Smoke Less, Pay Less (for Health Insurance)

Chloe Veth, Timmy Linquist, Andrew Swierenga

Abstract

We are interested in trying to predict healthcare costs based on certain factors, such as an individual's BMI, age, number of children, and the region of the US they are from. After using several analysis tests, we found that the strongest predictor of healthcare charges is smoking, followed by BMI, age, and the region of the US where you are from.

Introduction

Medical costs are known to be a large portion of people's expenses. The US has the highest healthcare expenditures of any country, with over four trillion dollars spent in 2020. That equates to personal health care expenditures of 10,202 U.S. dollars per resident. In 2020, healthcare made up 19.7% of the United States' GDP. With such a high portion of GDP going to healthcare, specifically health insurance, we want to learn about the possible trends in healthcare expenditures. Specifically, we want to learn if it is possible to predict insurance costs based on an individual's other demographic characteristics.

We found a data set that contains individual medical costs billed by health insurance, along with certain other variables, such as the individual's age, sex, BMI, # of children, smoker status, and the region of the United States they live in. We will determine if these characteristics are significant predictors of insurance costs.

Source: https://www.statista.com/topics/6701/health-expenditures-in-the-us/#dossierKeyfigures

```
#source: https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download
insurance <- read_csv("insurance.csv")
```

```
## Rows: 1338 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
insurance
```

```
## # A tibble: 1,338 x 7
##      age sex      bmi children smoker region    charges
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
##  1    19 female  27.9        0 yes    southwest  16885.
##  2    18 male    33.8        1 no     southeast   1726.
##  3    28 male    33          3 no     southeast   4449.
##  4    33 male    22.7        0 no     northwest  21984.
##  5    32 male    28.9        0 no     northwest   3867.
##  6    31 female  25.7        0 no     southeast   3757.
##  7    46 female  33.4        1 no     southeast   8241.
##  8    37 female  27.7        3 no     northwest   7282.
##  9    37 male    29.8        2 no     northeast   6406.
## 10    60 female  25.8        0 no     northwest  28923.
## # ... with 1,328 more rows
```

Above, we have created a data frame with our dataset. This dataset contains information on people's medical bills and respective characteristics.

Exploratory Data Analysis

```
#skim(insurance) #have to comment out to knit
```

It is determined that the data set contains no missing values in any of the variable columns. This insures that any models will not be lacking data inputs.

```
recipe_insurance <- recipe(charges ~., data = insurance) %>%
  step_dummy(sex, region, smoker)

prep(recipe_insurance)
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          6
##
## Training data contained 1338 data points and no missing data.
##
## Operations:
##
## Dummy variables from sex, region, smoker [trained]
```

```
insurance2 <- bake(prep(recipe_insurance), new_data = NULL)
glimpse(insurance2)
```

```
## Rows: 1,338
## Columns: 9
```

```
## $ age              <dbl> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 5~
## $ bmi              <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 25.740, 33.44~
## $ children         <dbl> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
## $ charges          <dbl> 16884.924, 1725.552, 4449.462, 21984.471, 3866.855, 3~
## $ sex_male         <dbl> 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1,~
## $ region_northwest <dbl> 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ region_southeast <dbl> 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,~
## $ region_southwest <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0,~
## $ smoker_yes       <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,~
```

This code transforms all variables into a numeric value in order to assess correlation values and run models on the data later in the experiment. When interpreting the new values for sex, smoker, and region variables, it has been determined that a value of 1 indicates male, 1 indicates that the patient is a confirmed smoker, and 1 indicates that a patient lives in the designated region.

```
cor(insurance2 %>%
      select_if(is.numeric))
```

