight downward trend as predicted values increase, suggesting some non-linearity or heteroscedasticity.</p>
<p>Model 3</p> <p>$\text{lm}(\text{charges} \sim \text{age} + \text{sex} + \text{children} + \text{region} + (\text{smoker} * \text{bmi}))$</p>	<p>Model 3 Actual vs Predicted</p>  <p>A plot of predicted vs actual values for Model 3. The x-axis is labeled 'Actual' and ranges from 0 to 50,000. The y-axis is labeled 'Predicted' and ranges from 0 to 30,000. A red diagonal line represents the identity line (y=x). Most data points are clustered below the identity line, indicating that the model tends to underpredict higher actual values.</p>  <p>A residual plot for Model 3. The x-axis is labeled 'predict(Model3)' and ranges from 0 to 60,000. The y-axis is labeled 'rstudent(Model3)' and ranges from -2 to 6. The data points show a clear downward trend, with residuals being positive for low predicted values and negative for high predicted values, indicating a strong negative linear relationship in the residuals.</p>

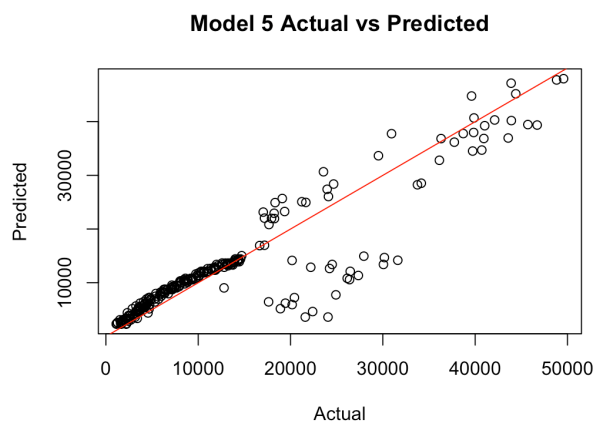
Model 4

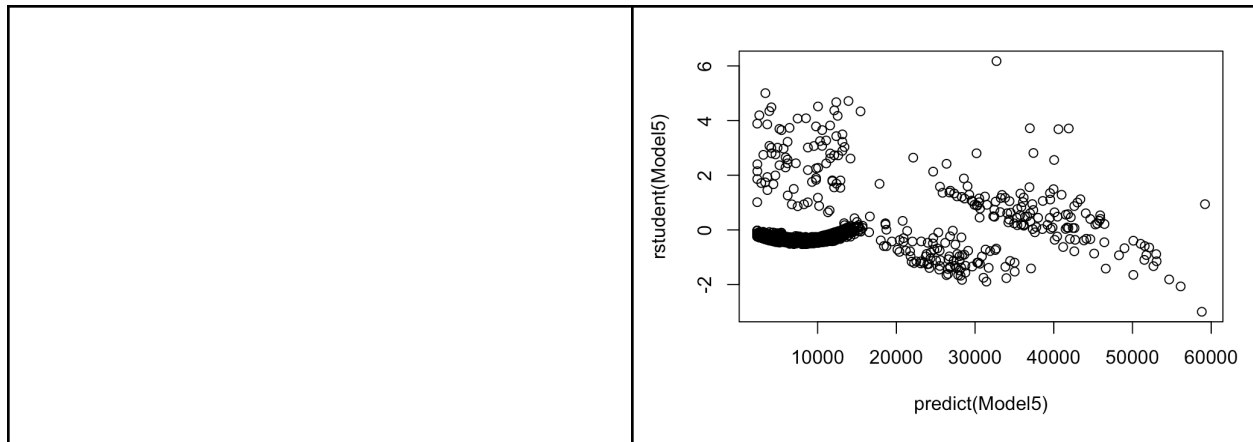
$\text{lm}(\text{charges} \sim \text{age} + \text{sex} + \text{children} + (\text{smoker} * \text{bmi}))$



Model 5

$\text{lm}(\text{charges} \sim \text{age} + \text{children} + (\text{smoker} * \text{bmi}))$





From completing the analysis of this data set and coming up with our model, it is clear that, while all of the 6 factors present were important in determining the charges that one would face regarding their insurance, the overall health of that person matters most, specifically with whether or not a person is a smoker or if they have an unhealthy BMI. This interaction between being a smoker and unhealthy BMI outweighs all other factors, as we saw in the correlation matrix, as well as Models 4 and 5.

Our data set and analysis ultimately has real world applications- in a country where people are drowning in medical debt and low health insurance coverage, people are forced to be proactive with their health in order to avoid these high insurance charges. There is an unsurprising correlation between being a smoker and also having an unhealthy BMI, as these two factors require medical attention to be treated. Though region wasn't an important factor in determining charges in this specific data set, it is clear that all of these variables are related to hospital visits, which ultimately leads to insurance charges.