United States Health Insurance Analysis

Amber D. Martinez

May 13, 2021

University of Wyoming, Honor's College

Senior Capstone Project

Abstract

This project uses a simulated dataset from a machine-learning textbook whose goal was to accurately predict total annual medical costs submitted to health insurance companies in the U.S. The effects of smoking and a person's Body Mass Index (BMI) were the focuses of the analysis but the study also included the effect of age, children being included on the insurance plan, and the region the policyholder lived in. Insurance companies calculate premiums with age, location, tobacco use, individual vs family enrollment, and the amount of coverage a person chooses to purchase. This study is important because health care is expensive in America and far too many people go without basic health care simply because they cannot afford it. There are bigger reasons why this is the case, but they are beyond this study. Using graphical analysis, various types of statistical analysis using the programming software R, and model building, this study will show that age, smoking, a BMI greater than 30, and having children significantly increases the total amount that people are paying in health care annually. Local and state resources for smokers and obese people will also be discussed after conclusions are drawn.

Keywords: total medical costs billed to insurance will be referred to as "charges" and is the response (y) variable. The predictor (x) variables are age, sex, BMI, smoking, region, and children.

Introduction

Being able to receive health care at a reasonable price is important for many people in the United States. Those who have health insurance have some control over their total cost by choosing which insurance company to insure through, but those without health insurance pay out of pocket for all medical bills and often forego necessary medical visits due to the cost. The purpose of this analysis is to show others that they can directly influence their own personal medical costs simply by living a healthier lifestyle. Understanding this idea can help people make smarter decisions and ultimately save on health care cost. From a simulated dataset that represents actual information from American health insurance holders, this analysis will determine which of the six variables -- age, sex, children, smoking, BMI, and region -- have a significant influence on medical costs.

Healthcare.gov[1] lists factors that are taken into consideration when insurance companies calculate health care premiums for their customers. Since this analysis contains similar factors, it is expected that similar conclusions can be drawn. First, being a smoker will increase total annual health care costs. Second, a higher BMI also increases total annual health care costs. Third, older individuals pay more in health insurance because, as one gets older, additional visits to a doctor are needed to remain healthy. The analysis is done with graphs, summary statistics, linear modeling, and various hypothesis tests. Graphs are shown in the results section, and tables are listed under the appendix at the end of the paper.

[1] According to healthcare.gov, most Insurance Companies in the US take into account age, location, tobacco use, individual vs family enrollment, and the type of plan a person chooses when calculating insurance premiums. The website also says that gender and medical history cannot be factors.

Methods

The distribution of the response variable ‘charges’ is shown in Figure 1. The data are mostly contained closer to zero with 980 of the 1338 observations falling between $0 and $15,000. The remaining 358 observations are those with charges greater than $15,000. These characteristics show that this dataset is ‘right skewed’. The observations greater than $15,000 are important to keep in the data because it shows high variation in how much people are paying. Figure 2 shows the correlation and scatterplot matrices for the quantitative variables in the data (age, BMI, children, and charges). It is necessary to check for high (> 0.80) positive or negative values between the explanatory variables shown in the upper diagonal in Figure 2. Such a value would indicate multicollinearity, which means that the two variables in question have a similar effect on

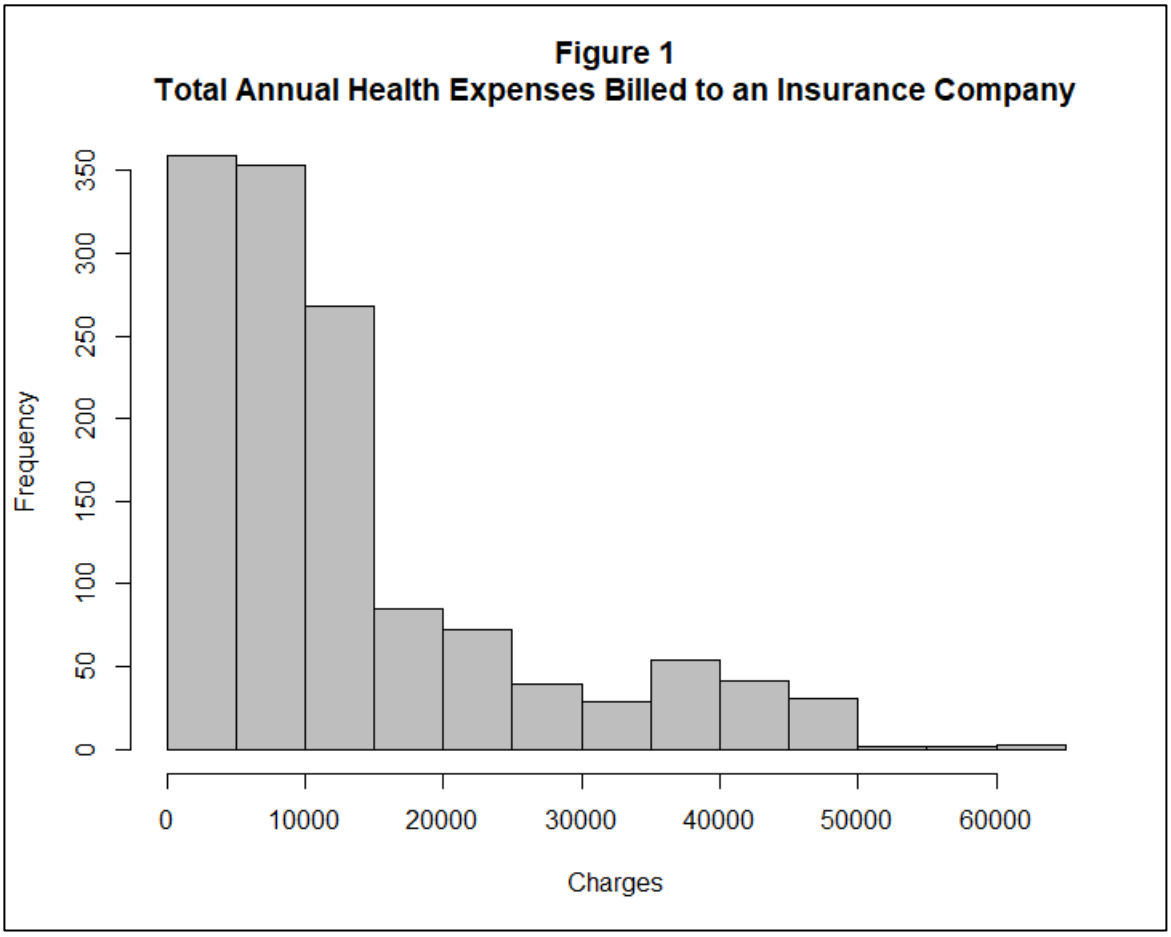
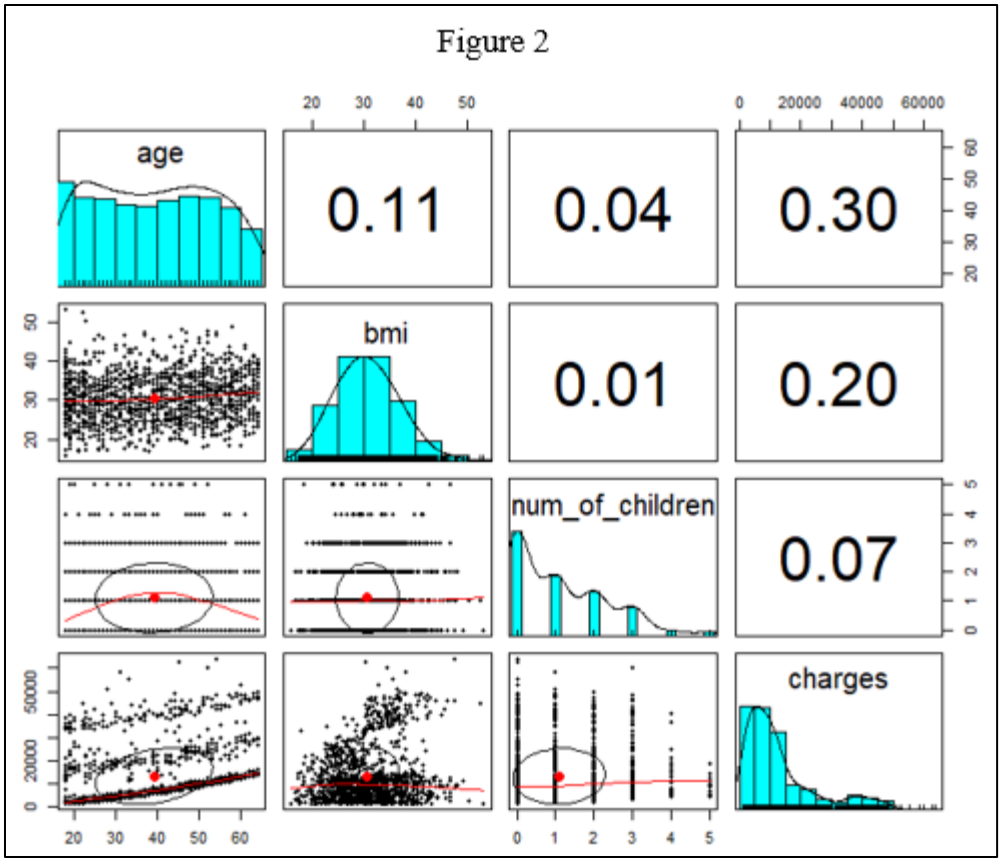