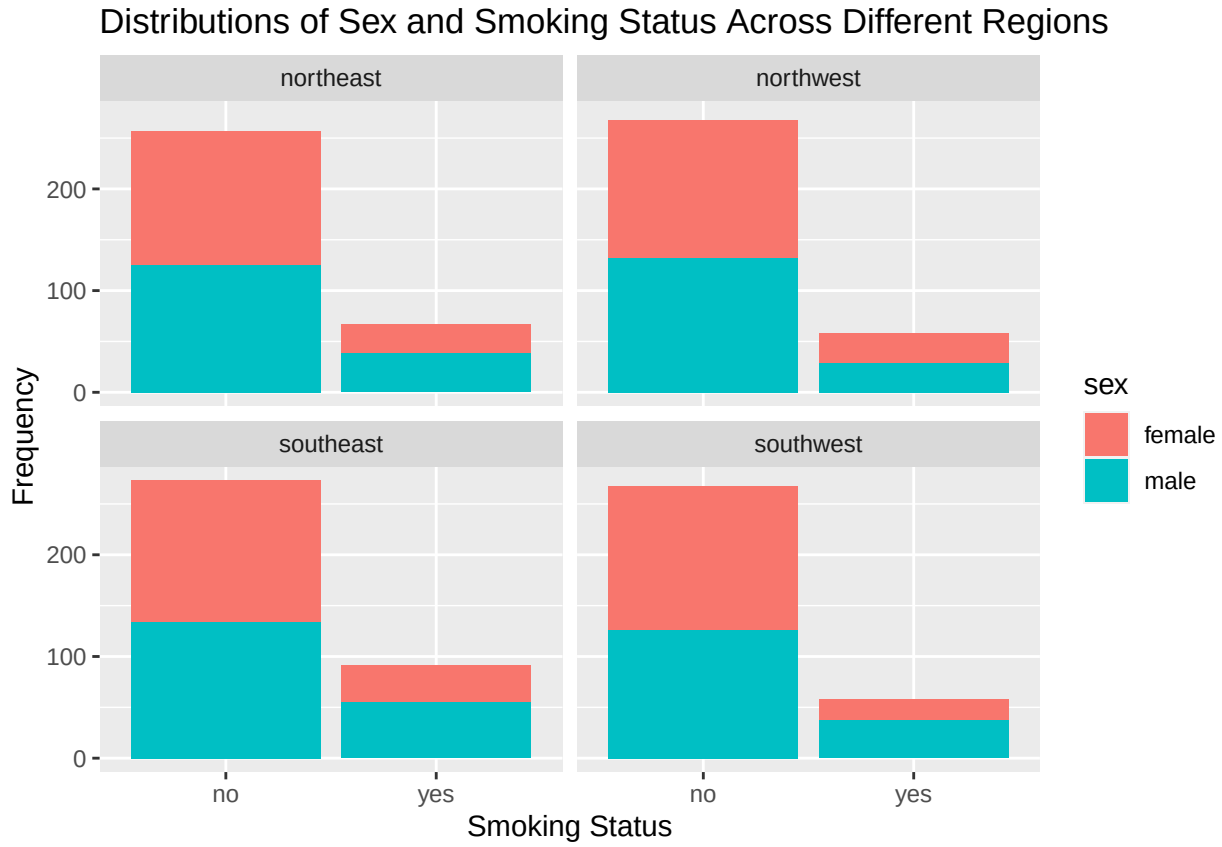
```
##                            age          bmi    children     charges
## age               1.0000000000  0.109271882  0.04246900  0.29900819
## bmi               0.1092718815  1.000000000  0.01275890  0.19834097
## children          0.0424689986  0.012758901  1.00000000  0.06799823
## charges           0.2990081933  0.198340969  0.06799823  1.00000000
## sex_male         -0.0208558722  0.046371151  0.01716298  0.05729206
## region_northwest -0.0004074234 -0.135995524  0.02480613 -0.03990486
## region_southeast -0.0116419406  0.270024649 -0.02306575  0.07398155
## region_southwest  0.0100162342 -0.006205183  0.02191358 -0.04321003
## smoker_yes       -0.0250187515  0.003750426  0.00767312  0.78725143
##                      sex_male region_northwest region_southeast
## age              -0.020855872    -0.0004074234      -0.01164194
## bmi               0.046371151    -0.1359955237       0.27002465
## children          0.017162978     0.0248061293      -0.02306575
## charges           0.057292062    -0.0399048640       0.07398155
## sex_male          1.000000000    -0.0111557280       0.01711688
## region_northwest -0.011155728     1.0000000000      -0.34626466
## region_southeast  0.017116875    -0.3462646614       1.00000000
## region_southwest -0.004184049    -0.3208292201      -0.34626466
## smoker_yes        0.076184817    -0.0369454740       0.06849841
##                  region_southwest   smoker_yes
## age                   0.010016234 -0.025018752
## bmi                  -0.006205183  0.003750426
## children              0.021913576  0.007673120
## charges              -0.043210029  0.787251430
## sex_male             -0.004184049  0.076184817
## region_northwest     -0.320829220 -0.036945474
## region_southeast     -0.346264661  0.068498410
## region_southwest      1.000000000 -0.036945474
## smoker_yes           -0.036945474  1.000000000
```

