Analyzing Hospital Bills Data

Raj Zaveri

9/25/2022

### Project Overview

I this project, I looked into seeing what goes into the total charge for medical bills. I analyzed different patients’ bill totals along with different attributes to see what are the most important factors that go into determining higher or lower charges.

### Dataset Description

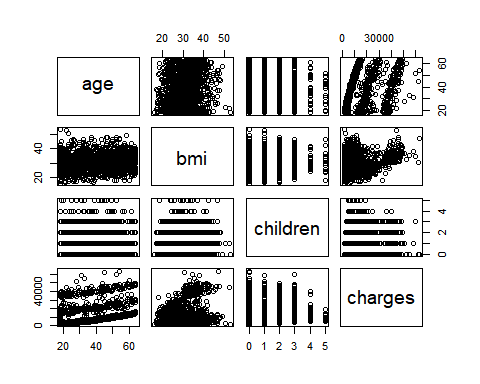
The dataset consists of approximately 1,500 records of patients’ medical bills which contain the following patient attributes: age, sex, bmi, children (number of children), smoker (yes/no), region (southwest, southeast, northwest, northeast), charges (total charge). The dataset consists of various data types including numeric, boolean and string

### Loading the Data

df = read.csv('C:/Users/rajza/Documents/Portfolio Projects/Health\_Data\_Analysis/predict-hospital-bills-main/predict-hospital-bills-main/insurance.csv', header=TRUE)  
num\_cols <- unlist(lapply(df, is.numeric))

### Exploratory Data Analysis

plot(df[,num\_cols])



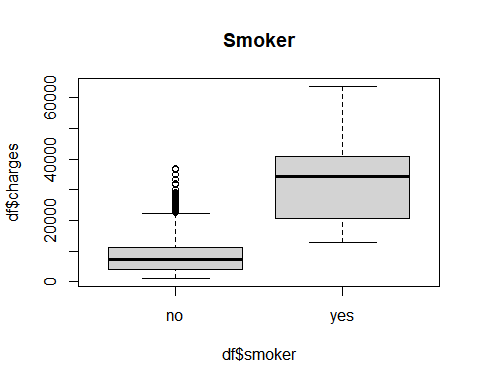
The first step I took in the exploratory data analysis process was to plot the numeric values against each other to try and understand what this dataset might look like. Some of the more significant aspects I noticed here was that as the number of children goes up, the charges actually were lower. Also, looking at bmi, there seems to be two different clusters of data points. Other than that, there seems to be correlation between age and charges.

round(cor(df[,num\_cols]),2)

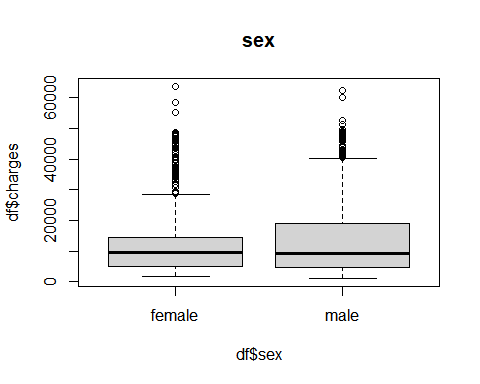
## age bmi children charges  
## age 1.00 0.11 0.04 0.30  
## bmi 0.11 1.00 0.01 0.20  
## children 0.04 0.01 1.00 0.07  
## charges 0.30 0.20 0.07 1.00

To continue with the analysis, I found the correlations between the numeric attributes. The values for charges however were not too high. The highest values were with age and bmi which were only at 0.30 and 0.20 respectively. Nevertheless, these were not the only attributes in the dataset.

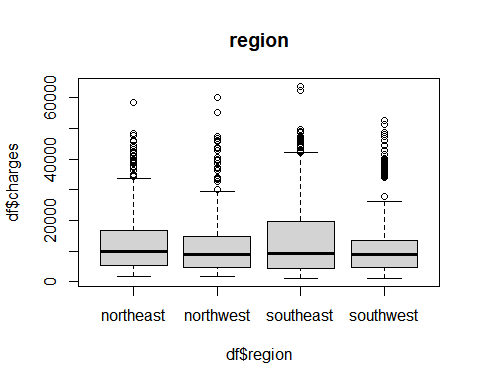
boxplot(df$charges ~ df$smoker, main ='Smoker')



boxplot(df$charges ~ df$sex, main ='sex')



boxplot(df$charges ~ df$region, main ='region')



For the other attributes, I created a few box plots to see if they had any impact on the total charges. The attribute ‘Smoker’ had a clear impact since the average charges for smokers and the IQR in general was much higher than for non smokers. ‘Sex’ also seemed to have an impact since there did seem to be some higher values for males and the average for males looked to be slightly higher. The southeast region also seemed to play a significant role when determining charges as it seemed higher than other regions. The southwest region also seemed to have quite a bit of values on the higher end.

### Linear Regression

model1 = lm(charges ~. , data =df)  
summary(model1)

##   
## Call:  
## lm(formula = charges ~ ., data = df)  
##   
## 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 \*\*\*  
## children 475.5 137.8 3.451 0.000577 \*\*\*  
## 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 \*   
## ---  
## 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

model1 = lm(charges ~ df$age + df$bmi + df$children + df$smoker + df$region , data =df)  
summary(model1)

##   
## Call:  
## lm(formula = charges ~ df$age + df$bmi + df$children + df$smoker +   
## df$region, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11367.2 -2835.4 -979.7 1361.9 29935.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11990.27 978.76 -12.250 < 2e-16 \*\*\*  
## df$age 256.97 11.89 21.610 < 2e-16 \*\*\*  
## df$bmi 338.66 28.56 11.858 < 2e-16 \*\*\*  
## df$children 474.57 137.74 3.445 0.000588 \*\*\*  
## df$smokeryes 23836.30 411.86 57.875 < 2e-16 \*\*\*  
## df$regionnorthwest -352.18 476.12 -0.740 0.459618   
## df$regionsoutheast -1034.36 478.54 -2.162 0.030834 \*   
## df$regionsouthwest -959.37 477.78 -2.008 0.044846 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6060 on 1330 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7496   
## F-statistic: 572.7 on 7 and 1330 DF, p-value: < 2.2e-16

The next step was to plot a Linear Regression model to be able to predict medical bill total charges based on certain significant attributes. We also wanted to find out what those attributes were. From the model, we can see that age, bmi, number of children, and region were definitely significant as they had p-values of less than 0.05. Age, BMI, number of children and smoker all being very highly significant. Sex was not significant.

### Conclusion

With this analysis of the dataset, I have been able to discover more insight about medical data. I was able to create a linear regression model which can help people accurately estimate and potentially determine how much they will charged at the hospital. One of the more surpising discoveries from the linear regression model was that sex did not actually play a significant role. My assumption going into this project was that women on average would be charged more than men, but this analysis proved me wrong. It was also a surprising discovery that the region also plays a role. Patients in the southern regions tend to be charged more.

Overall, analyzing this health data helped me gain more insight regarding the medical industry and motivated me to continue exploring this field even more.