Figure 2



the response variable, thus creating redundancies when used together. The good news for this dataset is that none of the quantitative explanatory variables show high collinearity. All the explanatory variables are positively correlated with the response variable (shown in the furthest right column of Figure 2) indicating that each has some effect on charges, with age having the highest value of 0.30.

Graphical analysis and formal linear hypothesis testing were the main forms of study in this project. The graphs and figures were used to infer possible conclusions about the variable/s in question and were then confirmed or rejected through a formal linear hypothesis test/s. The first variable analyzed was smoking and its effect on charges. Figure 3 is a simple boxplot showing charges for those who smoke and those who do not. It is plausible based on the upper bounds of

each plot that there could be a difference in charges for smokers and non-smokers. A linear hypothesis test was then done to formally check if there is a difference between the total charges of each group. If the model used is as follows

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6}$$

Where y is 'charges', x_1 is age, x_2 is sex, x_3 is BMI, x_4 is smoking, x_5 is region, and x_6 is children, this hypothesis test checks if $\beta_4 = 0$ in the null hypothesis (interpreted as smokers and non-smokers not being different from each other in total charges).

The analysis also took into consideration the effect that smoking and age combined might have on charges. Figures 4 and 5 show the linear relationship that age and charges have where smokers and non-smokers have been identified using different colors. The three different levels seen in the age scatterplot are explained when separated into smokers and non-smokers as shown in Figure 5. An interaction term between age and smoking, $\beta_7 x_1 x_4$, was included in the model.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_1 x_4}$$

A goodness of fit test was done to check if the model with the interaction is a good fit or if the interaction term can be left out. The null hypothesis is that the model with the interaction term is a good fit and the alternative hypothesis is that the model without the interaction term is a good fit. If the p-value of the interaction term is less than $\alpha = 0.05$, then we reject the null hypothesis, and conclude that the model with the interaction term is not a good fit to the data and should be removed.

The second variable analyzed was BMI. In the scatterplot of BMI and charges, shown in Figure 6, there appear to be two groups of data points: one is between $0 and roughly, $15,000 and has equal scatter throughout the range of ages, the other is a linear-like coned shaped scatter

that meets in the center at 30 BMI. The two separate groups do not really add up when looked at this way. Figure 7 shows the same graph now separated into two colors by smokers and non-smokers, and then further split into different graphs.

To gain a greater understanding of the effect BMI and smoking have on total charges, Figure 8 was created. It shows BMI separated into the 5 groups that the World Health Organization has created for BMI[2] (WHO). A new variable called 'BMI30' was added to the model. This variable only includes data values for people who had a BMI greater than or equal to 30. It is shown in the model as $\beta_8 x_7$.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_1 x_4 + \beta_8 x_7 + \beta_9 x_4 x_7}$$

Several tests of significance were used to test the effects of BMI alone ($\beta_3 = 0$), BMI greater than 30 ($\beta_7 = 0$), and the interaction between BMI greater than 30 and smoking ($\beta_9 = 0$; the interaction term is $\beta_9 x_4 x_7$) on 'charges'.

Two other variables were analyzed briefly outside of the main analysis simply because of curiosity. It is commonplace inside the United States to have health care costs vary between regions. Boxplots were created for each region and summary statistics were created for each (shown in figure 9 and table 5 respectively). A Tukey HSD Test[3] was used between the four regions to detect significant differences between each pair of regions.

The other variable of interest was children. Having medical bills for another person is likely to add to medical costs, and children often need to be taken to doctors since their immune

[2] WHO's website scales body mass index (for adults over 20 years old) in the following way: < 18.5 = underweight; 18.5-24.9 = normal weight; 25.0-29.9 = pre-obesity; 30.0-34.9= obesity class I; 35.0-39.9 = obesity class II; > 40.0 = obesity class III.

[3] This test is used to determine if groups are different from each other. It works by comparing all the possible pairs and outputting a significance value for each pair.

systems are not fully developed. It is plausible that anyone with a child is going to have higher medical costs. Figure 10 shows boxplots for each number of children included on an insurance plan (from zero to five). From the graph it seems that the more children a person has leads to lower charges, however, this is likely caused by a lack of observations for people with 4 and 5 children (who are only represented by 43 of the 1,338 observations). To compensate for this, any person with a child was included in a new group called "has_children", which makes up for 764 of the 1,338 total observations. This new variable shows up as $\beta_{10}x_8$ in the following model

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_1 x_4 + \beta_8 x_7 + \beta_9 x_4 x_7 + \beta_{10} x_8}$$

After adding so many interaction terms and new variables to the model, it is necessary to now determine which combination results in the best fit model. Backwards elimination was used on the model, where starting with all terms, the least significant variable is removed until all the insignificant variables are gone and all the remaining ones are significant. This process results in a final model that can be used to calculate annual medical charges.

Results

In the analysis for the effect of smoking, Figure 3 indicates an apparent difference between the highest charges for a non-smoker ($36,910.61) and the highest for a smoker ($63,770.43). The difference is more than $20,000! At first glance, it does appear that smoking drastically increase charges. A formal hypothesis test (programming results shown in Table 1) checks whether total charges between a smoker and non-smoker are the same. If the significance value given is less than a predetermined level – in this case 0.05 – the conclusion will be that the two are significantly different. The programming output shows the significance values under the "Pr(>|t|)" column. The value for "smokeryes" is less than 0.00001 which is clearly less than 0.05 so it can be concluded that there is a significant difference in total charges between smokers and non-smokers.