Viewing the correlation matrix of the given variables and outcome of total insurance charges, several important observations can be made. First, it can be seen that among the predictor variables, there is no indication of problematic correlation, with the highest correlation coefficient of -0.346 being between variables that indicate what region a patient is from. The correlation matrix also provides important insight

into which variables have the highest correlation coefficients with the target outcome variable "charges." It can be seen that age has a correlation of 0.299 with the total insurance charge amount and whether a patient is a smoker has a correlation of 0.787 with the total insurance charge amount.

Plots

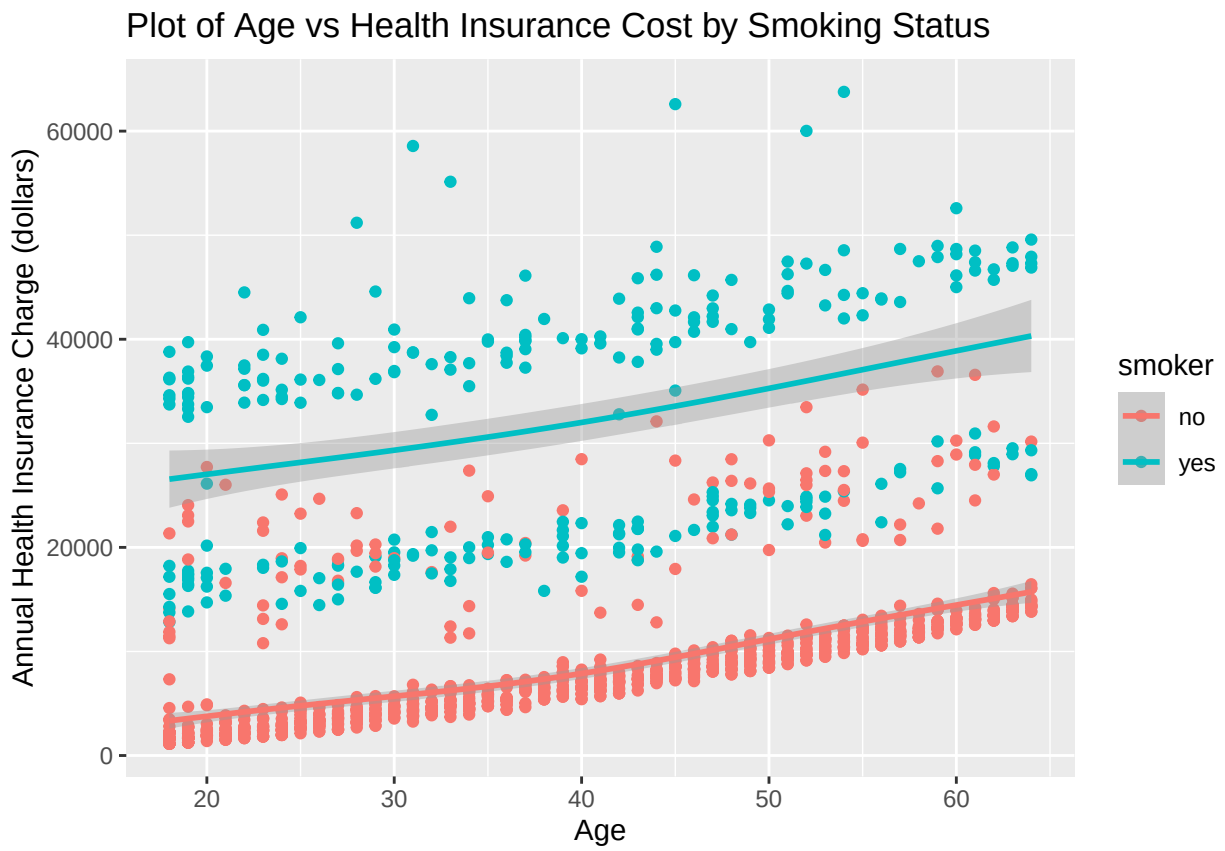
```
ggplot(insurance, aes(x = smoker, fill = sex)) +
    geom_bar() +
    labs(x = "Smoking Status",
         y = "Frequency",
         title = "Distributions of Sex and Smoking Status Across Different Regions") +
  facet_wrap(~region)
```



