Medical Insurance Regression Analysis

By: Zachery Gebreab, Kyle Keshmeshian,

John Moro Dutra, Kenneth Nguyen

1 Dataset Introduction

Medical insurance plays a critical role in providing financial protection and access to healthcare services for individuals and families. Understanding the factors that influence medical insurance charges is essential for insurers, policymakers, and healthcare professionals so they can make informed decisions regarding coverage, pricing, and resource allocation.

From Kaggle, we have a comprehensive dataset of medical insurance records to investigate the relationship between various demographic, lifestyle, and medical factors and insurance premiums. The dataset encompasses a diverse range of individuals across different age groups, genders, geographic regions, and health profiles. The dataset includes 2,772 observations of 7 variables.

We choose medical insurance charges, a continuous variable, to be our response. The rest of the variables would be our regressors including age, BMI, children, sex, region, and smoking status; 3 of which are numerical and 3 are categorical, respectively. The primary objective of this analysis is to develop a predictive model using linear regression techniques to estimate insurance premiums based on the available demographic and health-related variables.

1.1 Data Cleaning

Our dataset comes from Kaggle and was built to train a machine-learning model. The data is, for the most part, already clean. Of the 2,700 observations, none contain null values, incomplete data, or incorrect formatting. However, we did discover that more than half of the dataset contained duplicate values. After removing duplicate values from our dataset, our observations decreased to about 1,337.

2 Exploratory Data Analysis

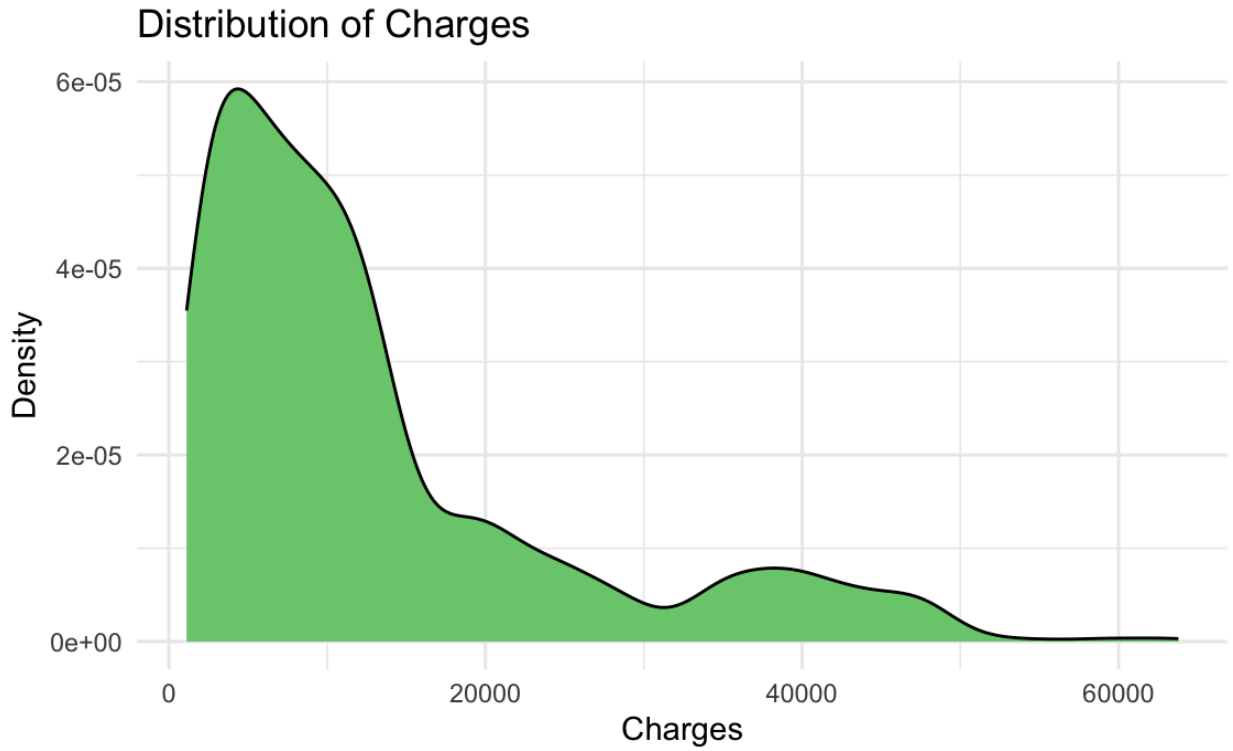


Figure 2.1 - Density Plot of the Charges

To begin our exploratory data analysis, we decided to take a look at the distribution of the charges by creating a density chart. Right away, we observed that the charges had a strongly right-skewed distribution. The majority of the charges remain under $20,000. Another small cluster of charges lies around the $30,000 to $50,000 range. At the tail end, charges get to even above $60,000. Our minimum price is right above $1,000. From this, we can assume that the average person's insurance will be under $20,000. When doing our variable selection, we will look to see what factors bring insurance charges past that threshold.

We then decided to look at the charges by region to see if there were any drastic differences between regions. When looking at the charges by region, the southeast had higher total insurance costs by around a million dollars more than the next leading region. The southwest and northwest regions have approximately $4,000,000 total with the northeast region having a slightly higher total.