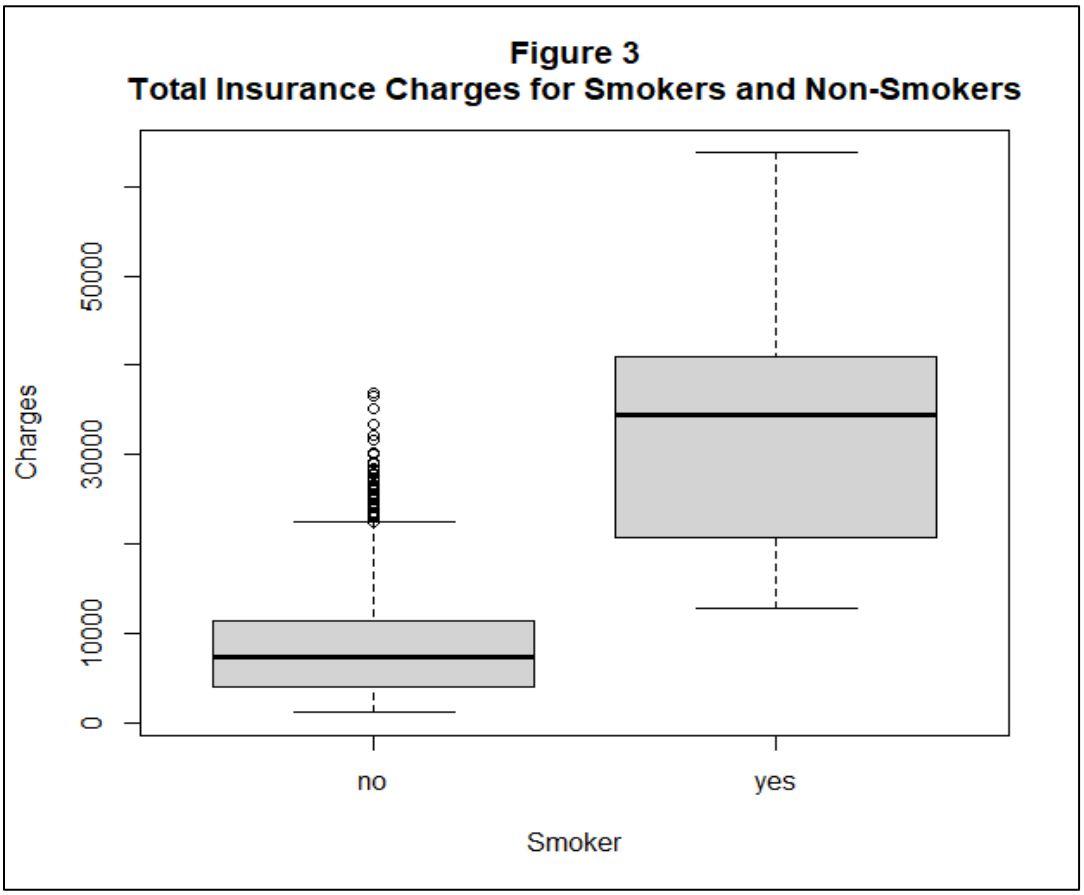
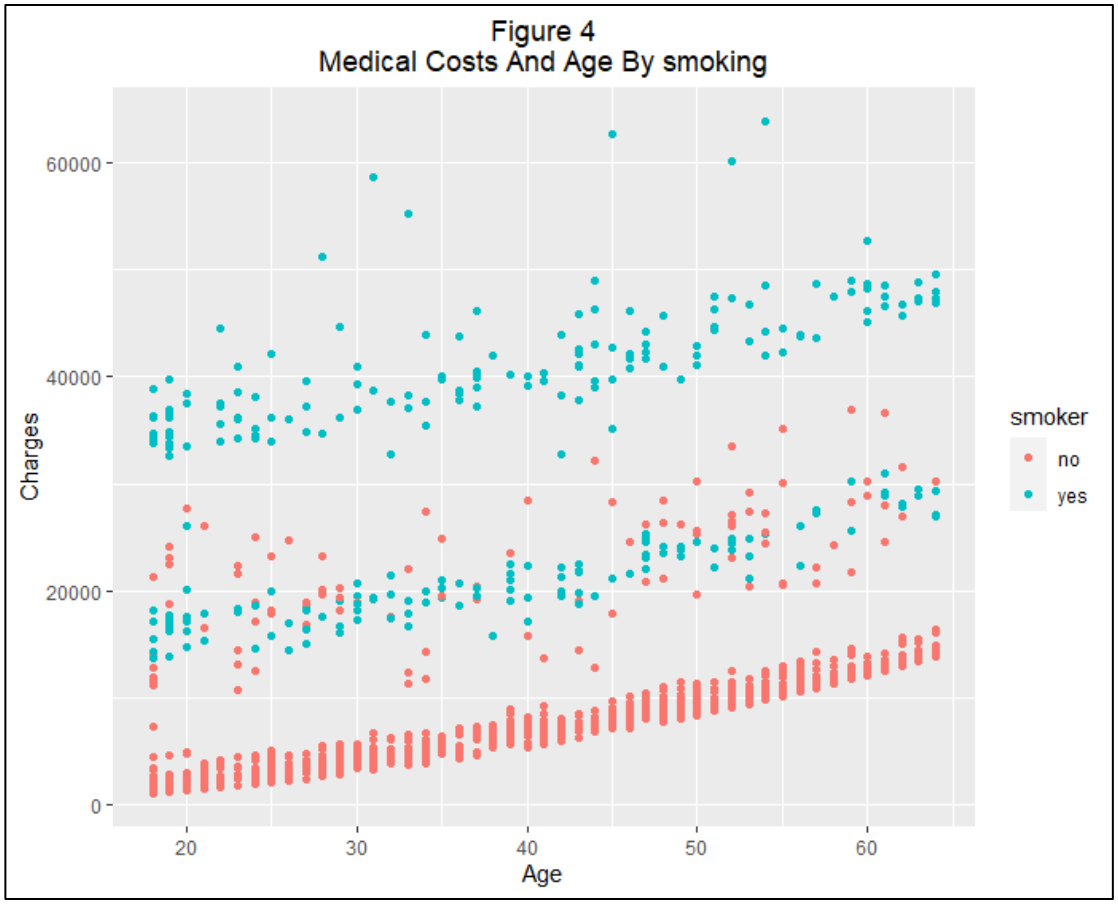
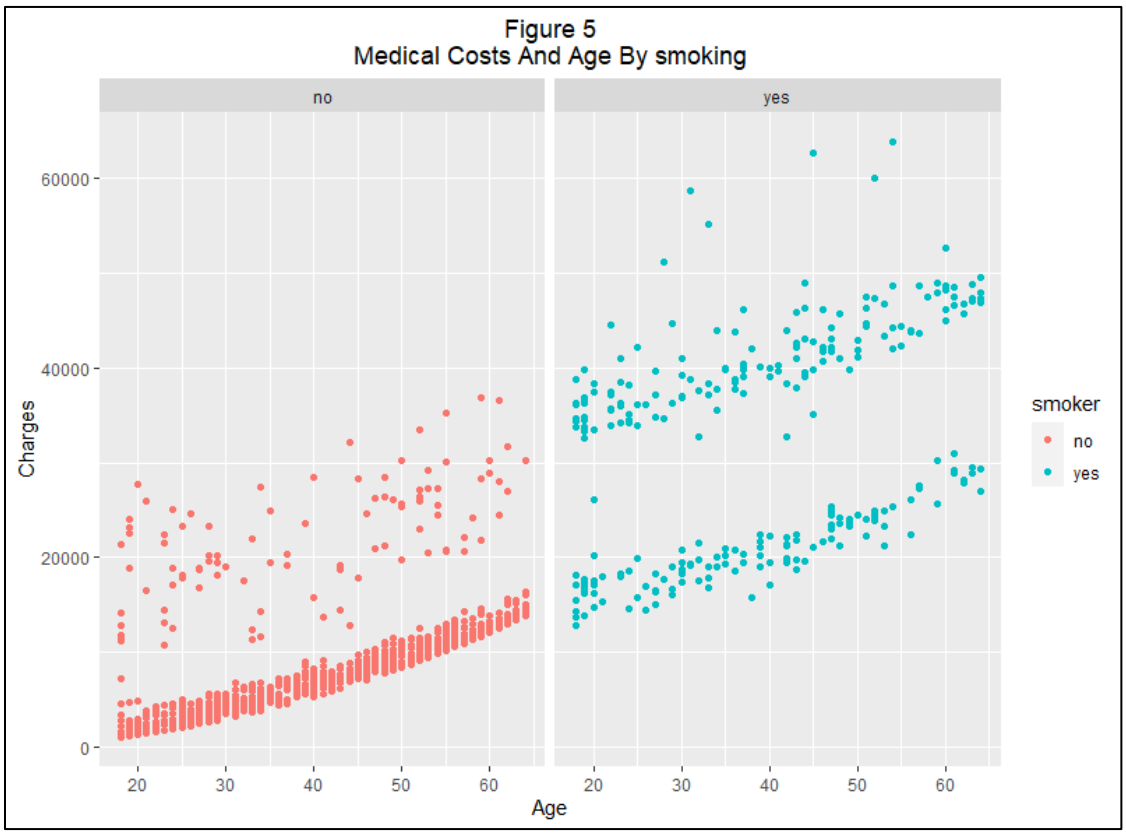