This graph provides informative insight into the distributions of several key variables across the four regions from which patient data is taken. It can be seen that while there are more patients who do not smoke than smoking patients, the approximate difference between these groups is consistent across all of the regions. In addition, within each of the smoking subgroups in each region, there are a similar proportion of males to females.

```
ggplot(data = insurance, aes(y = charges, x = age, color = smoker)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Age", y = "Annual Health Insurance Charge (dollars)" , title = "Plot of Age vs Health Insura
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



This plot shows that there is a generally positive correlation between age and insurance charges. It also demonstrates that smoking patients tend to have much greater insurance costs. Within the smoking subgroup, there appears to be a gap in the scatter plot which could be explained by whether or not a patient ends up needing major transplant surgery because of their history of smoking. Although, this explanation is mostly conjecture.

Statistical Analysis

Linear Model

```
fit <- lm(charges ~ age + bmi + children + sex_male + region_northwest + region_southeast + region_south
summary(fit) # show results
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex_male + region_northwest +
##     region_southeast + region_southwest + smoker_yes, data = insurance2)
##
## 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 ***
## bmi                 339.2       28.6  11.860  < 2e-16 ***
## children            475.5      137.8   3.451 0.000577 ***
## sex_male           -131.3      332.9  -0.394 0.693348
## region_northwest   -353.0      476.3  -0.741 0.458769
## region_southeast  -1035.0      478.7  -2.162 0.030782 *
## region_southwest   -960.0      477.9  -2.009 0.044765 *
## smoker_yes        23848.5      413.1  57.723  < 2e-16 ***
## ---
## 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
```

From the above linear regression, we see that several regressors have statistically significant, meaningful effects on the outcome variable, charges. We see that the estimates for age, bmi, children, and being a smoker all have positive estimated effects, with p-values less than 0.01. So, we can say with 99% confidence that these factors have strong positive effects on the amount of insurance charges. In addition, the dummy variables for an individual being from the southeast or southwest indicate negative effects on the cost of insurance, with p-values less than an alpha of 0.05. Out of the significant variables, it is clear that smoking has the largest effect on insurance cost, with an estimated increased charge of $23,848.50 per year. Age, bmi, and the number of children all have smaller estimated effects, less than $500 for each regressor. Finally, the estimated effect for being from the southeast or southwest is around -$1000.

With an Rsquared of 0.75, we know that these variables account for 75% of the variation in the outcome variable, insurance charges.

```
fit1 <- lm(bmi ~ children + smoker_yes + age, insurance2)
summary(fit1)
```

```
##
## Call:
## lm(formula = bmi ~ children + smoker_yes + age, data = insurance2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8084  -4.3356  -0.2609   4.1310  23.5353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.74241    0.51784  55.505  < 2e-16 ***
## children     0.04086    0.13780   0.297    0.767
## smoker_yes   0.09695    0.41124   0.236    0.814
## age          0.04735    0.01183   4.004 6.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.068 on 1334 degrees of freedom
## Multiple R-squared:  0.01205,    Adjusted R-squared:  0.009826
```