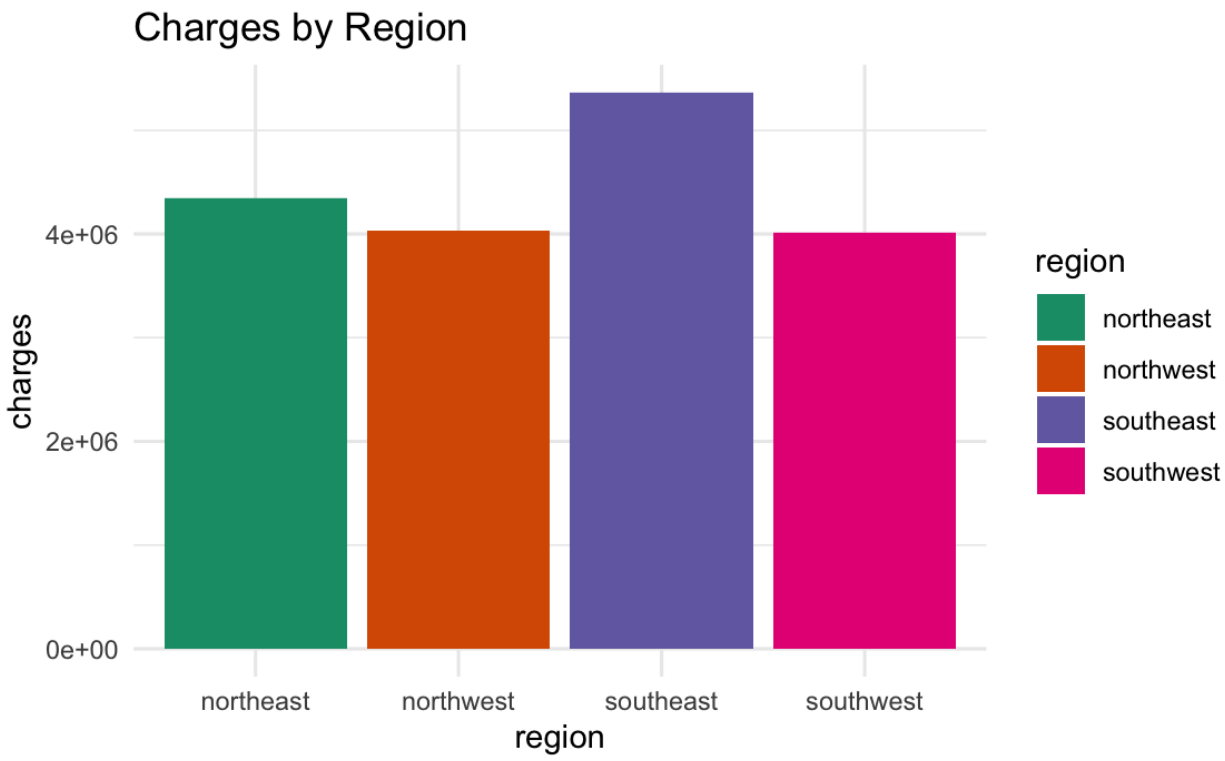


Figure 2.2 - Bar Plots of Charges Separated by Regions

We decided to further look at other categorical variables along with the region to see if we could find why the southeast has the highest charges. We decided to first look at smoking.

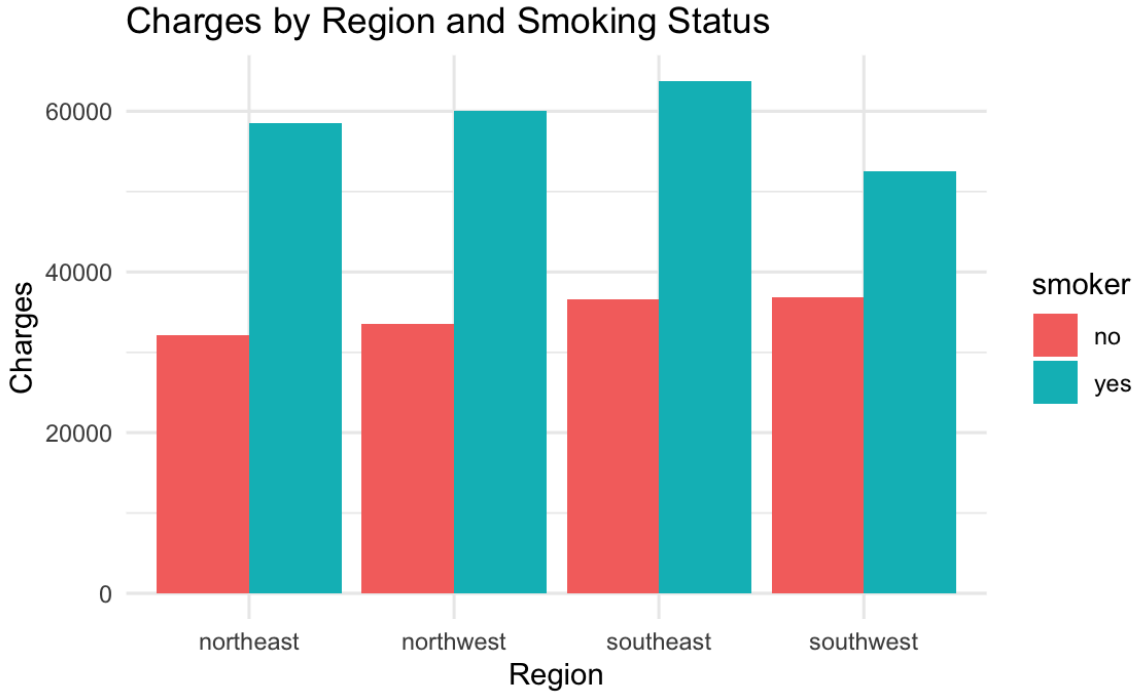


Figure 2.3 - Bar Plots of Charges by Smoking Status and Separated by Regions

Costs for smokers are nearly double that of the nonsmokers in every region. We believe this to be significant as smokers only make up 20% of the sample population, so we will keep an eye on smoking status when we build our model. Here the southeast region still has the highest charges for smokers reaching above

The next variable we looked at was the number of children. We turned the number of children into a categorical variable here to look at how having different amounts of children would influence insurance costs.