Figure 2 indicated that age and charges were positively linearly correlated with a value of 0.30, the three levels of 'charges' signify that there is something else influencing how age interacts with 'charges'. Figure 4 shows that the highest values in charges are those of smokers. Figure 5 shows that 20-year-old smokers are paying as much on average as 60+ year old non-smokers. Table 1 shows that age is a significant predictor of charges with a p-value of < 0.00001. Table 2 shows an interaction term between age and smoking and that it is not significant in predicting charges (p-value = 0.127 > 0.05). This shows that age and smoking by themselves are significant, but their interaction is not necessary to include in the model. From these analyses, we can conclude that smoking and age are significant predictors of increased charges.





BMI was analyzed next. Figure 6 does not do much in terms of interpretation. It shows two different groups as described in the methods section earlier but there is no clear interpretation. Figure 7 breaks the data into two groups by smoking and now there is room for interpretation. The graph of non-smokers shows that most of the data points are below $20,000 regardless of what the BMI is. The graph of smokers shows a slight increase in charges from 15 to 30 BMI, and that after 30 there are no observations less than $30,000 in charges. From these graphical observations, the data values for people with a BMI greater than 30 became of much greater interest.

Using the BMI values that WHO provides for weight status categories, Figure 8 shows two plots for smokers and non-smokers where the data have been broken into weight status. When

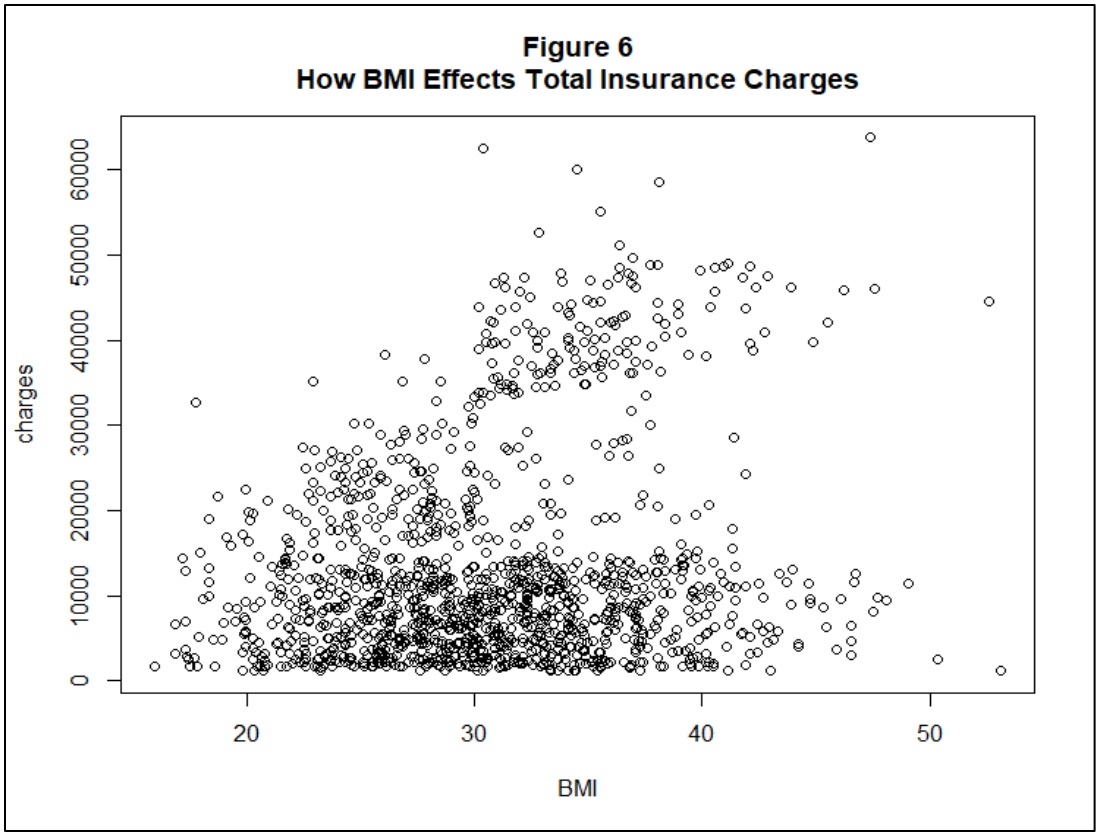


Figure 6
How BMI Effects Total Insurance Charges
charges
0
10000
20000
30000
40000
50000
60000
20
30
40
50
BMI