```
## F-statistic: 5.422 on 3 and 1334 DF,  p-value: 0.001044
```

I am also interested in seeing if there is meaningful correlation between the regressors. From this regression, we see that age has a small meaningful positive effect on bmi, but in our sample, both number of children and smoker status is unrelated to a person's bmi.

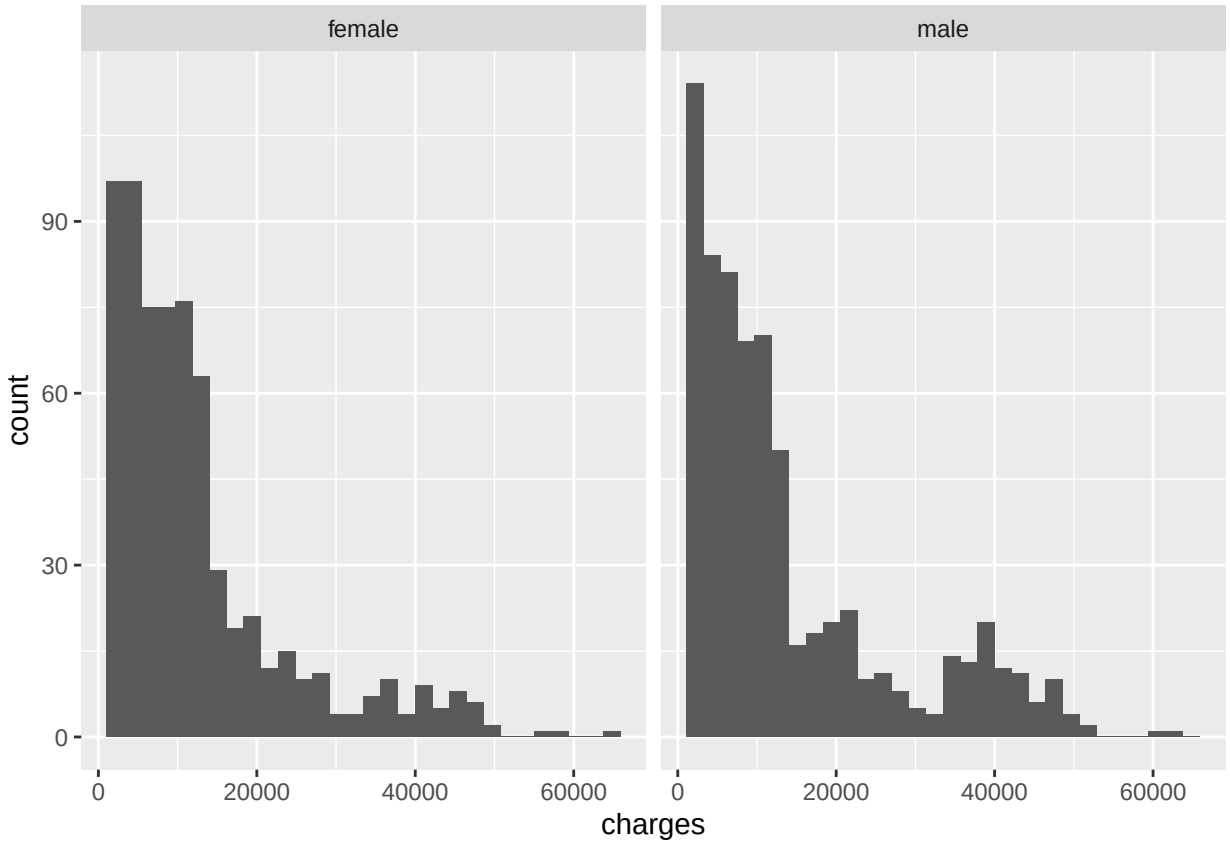
2 Sample T-Test and Confidence Interval for Sex

First, we want to investigate whether the mean insurance prices are different for males and females. To do this, we are going to perform a two-sample t-test. Our null hypothesis is that the mean insurance price is the same for males as it is for females. Our alternative hypothesis is that the price is different.

First, we need to check our assumptions. There are 3 assumptions. The first thing we are going to assume that there is independent random sampling. The second is the check for normality. Neither of these graphs appear to be very normal so we need to proceed with caution.

```
ggplot(data = insurance, aes(x = charges)) +
  geom_histogram() +
  facet_wrap(~sex)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The third assumption is the equal variance test.

```r
insurance %>%
  group_by(sex) %>%
  summarize(var=var(charges))
```

```
## # A tibble: 2 x 2
##   sex           var
##   <chr>       <dbl>
## 1 female 123848048.
## 2 male   168247513.
```

```r
168247513/123848048
```

```
## [1] 1.3585
```

The value we get is less than 3 so our final assumpton holds.

Now, we will perform the test.

```r
#First, we'll separate our dataset into 2 new data sets.
female <- insurance %>%
  filter(sex == "female")

male <- insurance %>%
  filter(sex == "male")

t.test(x = male$charges, y = female$charges, conf.level = .95, alternative = "two.sided", mu = 0, var.e
```

```
##
##  Two Sample t-test
##
## data:  male$charges and female$charges
## t = 2.0975, df = 1336, p-value = 0.03613
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    89.81229 2684.53238
## sample estimates:
## mean of x mean of y
##  13956.75  12569.58
```

The p-value here is 0.036 so at the 95% confidence interval, we reject the null hypothesis; it seems there is a significant difference in the mean charges for males and females. The confidence interval for the mean difference is (89.81, 2684.53). Because of this positive confidence interval, that does not include 0, it seems male insurance costs are generally higher than female insurance costs.