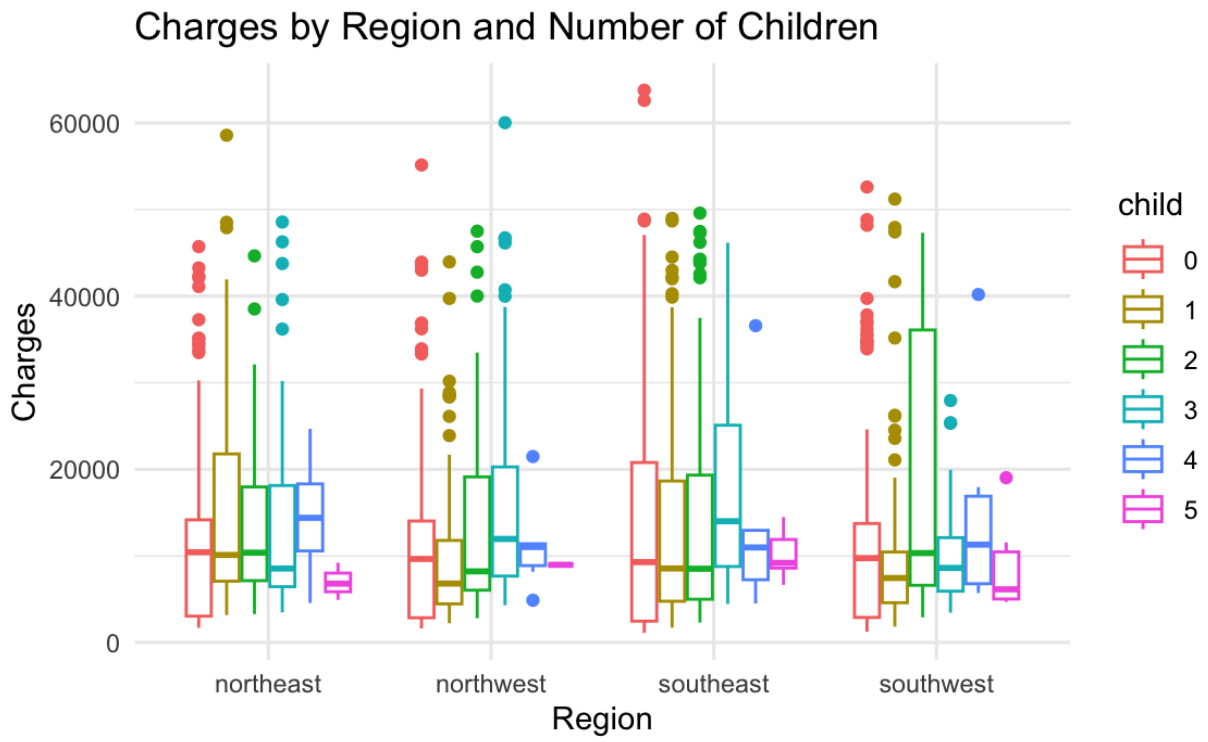


Figure 2.4 - Boxplots of Charges by Number of Children and Separated by Region

When looking at the interquartile ranges of charges across all counts of children, we see no significant degree of variation. We do see that most of them do lie under $20,000. The outliers and ranges are spread out and have no obvious pattern to them. For observations with 5 children, they usually have small ranges and these ranges lie under $10,000. However, we believe this to be because there are very few of them, so we cannot make any assumptions about them with confidence. When looking at the regions, we see the southeast region has generally higher ranges than the others, but no major differences.

To follow the number of children, we decided to look at BMI which is a continuous numerical variable, but we turned it into a categorical one to look at charges across the various BMI statuses: underweight, normal, overweight, and obese.

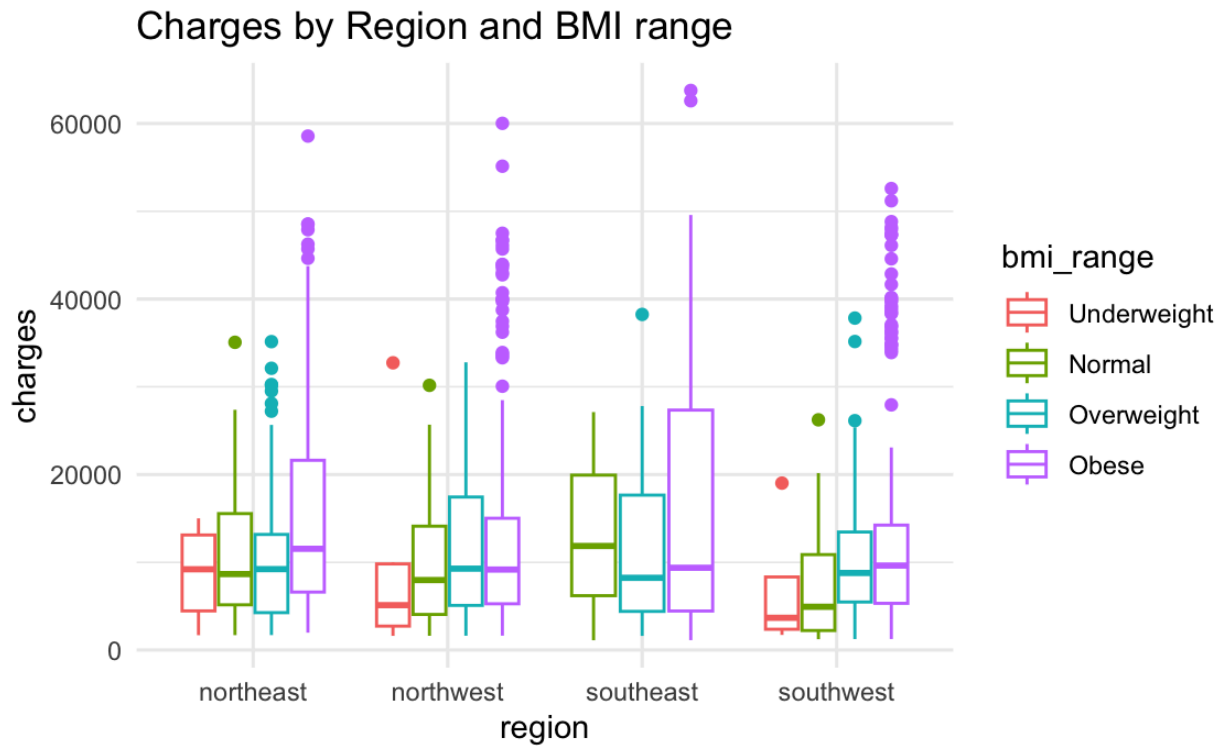


Figure 2.5 - Boxplot of Charges by BMI Status and Separated by Region

Our first observation when looking at the boxplots was that the obese weight status had the highest ranges and outliers. Across all regions, the obese status contains points that have the highest charges. The other statuses do not reach above $40,000, whereas the obese statuses have many points in the $40,000 to $50,000 range peaking at even above $60,000. When looking at the southeast region, we see that they have the highest range among all regions. Furthermore, they have the outliers with the highest charges in the entire dataset.

Of the categorical variables, the smoking status had the most significance, so we decided to keep an eye on this when we looked at the numerical variables age and BMI.