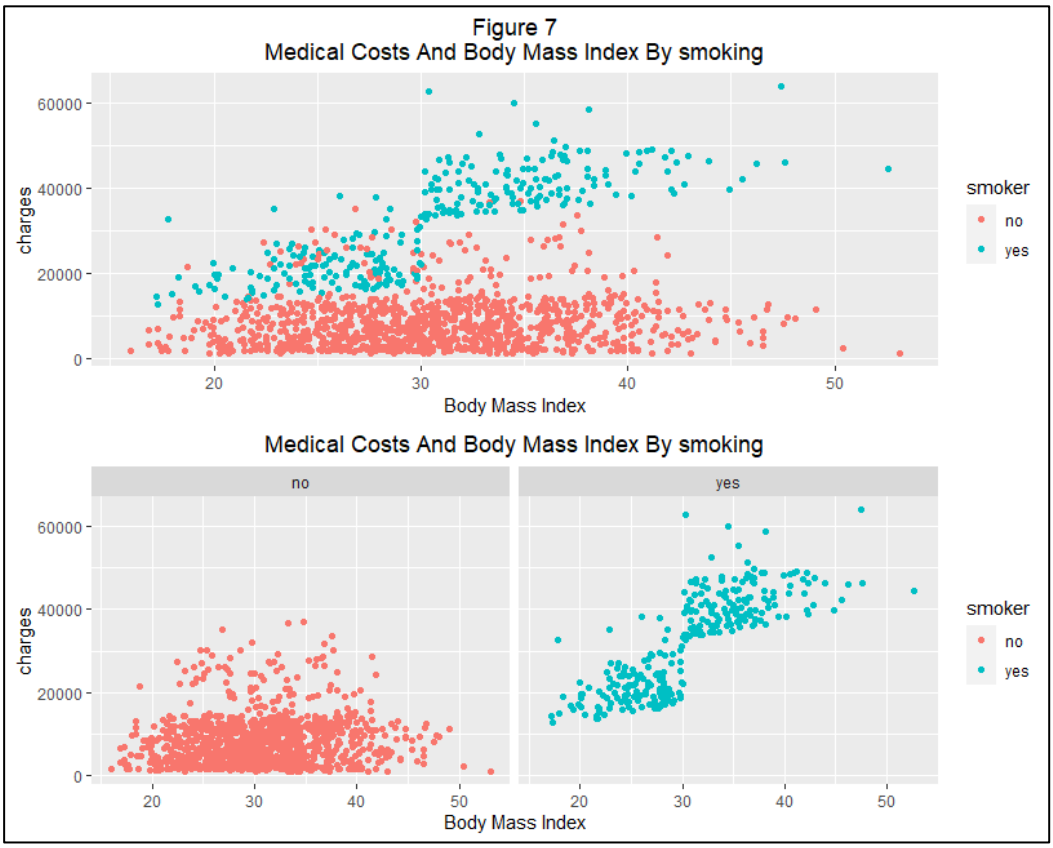
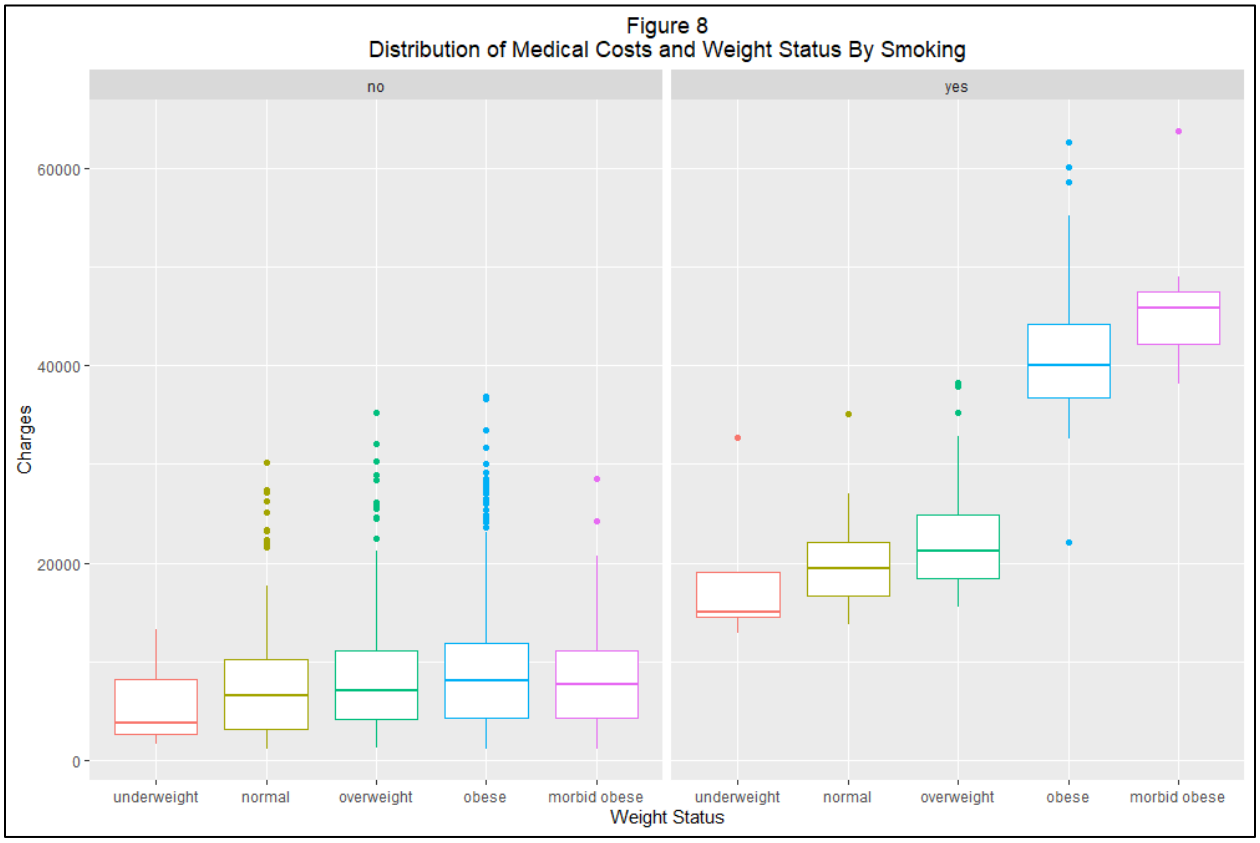


Figure 7
Medical Costs And Body Mass Index By smoking
charges
0
20000
40000
60000
20
30
40
50
Body Mass Index
smoker
no
yes
Medical Costs And Body Mass Index By smoking
no
yes
charges
0
20000
40000
60000
20
30
40
50
Body Mass Index
smoker
no
yes