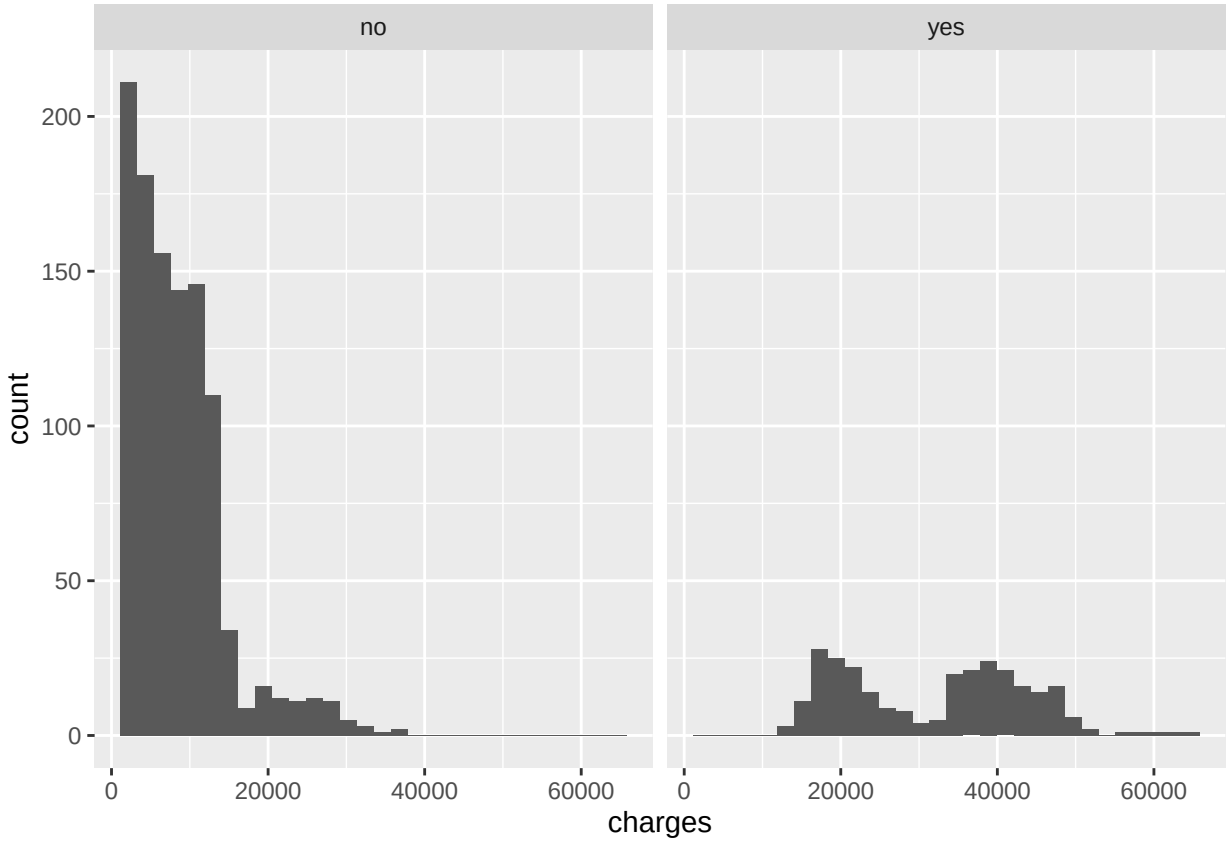
2 Sample T-Test and Confidence Interval for Smoking

One would think that smokers would have a higher insurance cost since their health is more at risk. To test this hypothesis, we will perform a similar test as before. Our null hypothesis is that there is no difference in the charges for smokers and non-smokers. Our alternative hypothesis is that smokers have a greater insurance cost.

First, we need to check our assumptions. There are 3 assumptions. The first thing we are going to assume that there is independent random sampling. The second is the check for normality. Neither of these graphs appear to be very normal so we need to proceed with caution.

```
ggplot(data = insurance, aes(x = charges)) +
  geom_histogram() +
  facet_wrap(~smoker)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The third assumption is the equal variance test.