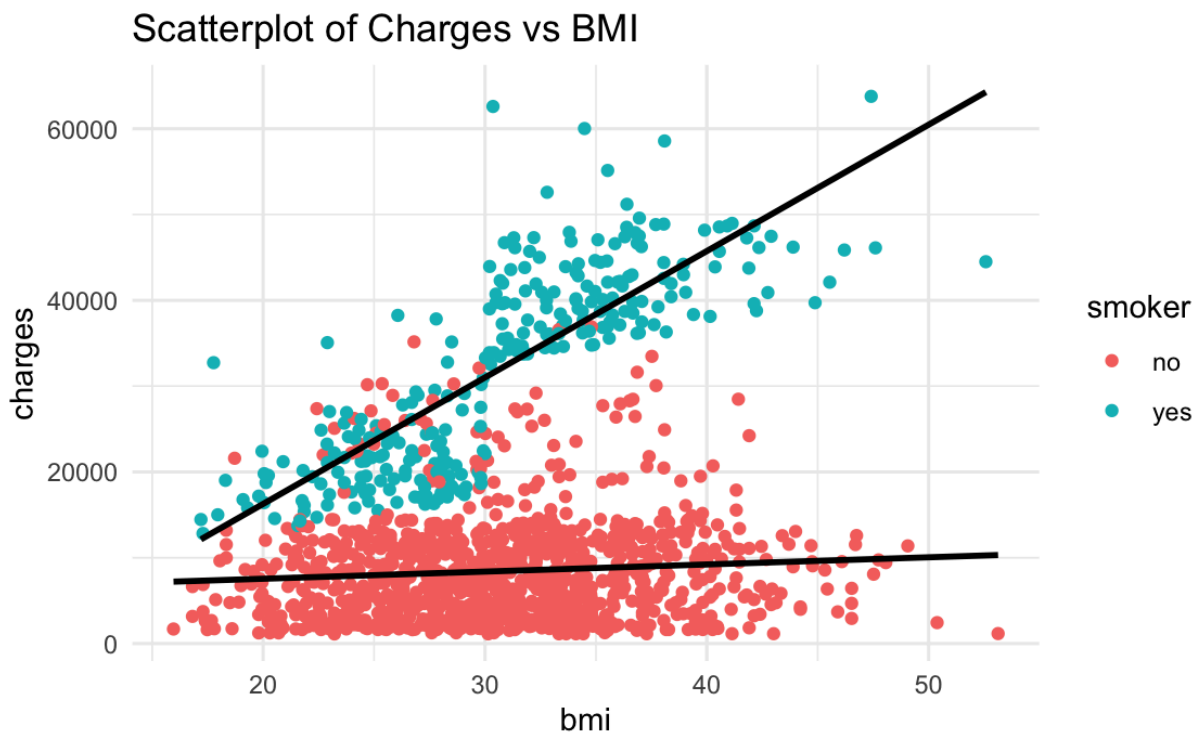
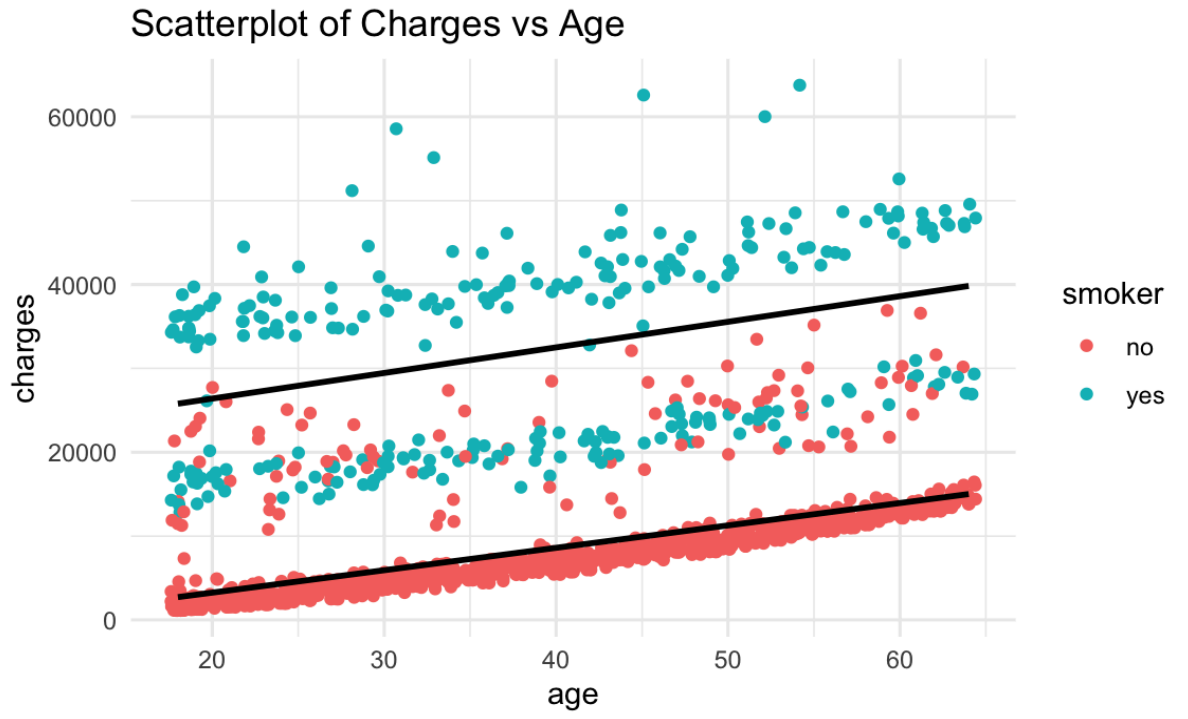


Figure 2.6 - Scatterplot of Charges vs BMI Separated by Smoking Status

When we look at the data points that are smokers only, we see a weak, but steep positive linear trend with the BMI. For non-smoker points, there appears to be minimal correlation between BMI and charges as points appear to be very spread out.

Figure 2.7 - Scatterplot of Charges vs Age Separated by Smoking Status



Overall, for charges vs age, there is a clear positive trend. There does also appear to be a weak linear trend. However, when looking at only non-smokers, there is a more definite positive linear trend between age and medical insurance. Furthermore, it can be argued that there appear to be multiple linear trends in this scatterplot: one clear trend from about $1,000 at 20 years old to $15,000 at 65 years old of non-smokers, a weaker one from about $15,000 at 20 years old to $30,000 at 65 years old filled with both smokers and non-smokers, and another weak one from about $35,000 at 20 years old to $50,000 at 65 years old of only smokers. From this, we believe age will be another significant factor in our model for medical insurance costs.

3 Variable Selection and Model Building

Based on the results of our exploratory analysis, the regressors that showed the most influence on charges were smoking status, age, and BMI, so these were very likely going to be in the model. We decided to start with a full model of the lm function.

```
Call:
lm(formula = charges ~ age + bmi + sex + children + smoker +
    region, data = medic)

Residuals:
     Min       1Q   Median       3Q      Max
-11305.1  -2850.3   -979.9   1395.0  29992.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11936.56     988.23 -12.079  < 2e-16 ***
age                256.76      11.91  21.555  < 2e-16 ***
bmi                339.25      28.61  11.857  < 2e-16 ***
sexmale           -129.48     333.20  -0.389 0.697630
children           474.82     137.90   3.443 0.000593 ***
smokeryes        23847.33     413.35  57.693  < 2e-16 ***
regionnorthwest   -349.23     476.82  -0.732 0.464053
regionsoutheast  -1035.27     478.87  -2.162 0.030804 *
regionsouthwest   -960.08     478.11  -2.008 0.044836 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6064 on 1328 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7492
F-statistic:   500 on 8 and 1328 DF,  p-value: < 2.2e-16
```

Figure 3.1 - Full Model Summary

Our initial observation was that we had an adjusted r-squared of 0.7492 which is reasonably good. However, we believe it can be better. From the full model, we see that age, BMI, children, and smoker

are all statistically significant, as their p-values are all well under 0.05. To continue, we decided to use the stepAIC function in R to help reduce our model. We used direction = “backward”. This would end up removing sex from our model.

```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region,
    data = medic)

Residuals:
     Min       1Q   Median       3Q      Max
-11366.5  -2841.4   -976.9   1364.0  29936.4

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11987.42     979.21 -12.242  < 2e-16 ***
age                256.88      11.91  21.577  < 2e-16 ***
bmi                338.73      28.57  11.856  < 2e-16 ***
children           473.86     137.83   3.438 0.000604 ***
smokeryes        23835.21     412.04  57.847  < 2e-16 ***
regionnorthwest   -348.25     476.66  -0.731 0.465152
regionsoutheast  -1034.63     478.71  -2.161 0.030852 *
regionsouthwest   -959.42     477.95  -2.007 0.044914 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7494
F-statistic: 571.8 on 7 and 1329 DF,  p-value: < 2.2e-16
```