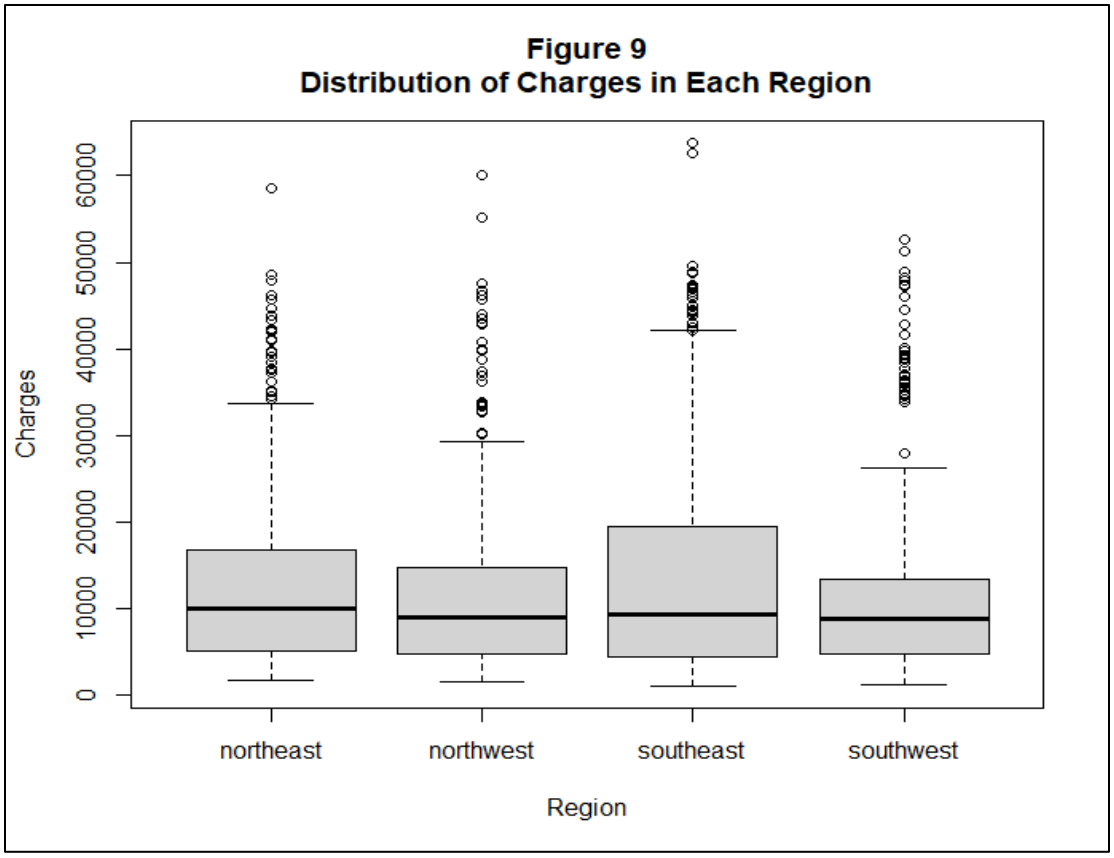
someone does not smoke, their charges increase slightly as their BMI increases (looking at the average values represented by the line inside the boxes). Smoking by itself increases total charges, but these graphs show that those charges increase even more when a smoker also has a BMI greater than or equal to 30. This is also shown in the estimates for smoker and smoker:bmi30 in Table 4. Smokeryes shows an estimated value of 13,402.363 and smokeryes:bmi30 shows an estimated value of 19,794.852. This indicates that smokers with a BMI greater than 30 pay about $6,000 more each year than smokers with a BMI less than 30.

When BMI, BMI30, and an interaction between BMI30 and smoking are included in a linear model and then tested for significance, the results (shown in Table 4) show that BMI and the interaction between BMI30 and smoking are both significant predictors of charges (significance level < 0.05), which is to be expected. It also shows that the interaction between



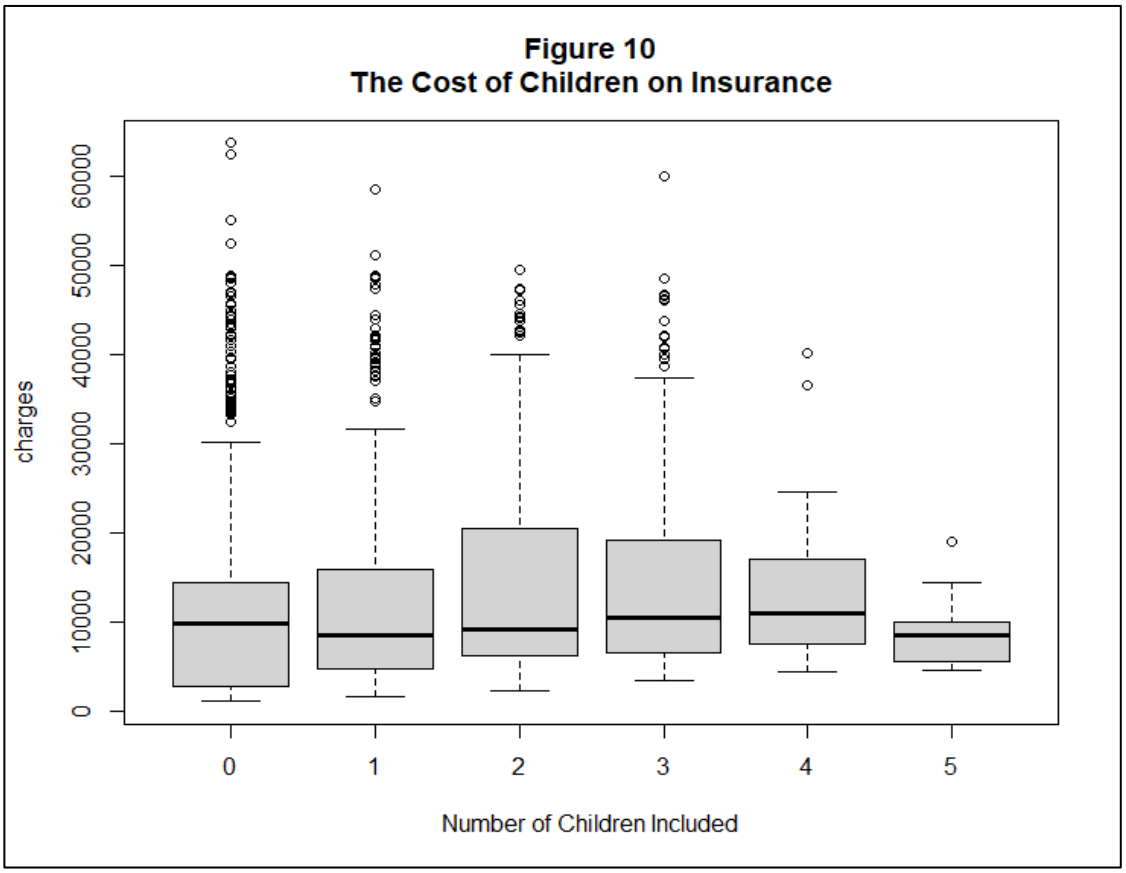
BMI30 and smoking is much more significant (p-value < 0.0001) in increasing total charges than BMI (p-value = 0.0009) or BMI30 (p-value = 0.04) by themselves[4].