```
insurance %>%
  group_by(smoker) %>%
  summarize(var=var(charges))
```

```
## # A tibble: 2 x 2
##   smoker        var
##   <chr>       <dbl>
## 1 no      35925420.
## 2 yes    133207311.
```

```
133207311/35925420
```

```
## [1] 3.707885
```

The value we get is greater than 3 so we will proceed with caution since 2 out of our 3 assumptions fail. We are still going to assume equal variance in our t.test call for the sake of staying consistent with what we have learned thus far in the class.

Now, we will perform the test.

```
#First, we'll separate our dataset into 2 new data sets.
smoker <- insurance %>%
  filter(smoker == "yes")

nonsmoker <- insurance %>%
  filter(smoker == "no")

t.test(x = smoker$charges, y = nonsmoker$charges, conf.level = .95, alternative = "greater", mu = 0, va
```

```
##
##  Two Sample t-test
##
## data:  smoker$charges and nonsmoker$charges
## t = 46.665, df = 1336, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  22782.97      Inf
## sample estimates:
## mean of x mean of y
## 32050.232  8434.268
```

The p-value here is so small (2.2e-16) that we reject the null hypothesis. There is evidence to say that smokers have a higher insurance charge.

ANOVA test for difference in prices by region

Next, we want to see if the charges are different for different regions. Since there are more than 2 regions, we are going to use an ANOVA test.

Before performing the test, we need to check our assumptions. First, we will assume independent, random sampling. Then we make sure that the equal variance assumption holds.

```
insurance %>%
  group_by(region) %>%
  summarize(var=var(charges))
```

```
## # A tibble: 4 x 2
##   region           var
##   <chr>          <dbl>
## 1 northeast 126693103.
## 2 northwest 122595316.
## 3 southeast 195191596.
## 4 southwest 133568389.
```

```
195191596/122595316
```

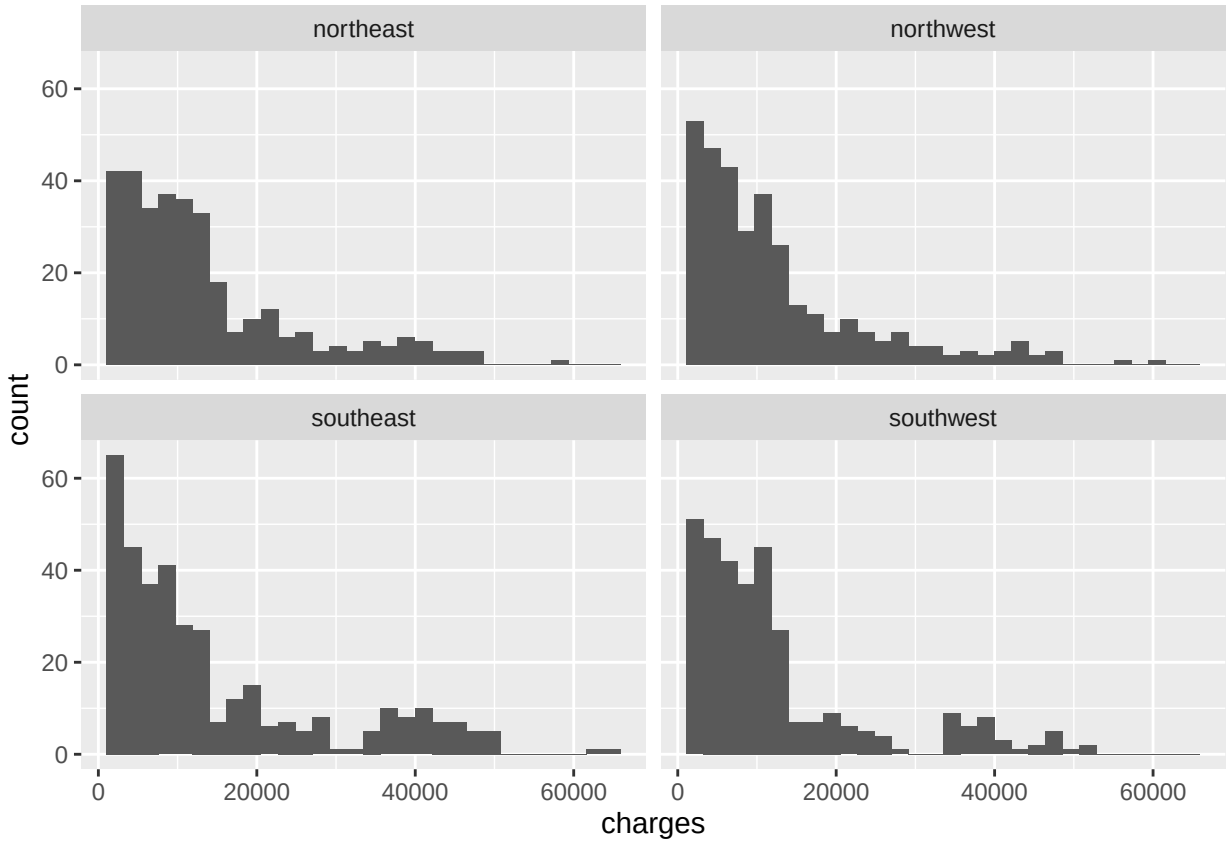
```
## [1] 1.592162
```

We divide the largest variance by the smallest variance and the number we get is 1.59 which is less than 3.

Now, we will check for our third and final condition, normality.