Figure 3.2 - Reduced Model Summary After Applying stepAIC(direction = “backward”)

After using stepAIC to reduce our model, the sex variable is removed while the other variables remain in the model. Our adjusted r-squared does improve by 0.0002 which is not major; however, it is still better.

```
Analysis of Variance Table

Model 1: charges ~ age + bmi + children + smoker + region
Model 2: charges ~ age + sex + bmi + children + smoker + region
  Res.Df        RSS Df Sum of Sq     F Pr(>F)
1   1329 4.8844e+10
2   1328 4.8838e+10  1   5553651 0.151 0.6976
```

Figure 3.3 - Anova Table comparing our full and reduced models

When comparing the two models using ANOVA, we get a p-value of 0.6976, which is over 0.05. Therefore, the null hypothesis is rejected, proving sex isn't a significant variable

```
> vif(fit)
                  GVIF Df GVIF^(1/(2*Df))
data1$age     1.016794  1        1.008362
data1$sex     1.008944  1        1.004462
data1$bmi     1.106742  1        1.052018
data1$children 1.004017  1        1.002006
data1$smoker  1.012100  1        1.006032
data1$region  1.099037  3        1.015864
```

Figure 3.4 - VIF(Variance Inflation Factor)

Our VIF threshold is 4.011, given by the equation $1/(1 - R^2)$ where R^2 is the multiple R-squared value from our reduced model. The output shows that all of our predictor variables have VIFs that are very close to 1, which is notably lower than our VIF threshold. This indicates an extremely low level of multicollinearity between our variables.

4 Residual Analysis

4.1 Transformation

After we removed the sex variable, we performed a residual analysis on the rest of the data. The first step in our analysis was to check if our data would benefit from a transformation. We tried various methods and we settled on a log transformation of our response variable, charges. In Figure 4.1 we can see that after applying the log transformation, our charges resembled a normal distribution.

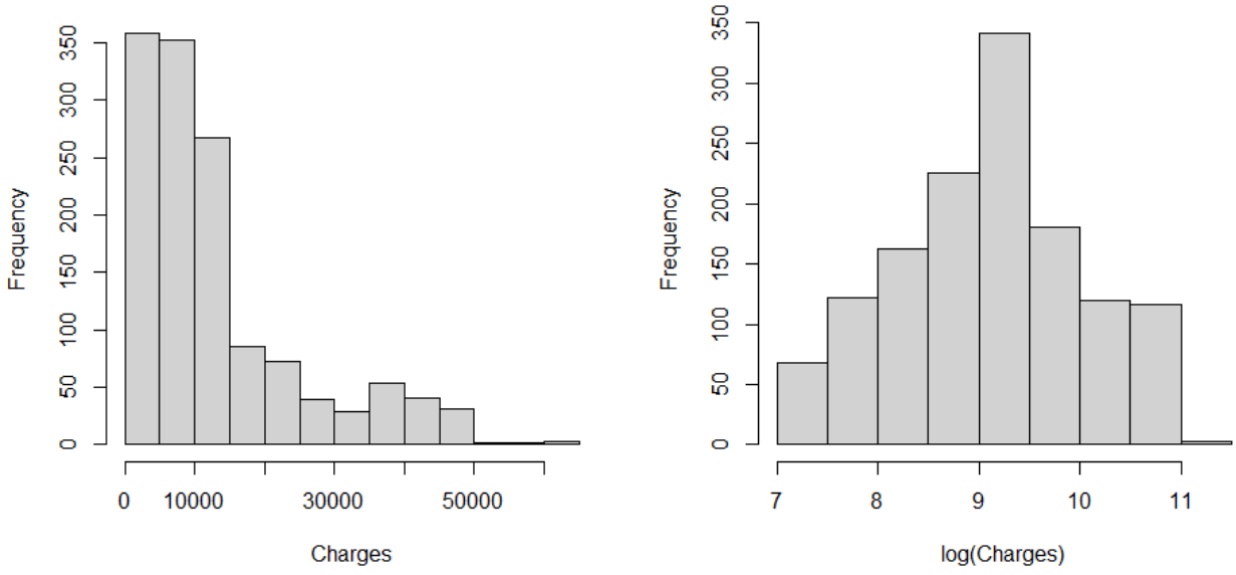


Figure 4.1.1 - Histogram of the "Charges" response variable before and after applying a log transformation

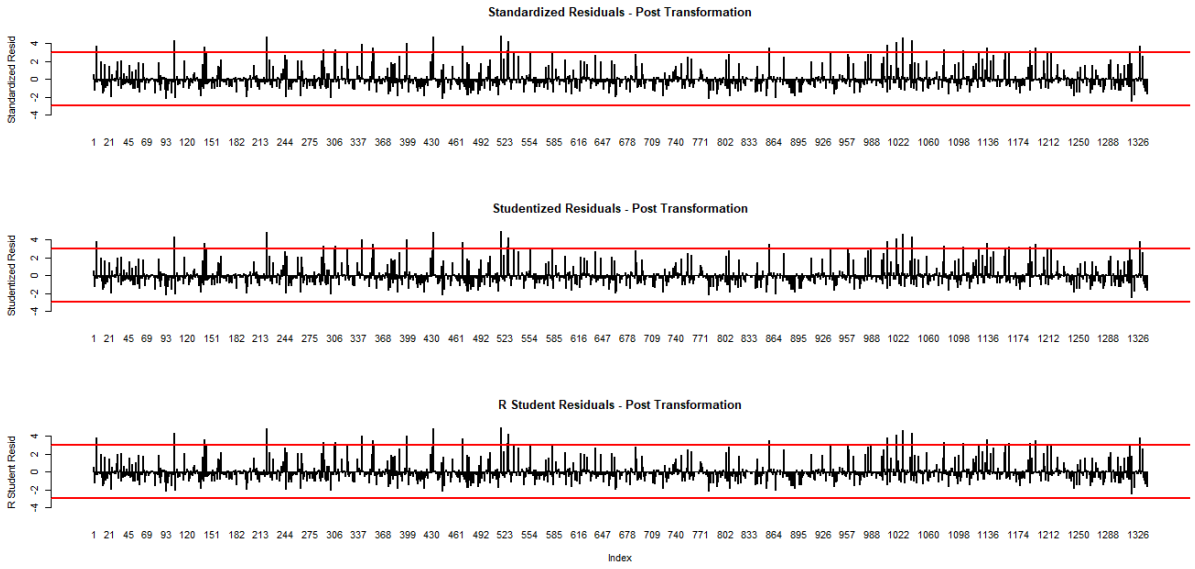
We know that this transformation was effective because our AICs and BICs decreased significantly. Pre-Transformation we had an AIC of about 27096 and a BIC of about 27148. After the transformation, we have an AIC of about -594 and a BIC of about -542.

4.2 Residuals

Figure 4.2.1 - The Standardized, Studentized, and R-Student Residuals for each observation

In Figure 4.2.1, from top to bottom, we have the Standardized residuals, Studentized residuals, and R-Student residuals. At first glance we see that the graphs look very similar to each other, suggesting that the model is well-behaved. Although there are a few observations that exceed the red cutoff line, we decided that they were not severe enough to be considered outliers in this case.

Figure 4.2.2 - The Standardized, Studentized, and R-Student Residuals for each observation after a Log Transformation



Looking at the same graph post-transformation, a few of the outliers from before now sit below the red, but at the same time, we have some different observations that could be potential outliers.

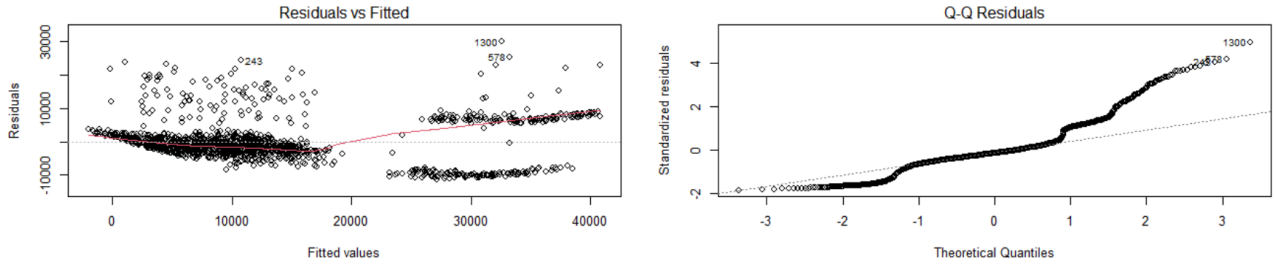


Figure 4.2.3 - Residuals vs Fitted Values on the left, QQ Plot on the right

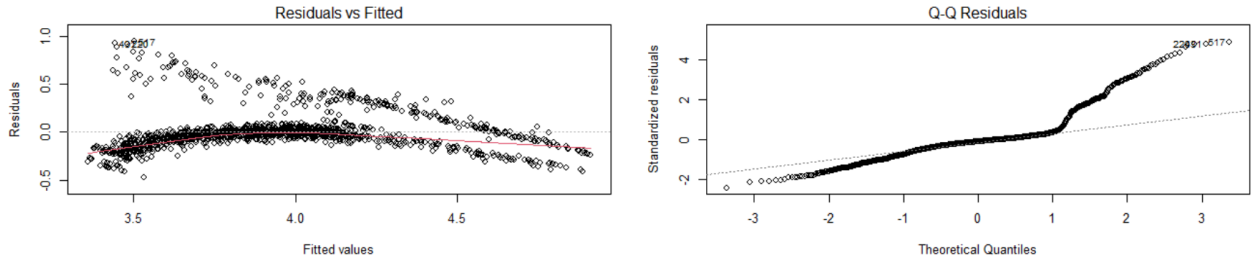


Figure 4.2.4 Residuals vs Fitted Values on the left, QQ Plot on the right after Log Transformation

Before any transformations, our residuals vs fitted values in Figure 4.2.3 look far from what we want. Ideally, we would have all the observations grouped and lined up around the red line in the center, but in our case, we have three separate clusters that don't seem to follow one specific trend. In situations like these, we know we have to transform the data. The same goes for our QQ plot. Ideally, it would follow a positive linear trend, but we once again need to transform the data.

After applying our log transformation in Figure 4.2.4 we managed to get down to two distinct groups and this time there is a downward trend. Our QQ plot also slightly improved, as we corrected the first major curve in the line. Although both our plots after transformation are still not ideal, we seek to further improve this in the reflection section of this report.

5 Influential Analysis

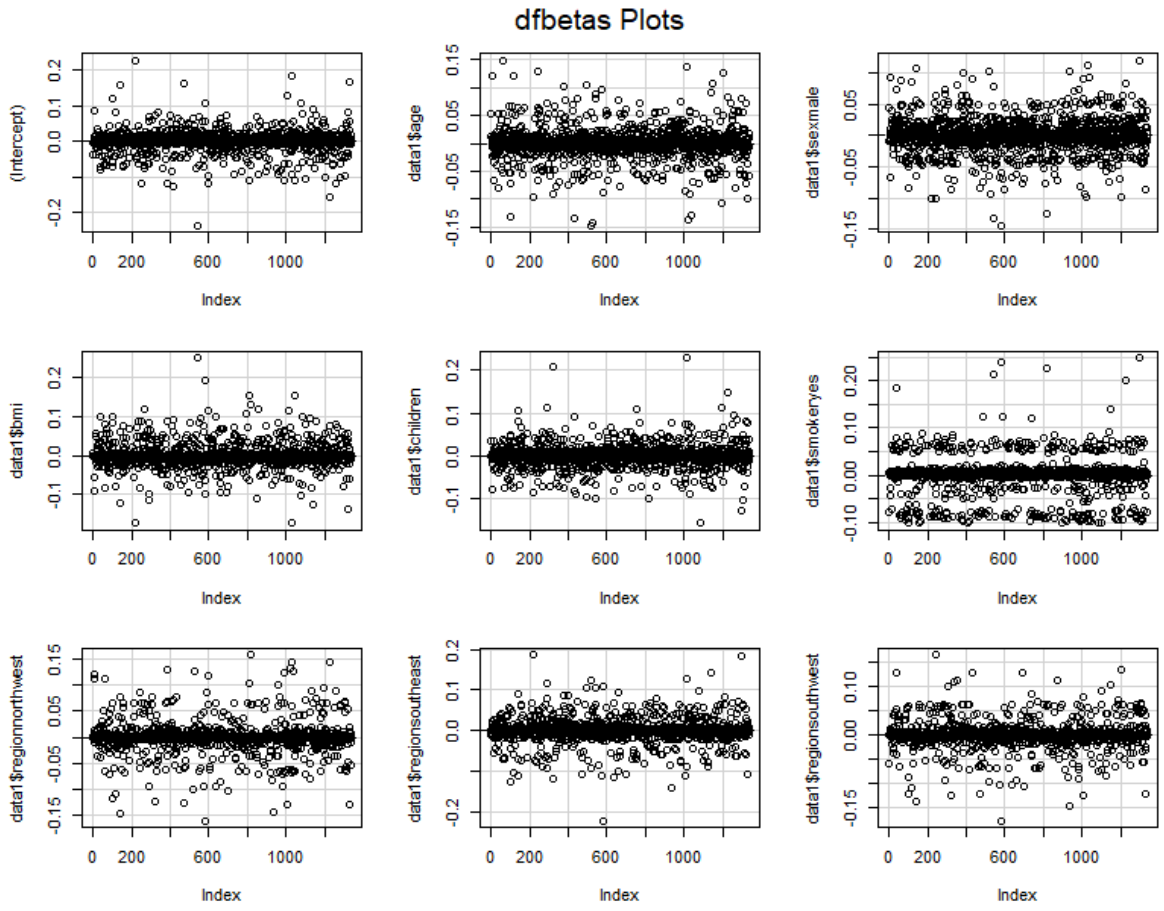


Figure 5.1 DFBetas Plots for Intercept, Age, Sex(Male), BMI, Children, Smokers, and the Northwest, Southeast, and Southwest regions

A pattern that repeats in most of our graphs is the concentration of observations. This tells us that the majority of points are of comparable influences. This is the case for most graphs in Figure 5.1, except for the one corresponding to the "smoker" variable. This discrepancy suggests a significant distinction in influence based on whether the user is categorized as a smoker or non-smoker.

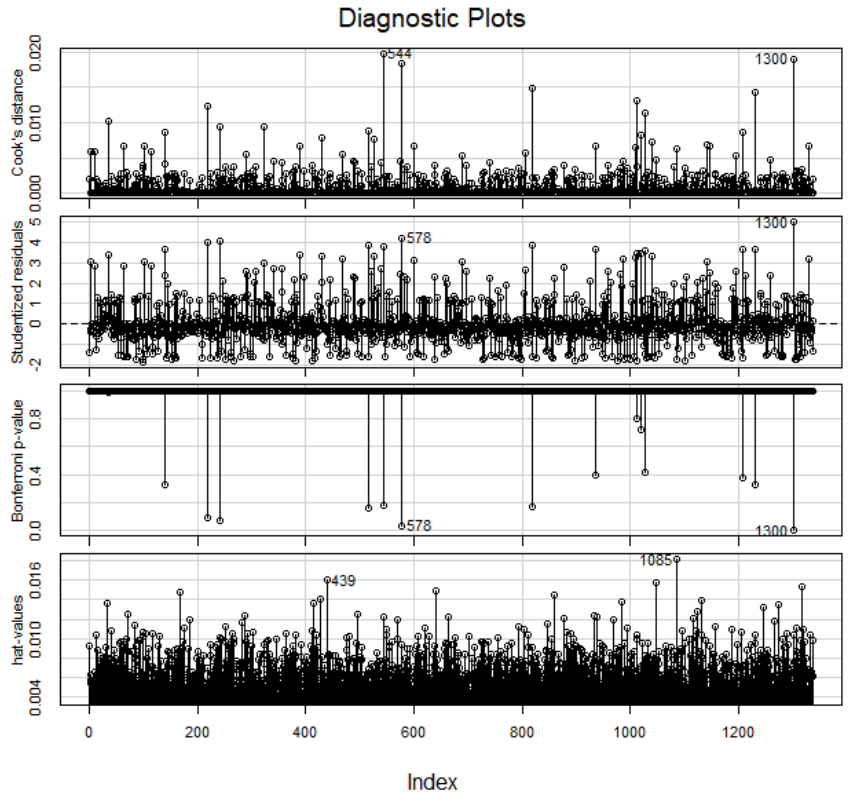


Figure 5.2 - Influence Index Plots with Cook's Distance, Studentized Residuals, Bonferroni p-values, and Hat-values

The first graph in Figure 5.2 is the Cook's Distance. From Cook's Distance, we can visualize which points are the most influential to our model. The most influential observations are 544, 578, and 1300. Observation #544 is 54 years old, has a BMI of 47.71, has no children, smokes, lives in the Southeast, and has a charge of $63,770.42. Observation #578 is 31 years old, has a BMI of 38.09, has one child, smokes, lives in the Northeast, and has a charge of $58,571.07. Observation #1300 is 45 years old, has a BMI of 30.36, has no children, smokes, lives in the Southeast, and has a charge of $62,592.87. If we keep in mind that a BMI of above 30 is considered obese, the common ground that these three points share is that they all smoke, are obese, and their cost is around $60,000. This observation hints that the primary factors that increase insurance costs are a high BMI and smoking.

The last graph in Figure 5.2 is the visualization of our model's hat-values. From this graph, we can see our potential leverage points and look further into them. It's important to identify leverage points because with them we can find what is influencing our regression coefficients the most. By this criteria, observations 439 and 1085 stand out the most on our graph. Observation #439 is 52 years old, has a BMI of 46.75, five children, doesn't smoke, and lives in the Southeast with a charge of $12,592.53.

This observation has a high likelihood of being a leverage point because although she has a high BMI, her charge isn't abnormally high. Observation #1085 is 39 years old, has a BMI of 18.3, five children, smokes, and lives in the Southwest and a charge of $19,023.26. Both of these observations have five children which is rare for our dataset which could be why they are so impactful.

6 Conclusion

After completing our analysis, we concluded that the most influential factors affecting the price of medical insurance are age, BMI, children, and smoker. Smoker was by far the most influential predictor of the actual insurance charges followed by children, BMI, and age. Our most significant predictors of the model were age, BMI, and smokers. Our full model looks like this:

Charges = -12098.82 + 257.77(age) + 321.87(bmi) + 472.98(children) + 23810.40(smoker)

We removed the "region" variable because while it does have significance to the model, we do not know if the region truly affects insurance charges for medical reasons. It could purely be because one region is more economically well off than another. Furthermore, R splits up the categorical regions variable into dummy variables. Overall the region variable is very difficult to interpret for our results.

7 Reflection

We believe this project did put our regression and analysis skills to the test. We are grateful to have been able to apply everything we have learned this semester. However, going back, we would have picked a different dataset with more numerical variables. When we first found our medical insurance dataset, we thought building a model off of it would be easy as the description said it is designed to train machine learning models. This was not entirely the case as we found a lot of difficulty in working with the limited variables, half categorical. Furthermore, halfway through our initial analysis, we discovered that more than half of the dataset were duplicates. Alternatively, the numerical variables were very easy to work with, and smoker had a clear influence on the model.

Initially, when we presented our reduced model the residual vs fitted plot and QQ plot appeared to have multiple different lines of best fit. We tried different transformations to fix this; however, no transformation was able to fully correct the residual plots. After our presentation, the professor advised us to test the interaction between a dummy variable with the covariate. We tried crossing BMI with smokers, age with smokers, and children with females. Of these three, only BMI with smokers and

children with females contributed to the model. We will note that when they were added to the model, the adjusted r-squared increased to 0.8399 which is much better than our original adjusted r-squared of 0.7494. Even though this was the case, when we did the residual and QQ plots, the outputs were identical. We recognize this as a limitation to our findings, but due to the insignificant change in these visualizations, we decided to include it in our reflection.

Charges = -1796.732 - 586.515(regionnorthwest) - 1144.896(regionsoutheast) - 1213.175(regionsouthwest) - 21123.721(smokeryes) + 264.131(age) + 352.551(children) + 1466.133(bmiSmoke) + 339.872(sexBirth)

For future work, we could look at more interactions between variables as there appeared to be multiple lines of best fit in the residuals plot, so we could find the other lines of best fit. Furthermore, once we had reached an end in our research due to the unfortunate residual plots, we attempted to find more analytical interpretations of this dataset. The only one that we were able to find that had appropriate plots utilized a polynomial regression model, demonstrating that we might not yet have all the resources necessary to analyze this data as best that we can.

Appendix

References:

Rahul Vyas. (2020). Medical Insurance Cost Prediction [Data file]. Retrieved from Kaggle datasets. URL: https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction/data

Responsibilities:

Zach Gebreab: Variable selection and Model fitting, Introduction, Conclusion

Kyle Keshmeshian: Residual Analysis, Influential Analysis, Transformation

John Moro Dutra: Influential Analysis, Conclusion, Variable Selection and Model Fitting, Data Cleaning

Kenneth Nguyen: Introduction, Exploratory Analysis, Reflection, Conclusion

Code:

```
library(ggplot2)

library(MASS)

library(dplyr)

library(car)

medic <- read.csv(file.choose())

medic <- unique(medic)

##Figure 2.1

##density plot of charges

ggplot(medic, aes(x = charges)) +

 geom_density(fill = "palegreen3") +

 labs(title = "Distribution of Charges",

     x = "Charges",
```

```
    y = "Density") +
  theme_minimal()
##Figure 2.2
##box/barplot region and charges
ggplot(medic, aes(x = region, y = charges, fill = region)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Dark2") +
  theme_minimal()
  ##Figure 2.3
  ##barplot of charges vs region and smoking status
ggplot(medic, aes(x = region, y = charges, fill = smoker)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Charges by Region and Smoking Status",
       x = "Region",
       y = "Charges") +
  theme_minimal()
##Figure 2.4
##boxplot of charges by region and number of children
medic <- medic %>%
  mutate(child = factor(children, levels = 0:5, labels = c("0", "1", "2", "3", "4", "5")))
ggplot(medic, aes(x = region, y = charges, col = child)) +
  geom_boxplot() +
  labs(title = "Charges by Region and Number of Children",
       x = "Region",
       y = "Charges") +
```

```
  theme_minimal()

##Figure 2.5

##boxplot of charges vs region and BMI

medic <- medic %>%

  mutate(bmi_range = cut(bmi, breaks = c(-Inf, 18.5, 24.9, 29.9, Inf), labels = c("Underweight",
"Normal", "Overweight", "Obese")))

ggplot(medic, aes(x = region, col = bmi_range, y = charges)) +

  geom_boxplot() +

  labs(title = "Charges by Region and BMI range") +

  theme_minimal()

##Figure 2.6

##scatterplot charges vs BMIi and smoker

ggplot(medic, aes(x = bmi, y = charges, color = smoker)) +

  geom_jitter() +

  geom_smooth(method = "lm", se = FALSE, aes(group = smoker), color = "black") +

  labs(title = "Scatterplot of Charges vs BMI") +

  theme_minimal()

##Figure 2.7

##scatterplot charges vs age and smoker

ggplot(medic, aes(x = age, y = charges, color = smoker)) +

  geom_jitter() +

  geom_smooth(method = "lm", se = FALSE, aes(group = smoker), color = "black") +

  labs(title = "Scatterplot of Charges vs Age") +

  theme_minimal()

##Figure 3.1

##Full model summary
```

```
fit <- lm(charges ~ age + bmi + sex + children + smoker + region, data = medic)
summary(fit)
##Figure 3.2
##Reduced Model Summary
fitM <- stepAIC(fit, direction = "backward")
summary(fitM)
##Figure 3.3
##Anova Table comparing our full and reduced models
anova(fitM, fit)
##Figure 3.4
##Variance Inflation Factor
vif(fit)
##Figure 4.1.1
hist(medic$charges, xlab = 'Charges')
hist(log(medic$charges), xlab = 'log(Charges)')
##Figure 4.2.1
range(stdres(fit))
barplot(height = stdres(fit), names.arg = 1:1337,
      main = "Standardized Residuals - Pre Transformation",
      ylab = "Standardized Resid", ylim = c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)
range(studres(fit))
barplot(height = studres(fit), names.arg = 1:1337,
      main = "Studentized Residuals - Pre Transformation",
```

```
    ylab = "Studentized Resid", ylim = c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)
range(rstudent(fit))
barplot(height = rstudent(fit), names.arg = 1:1337,
    main = "R Student Residuals - Pre Transformation", xlab = "Index",
    ylab = "R Student Resid", ylim = c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)
##Figure 4.2.2
fit_log <- lm(log(charges) ~ age + bmi + sex + children + smoker + region, data = medic)
range(stdres(fit_log))
barplot(height = stdres(fit_log), names.arg = 1:1337,
    main = "Standardized Residuals - Post Transformation",
    ylab = "Standardized Resid", ylim = c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)

range(studres(fit_log))
barplot(height = studres(fit_log), names.arg = 1:1337,
    main = "Studentized Residuals - Post Transformation",
    ylab = "Studentized Resid", ylim = c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)
```

```
range(rstudent(fit_log))
barplot(height = rstudent(fit_log), names.arg = 1:1337,
      main = "R Student Residuals - Post Transformation", xlab = "Index",
      ylab = "R Student Resid", ylim = c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)
##Figure 4.2.3
plot(fit, which = 1:2)
##Figure 4.2.4
plot(fit_log, which = 1:2)
##Figure 5.1
dfbetasPlots(fit, intercept = T)
##Figure 5.2
influenceIndexPlot(fit)
##Interaction between age and smoker, bmi and smoker, female and number of children
medic$ageSmoke <- medic$age * as.numeric(medic$smoker == "yes")
medic$bmiSmoke <- medic$bmi * as.numeric(medic$smoker == "yes")
medic$sexBirth <- medic$children * as.numeric(medic$sex == "female")
newFit <- lm(charges ~ region + smoker + age + children + bmiSmoke + sexBirth, data = medic)
summary(newFit)
```