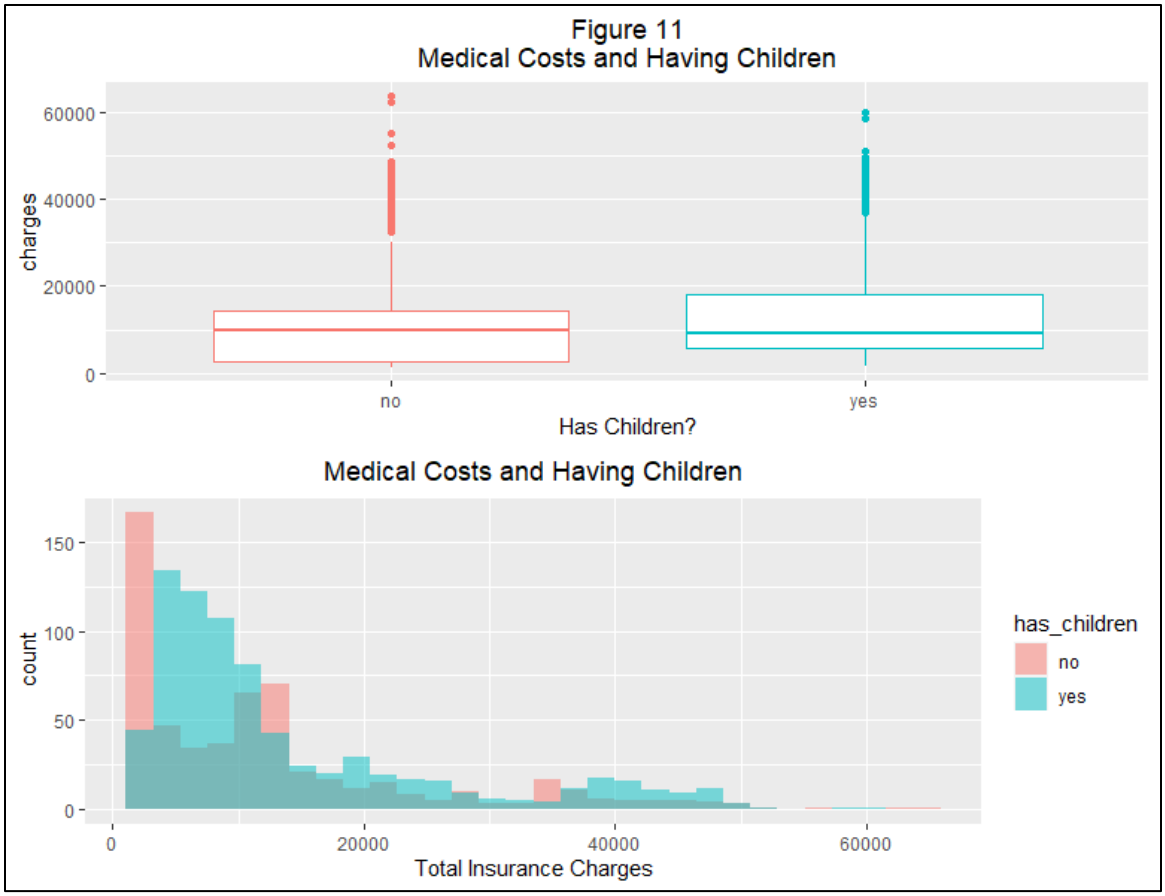
Moving to the analysis of region, Figure 9 shows the boxplots of each region and Table 6 shows the results from a Tukey HSD test between all possible pairs of the 4 regions. From the graph, it appears that there may not be a difference between each region since many of the averages are close to each other. If the 3rd quartiles are studied, it is possible that the southeast region could be different than the others because it is much higher than the 3rd quartile for the other regions. The Tukey HSD test shows that the southwest and southeast regions are different from each other. Looking back in the summary statistics, those two regions differ by about $5,500 in the 3rd quartile and about $11,000 in the max.



[4] Obesity can lead to cardiovascular issues and diseases if not treated quickly, and smoking directly weakens the lungs, so it is logical that the two combined would lead to increased medical issues that need treatment.

'Children' was analyzed next in a similar fashion as the other variables. Figure 10 shows boxplots of each number of children ranging from zero to five. At first glance it looks like more children leads to lower medical charges. Plotting the new variable 'has_children' next to people without children in figure 11, the mean charges appear to be approximately the same for each group and the 3rd quartile for has_children is higher than that of no children. The histogram shows that a good chunk of the people who do not have children are contained within the first bar (approximately $1,000 to $3,000) and that there are more people with children who are paying $3,000 to $12,000 than those without children. In a test between those with children and those without children, shown in table 7, 'has_children' resulted in a significance value of < 0.00001. It can be concluded that there is a significant difference in charges between people with children





and people without children. And based on the graphs in Figure 11, it is likely that people with children are paying more in 'charges'.

Through the several forms of analysis used in this project, the model used now contains ten predictors. Knowing that some of them are not significant predictors of charges, backwards elimination was used to delete terms from the model until all predictors left over were significant. The final model with estimated values used to predict medical charges is

$$\begin{aligned} \boldsymbol{charges} = & -\mathbf{4879.388} + \mathbf{263.808}\boldsymbol{age} - \mathbf{488.092}\boldsymbol{sexmale} + \mathbf{98.637}\boldsymbol{bmi} \\ & + \mathbf{13431.633}\boldsymbol{smokeryes} + \mathbf{515.972}\boldsymbol{children} \\ & + \mathbf{18929.499}\boldsymbol{smokeryes}\mathbf{:}\boldsymbol{bmi}\mathbf{30} \end{aligned}$$

The Adjusted R^2 value found in table 8 also shows that this model explains 86.17% of the variation in 'charges' which is a very good value. When using this equation to calculate medical charges for someone, refer to the following table for numerical equivalence of categorical variables:

'sexmale'	replace with '1' if person is male	'0' if female
'smokeryes'	replace with '1' if person smokes	'0' if non-smoker
'smokeryes:bmi30'	replace with '1' if person smokes and has bmi ≥ 30	'0' otherwise

If this equation were calculated using a 22-year-old female with a BMI of 28, who does not smoke and does not have any children, with the table above being used, her charges would be:

$$\boldsymbol{charges} = \mathbf{-4879.388 + 263.808(22) - 488.092(0) + 98.637(28) +}$$

$$\mathbf{13431.633(0) + 515.972(0) + 18929.499(0) \approx \$3686.22}\ \textbf{annually.}$$

Through an exploratory analysis and many hypothesis tests, it can be concluded that out of the 6 explanatory variables that are included in this study, smoking, age, BMI, and children are all significant predictors of charges. Smoking is the most significant predictor, with the interaction of BMI30 and smoking, and BMI following closely behind. Charges increase linearly as age increases and each child included adds $515.97 on average to charges. It was also concluded that the region someone lives in does not significantly change the amount of money a person spends on medical costs.

Discussion

This analysis found that smoking and a high BMI are two of the largest drivers of increased health care cost. It is imperative to discuss the resources both locally and statewide that are offered to people who fall into these categories. Wyoming has a statewide tobacco cessation program called *Wyoming Quit Tobacco* (WQT) that is low cost and has a representative in many of the more populated counties. It offers personalized quit plans, coaching support, free nicotine replacement therapies (patches, gum, etc.), and free or low-cost medications that can help people quit smoking. This website also features a hotline for people to call as well as a contact list for all the representatives in Wyoming. At a local scale, the Cheyenne Regional Medical Center has an informational section on their website[5] that provides people with basic facts, symptoms that may occur from withdrawal, and several management options for people who want to stop smoking. This website provides great information and facts but lacks direct resources for those who need more help than do-it-yourself options.

When researching resources and programs that are available in Wyoming for obese people, I was not able to find anything that offered help to those considered obese or morbidly obese. Many websites and organizations locally and nationally, have most of their focus on reporting obesity statistics or raising money for various studies relating to obesity. Unlike the forward focus that is being put into helping people quit smoking, there are hardly any organized groups or programs geared towards helping people lose weight. There are programs directed at preventing obesity in children, that are full of solutions, names of pediatric doctors, meal plan suggestions, and exercise suggestions. It is admirable that so much effort is going into helping children stay

[5] *https://www.cheyenneregional.org/location/heart-vascular-surgical-services/patient-resources/risk-factors/quitting-smoking/*

away from an unhealthy weight, and hopefully those efforts help reduce the percentage of obese people in the upcoming generations. It is concerning however, that there is hardly any support for the adults (age 45-64) who make up an overwhelming percentage of obese people. This may be because it takes a great deal of time, money, and dedication to change one's lifestyle and many people do not feel that they have what it takes to make that kind of change. One solution that may be taken into consideration is lowering the cost of organic, healthy food, as well as gym membership prices. This would help to better incentivize the change in lifestyle that will help to lower obesity in America and ultimately the cost of health care.

Conclusion

Understanding how smoking, age, children, and BMI affect medical costs can help people make smarter decisions about their health and provide additional income to be used for health care. Age cannot be helped when considering medicals costs as it cannot be stopped or postponed. But not smoking and keeping one's weight at a healthy level are things that can be controlled in many instances. Programs and resources have been funded and put into place to help people stop smoking and to prevent obesity in children. More resources and money should be placed towards developing programs and weight management resources for those with an unhealthy weight. As these kinds of conclusions are more fully understood by American citizens, the government, and health care providers, health care cost and the overall health of Americans will greatly improve.