```
ggplot(data = insurance, aes(x = charges)) +
  geom_histogram() +
  facet_wrap(~region)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



None of these graphs appear to be normal, so we need to proceed with caution.

Now, we will perform the ANOVA test.

```
anova_data <- aov(charges~region, data=insurance)

summary(anova_data)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## region         3 1.301e+09 433586560    2.97 0.0309 *
## Residuals   1334 1.948e+11 146007093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that all of the regions have equal insurance charge means. The alternate hypothesis is that at least one of the means is different. Since our p-value of 0.0309 is small, we have evidence to reject

the null hypothesis. At least one one of the means is different, indicating region may have an affect on a person's insurance charge.

Now, to take this observation even further, we can perform a Tukey test.

```
TukeyHSD(anova_data, conf.level = .95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = charges ~ region, data = insurance)
##
## $region
##                             diff         lwr        upr     p adj
## northwest-northeast  -988.8091 -3428.93434 1451.31605 0.7245243
## southeast-northeast  1329.0269 -1044.94167 3702.99551 0.4745046
## southwest-northeast -1059.4471 -3499.57234 1380.67806 0.6792086
## southeast-northwest  2317.8361   -54.19944 4689.87157 0.0582938
## southwest-northwest   -70.6380 -2508.88256 2367.60656 0.9998516
## southwest-southeast -2388.4741 -4760.50957  -16.43855 0.0476896
```

From this, we can see that only one of the p-values is less than 0.05. There is a significant difference in the mean insurance charges of the southwest and southeast regions. The confidence interval for the difference in the mean prices is (-4760.51, -16.44). The interval does not contain 0, which is consistent with the fact that there is a significant difference.

Conclusion

First, we did exploratory data analysis on the dataset. From first glance, there appeared to be some correlation between some of the different variables. It seemed that some variables may affect insurance charge. From this, we created hypotheses and tested if what we may have observed was actually true. We used two 2-sample t-tests, an anova and a tukey test, and we fit a linear model to the data. Using all these tools allowed us to demonstrate a lot of what we have learned this semester, as well as determine if variables affect the insurance cost. The linear model showed that sex was not a significant predictor for insurance cost, but the two-sample t-test showed that there is a significant difference in insurance costs for males and females. The fact that we used both of these models/ methods allowed us to do a more detailed analysis of the data. The linear model and the two-sample t-test showed us different things. There is a significant difference in the insurance costs of males and females, but sex is not a significant predictor in relation to the other predictors. The linear model showed that smoking is the most significant predictor, followed by age, bmi, and number of children. Because we also did a two-sample t-test on smoking, we were able to see that there is a significant difference in insurance costs between those who smoke and don't smoke. The linear model also showed region may have an affect and the anova test confirmed that the southeast and southwest have different insurance costs. Combining our conclusions from all of these tests and the linear model helped up to see that there are variables that have significant effects on a person's insurance cost, some that can be controlled such as smoking and others that cannot such as age or sex. One thing to note is that not all assumptions were met in each test, so take this analysis with caution. But, the moral of the story is if you want to save money on insurance, the first thing we would recommend doing is not smoking since smoking is the most significant predictor and there is a significant difference in insurance costs between smokers and nonsmokers.