Appendix of Tables

Table 1
Testing 'smoker' as a significant factor

```
lm(formula = charges ~ age + sex + bmi + smoker + region +
num_of_children, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
num_of_children    475.5      137.8   3.451 0.000577 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Table 2
Testing the interaction term age:smoker

```
lm(formula = charges ~ age + sex + bmi + smoker + region +
num_of_children + age * smoker, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11296.5  -2832.8   -970.8   1420.9  29775.1

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11612.13    1010.15 -11.495  < 2e-16 ***
age                247.73      13.31  18.617  < 2e-16 ***
sexmale           -135.16     332.79  -0.406 0.684706
bmi                340.62      28.60  11.910  < 2e-16 ***
smokeryes        22105.00    1213.13  18.222  < 2e-16 ***
regionnorthwest   -362.54     476.08  -0.762 0.446480
regionsoutheast  -1060.06     478.73  -2.214 0.026977 *
regionsouthwest   -943.31     477.82  -1.974 0.048566 *
num_of_children    471.41     137.76   3.422 0.000641 ***
age:smokeryes       45.13      29.53   1.529 0.126626
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6059 on 1328 degrees of freedom
Multiple R-squared:  0.7514,    Adjusted R-squared:  0.7497
F-statistic: 445.9 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Table 3
Testing BMI30 as a significant factor

```
lm(formula = charges ~ age + sex + bmi + smoker + region +
num_of_children +  bmi30, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11944.9  -3431.6   -102.5   1538.8  28489.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7652.55    1279.06  -5.983 2.81e-09 ***
age               257.20      11.78  21.826  < 2e-16 ***
sexmale          -161.10     329.78  -0.489  0.62527
bmi               149.07      46.24   3.224  0.00130 **
smokeryes       23847.20     409.16  58.283  < 2e-16 ***
regionnorthwest  -388.48     471.72  -0.824  0.41035
regionsoutheast  -885.06     474.94  -1.864  0.06261 .
regionsouthwest  -949.51     473.32  -2.006  0.04505 *
num_of_children   477.73     136.47   3.501  0.00048 ***
bmi30            2855.09     548.91   5.201 2.29e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6004 on 1328 degrees of freedom
Multiple R-squared:  0.7559,    Adjusted R-squared:  0.7542
F-statistic: 456.9 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Table 4
Testing if the interaction term smokeryes:bmi30 is significant

```
lm(formula = charges ~ age + sex + bmi + smoker + region +
num_of_children + bmi30 + bmi30 * smoker, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-18234.3  -1826.1  -1251.6   -447.5  24803.9

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4745.546    959.685  -4.945 8.59e-07 ***
age               263.242      8.805  29.897  < 2e-16 ***
sexmale          -491.179    246.563  -1.992 0.046565 *
bmi               115.035     34.560   3.329 0.000897 ***
smokeryes       13402.363    443.910  30.192  < 2e-16 ***
regionnorthwest  -266.836    352.410  -0.757 0.449079
regionsoutheast  -825.000    354.800  -2.325 0.020209 *
regionsouthwest -1224.315    353.684  -3.462 0.000554 ***
num_of_children   520.402    101.958   5.104 3.81e-07 ***
bmi30            -865.057    425.775  -2.032 0.042381 *
smokeryes:bmi30 19794.852    610.092  32.446  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4485 on 1327 degrees of freedom
Multiple R-squared:  0.8639,    Adjusted R-squared:  0.8628
F-statistic: 842.1 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Table 5
Summary Statistics for the variable 'region'

```
> summary(sw)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1242    4751    8799   12347   13463   52591
> summary(se)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1122    4441    9294   14735   19526   63770
> summary(nw)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1621    4720    8966   12418   14712   60021
> summary(ne)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1695    5194   10058   13406   16687   58571
```

Table 6
Tukey HSD test for region

```
        Simultaneous Tests for General Linear Hypotheses
         Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = charges ~ region, data = insurance)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
northwest - northeast == 0  -988.81     948.63  -1.042   0.7245
southeast - northeast == 0  1329.03     922.91   1.440   0.4744
southwest - northeast == 0 -1059.45     948.63  -1.117   0.6791
southeast - northwest == 0  2317.84     922.16   2.513   0.0581 .
southwest - northwest == 0   -70.64     947.90  -0.075   0.9999
southwest - southeast == 0 -2388.47     922.16  -2.590   0.0480 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Table 7
Testing if has-children is a significant predictor of 'charges'

```
lm(formula = charges ~ age + sex + bmi + smoker + region + age:smoker +
    bmi30 + smoker:bmi30 + has_children, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-18228.7  -1889.8  -1188.9   -376.8  24307.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4764.822    977.438  -4.875 1.22e-06 ***
age               262.853      9.893  26.569  < 2e-16 ***
sexmale          -485.030    247.389  -1.961 0.050136 .
bmi               114.733     34.689   3.308 0.000967 ***
smokeryes       13325.469    941.738  14.150  < 2e-16 ***
regionnorthwest  -265.007    353.682  -0.749 0.453820
regionsoutheast  -849.519    356.232  -2.385 0.017231 *
regionsouthwest -1204.534    354.926  -3.394 0.000710 ***
bmi30            -846.653    427.272  -1.982 0.047738 *
has_childrenyes  1054.127    249.173   4.231 2.49e-05 ***
age:smokeryes       2.479     21.972   0.113 0.910192
smokeryes:bmi30 19762.375    613.172  32.230  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4500 on 1326 degrees of freedom
Multiple R-squared:  0.8631,    Adjusted R-squared:  0.8619
F-statistic: 759.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Table 8
Final model after backwards elimination

```
lm(formula = charges ~ age + sex + bmi + smoker + num_of_children +
    smoker:bmi30, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-18888.7  -1874.5  -1240.6   -479.2  24621.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4879.388    935.471  -5.216 2.12e-07 ***
age               263.808      8.835  29.860  < 2e-16 ***
sexmale          -488.092    247.609  -1.971  0.04891 *
bmi                98.637     33.705   2.926  0.00349 **
smokeryes       13431.633    444.812  30.196  < 2e-16 ***
num_of_children   515.972    102.300   5.044 5.20e-07 ***
smokeryes:bmi30 18929.499    644.586  29.367  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4504 on 1330 degrees of freedom
Multiple R-squared:  0.8624,    Adjusted R-squared:  0.8617
F-statistic:  1191 on 7 and 1330 DF,  p-value: < 2.2e-16
```

References

Lantz, Brett. Machine Learning with R, Packt Publishing, Limited, 2013. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uwy/detail.action?docID=1343653. Pg 204 – 217.

U. (n.d.). How health insurance marketplace® plans set your premiums. Retrieved April 07, 2021, from https://www.healthcare.gov/how-plans-set-your-premiums/

Quitting Smoking. (2016, November 04). Retrieved April 15, 2021, from https://www.cheyenneregional.org/location/heart-vascular-surgical-services/patient-resources/risk-factors/quitting-smoking/

1-800-QUIT-NOW. (2021, January 05). Retrieved April 15, 2021, from https://health.wyo.gov/publichealth/prevention/tobacco-prevention/wqtp/

World Health Organization. (n.d.). Body mass index - BMI. World Health Organization. https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi.