Medical Cost Analysis Report

Boren Zheng

5/9/2019

1. Introduction

The dataset that is analyzed in this project is titled "Medical Cost Personal Dataset" and contains a set of 1338 observations. The dataset is saved as insurance.csv and was downloaded from the website, kaggle.com. It is comprised of seven different variables: age, sex, BMI, (number of) children, smoker (yes or no), region (in the US), and charges (USD). Personal medical cost data could help an insurance company to find the correlation between medical cost and different variables in order to design new insurance products and attract more customers. Data similar to this is typically used in the healthcare field. For example, patient data, such as the variables listed, are documented and recorded in insurance databases. Healthcare cost datasets are how an insurance company would monitor cost trends and perform analysis around specific variables to determine a potential relationship between variables and cost. This dataset provides straightforward data and variables that allow the user to analyze the data in numerous ways to try and develop potential correlations between the variables provided and medical costs of each observation.

2. Motivating Data

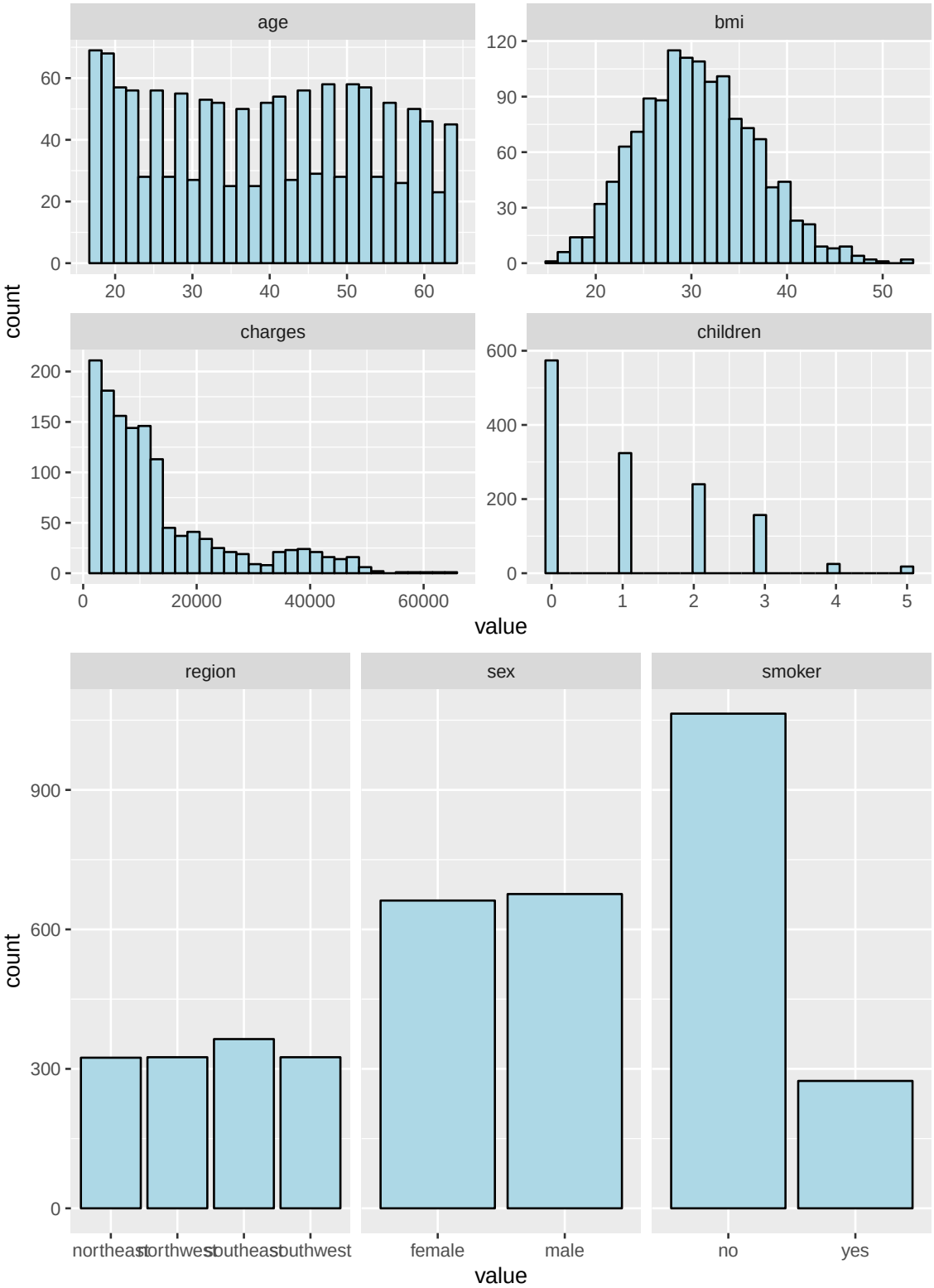
The purpose of this project is to analyze correlation and trends of specific variables when compared to cost. Any correlation can be viewed as a trend and then be used as cost prediction for future observations. This information is of high value to insurance companies in particular health insurance. For example, this dataset contains the variable "smoking status." Smoking has been known to cause health problems such as lung cancer and COPD, thus potentially increasing the medical costs of someone who smokes. Within this dataset, the variable is recorded as "yes" or "no" for smoking status. The cost of those patients who smoke can then be compared to the cost of those patients who do not smoke to support the hypothesis that people who smoke will have higher medical costs. Variables that are utilized to predict cost can then be used by actuaries to help determine the "risk" the insurance company is in by offering coverage to a patient and then access their insurance premiums accordingly. The objective of this study is to determine and analyze potential predictors of cost within the dataset through statistical methods in R. We hypothesize that BMI and age will have a direct correlation to cost in that the higher a patient's BMI and/or age, their medical costs will be higher than patients with lower BMI and/or younger age. Our second hypothesis is that smoking will have the highest relationship with cost and will be the best predictor of cost compared to those who do not smoke. Our last hypothesis is that sex, region, and number of children should not have a significant difference in distribution of cost since observations should be relatively even for those variables.

3. Real Data Analysis

Data Distribution

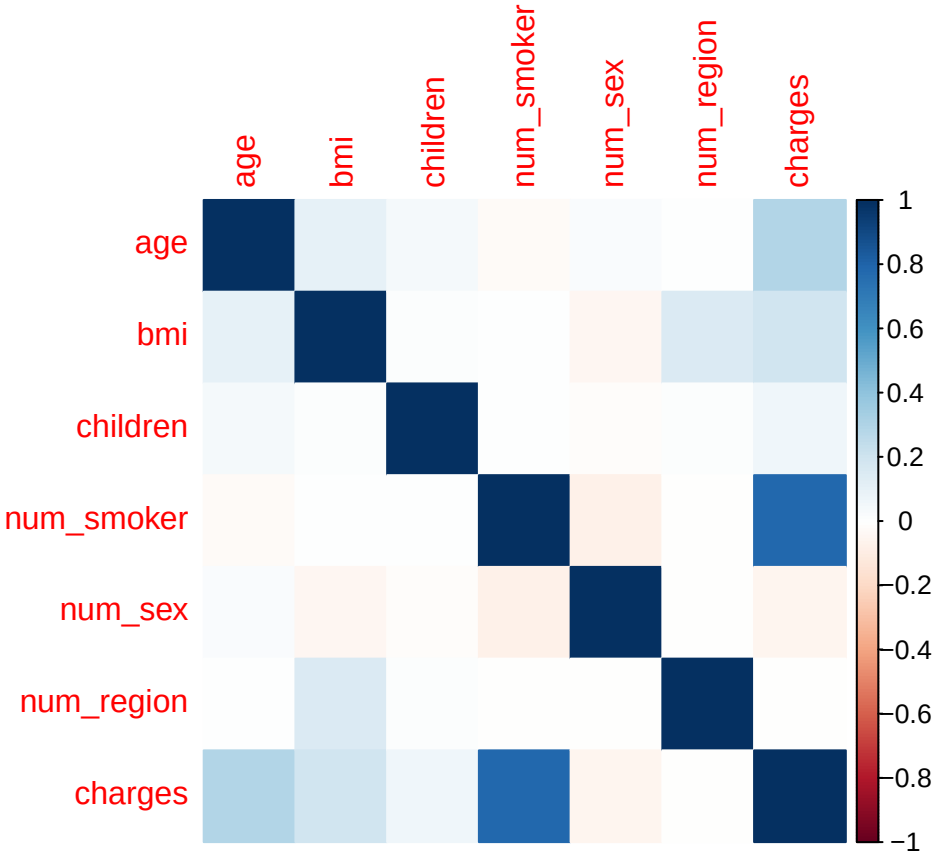
Before data analysis was conducted, the distribution of the variables in the dataset had to be determined to provide reasoning for any potential discrepancies in the analysis. Age has an even distribution except for a higher population of young adults (18-19 years old). BMI follows a relatively normal distribution curve, which is expected in an average population. Charges follows a random distribution showing most of the observations below 20,000. However, after 20,000, the observations start to decline and then increase around 40,000. This random distribution is not the highest of concern since the analysis is determining correlations for values within the cost variable. Children follows a decreasing distribution, demonstrating a decrease in observations as the number of children increase. Smoking status is shown as around one in four observations

smoke within this dataset. This distribution is expected and will be a focal point of the data analysis. The last two variables, region and sex, demonstrate a relatively even distribution, there are slightly more male observations as well as more observations in the southeast region. This should not have a significant impact in the analysis but is worth noting.



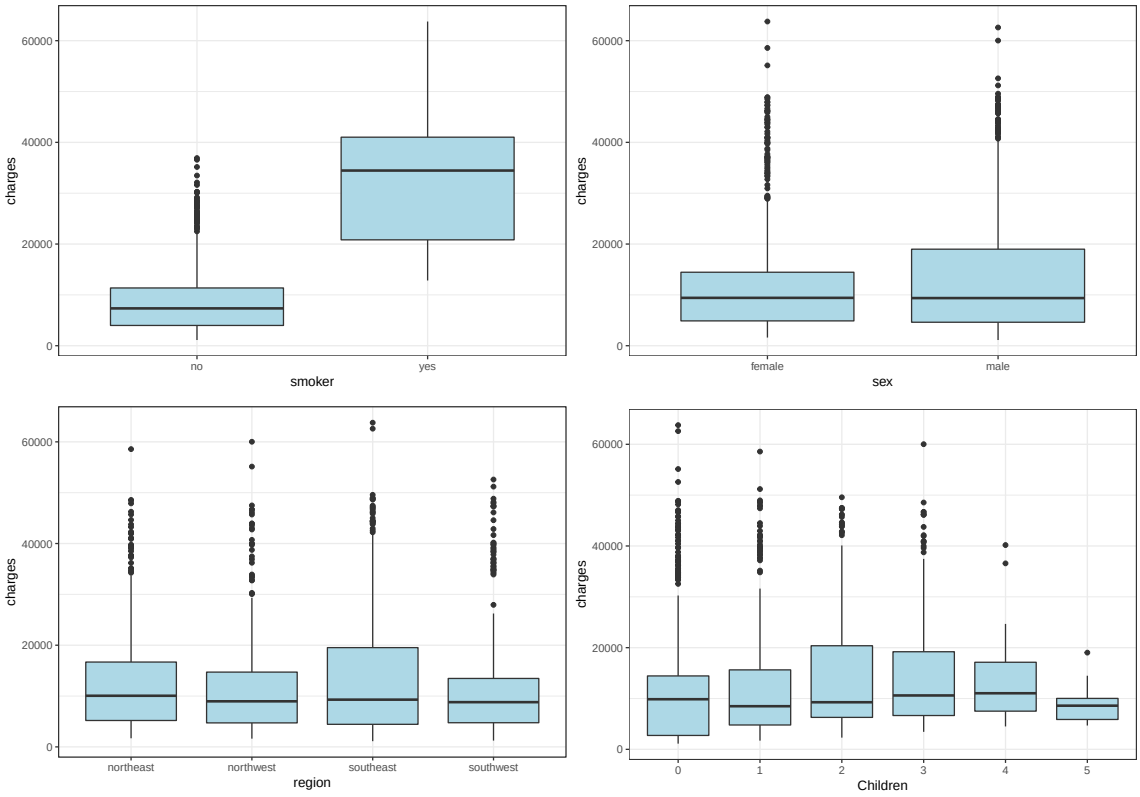
Correlation matrix

The factor variables smoker, sex, and region were converted into numeric. A correlogram was plotted to find the correlation between the variables and charges. In this plot, correlation coefficients is colored according to the value. It was determined that smoker, bmi and age have direct correlation with charges. Analysis was then performed on these variables.



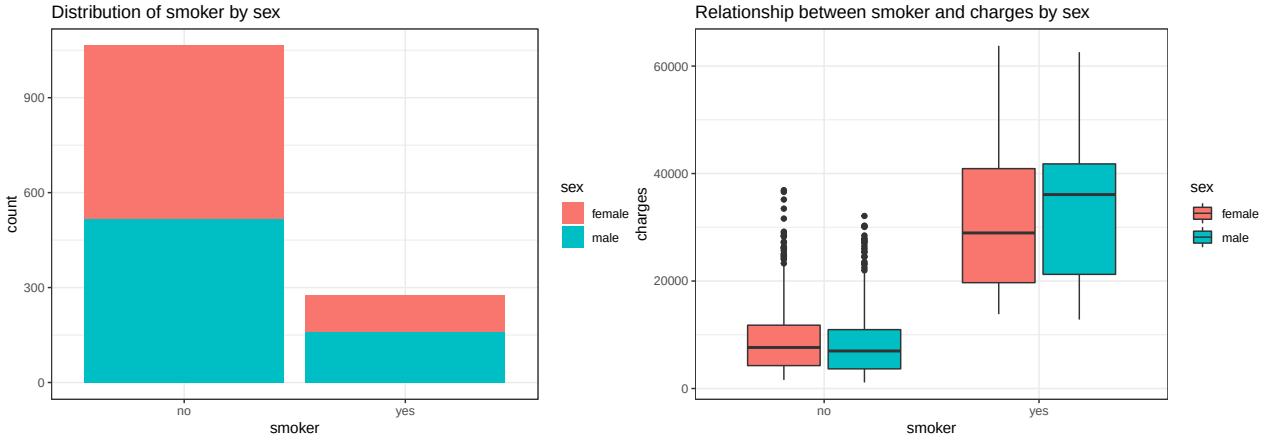
Analysis of Categorical Variables

Analysis was run on the three categorical variables in the dataset to determine relationship between charges. As shown by the correlogram, sex and region do not have correlation with charges, however, it was important to further defend this observation. For charges plotted with sex, there exists a relatively even relationship between cost and sex. As noted, and observed in the variable distribution, there exists a slightly higher amount of observations of males in the dataset, which may be represented by the higher range compared to females. The distribution of charges based on region demonstrates a relatively even distribution as well with no significant difference. For the southeast region, it can be seen that a higher range and higher maximum exists (as shown by the box plot). However, as mentioned, the southeast region contains more observations, so this result is expected. Number of children does not demonstrate any significant correlation with charges. The difference in the range of the box plot is defended by the distribution of the number of children variable within the dataset. The last categorical variable that was plotted with cost was smoking status. The box plot distribution demonstrates a much higher average cost, median cost, maximum, minimum and range thus demonstrating smoking status (yes) is highly indicative of medical charges.



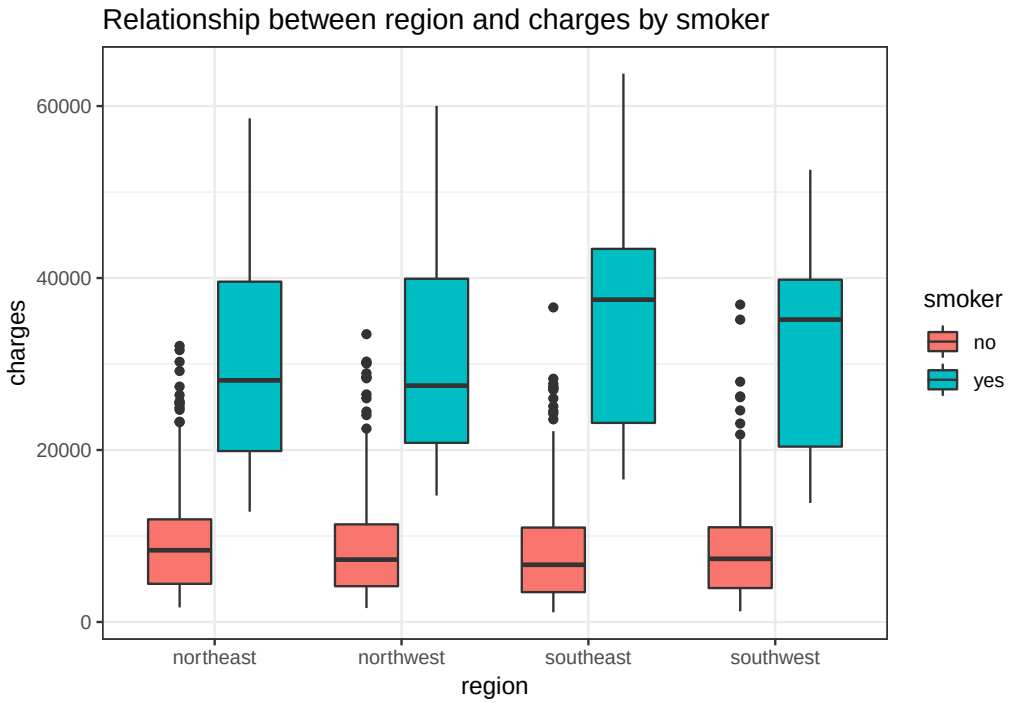
Variables Analyzed with Smoking Status

As discovered in the analysis, smoking has a very high correlation with charges. Further analysis was performed around smoking status with another variable correlated with cost. The data contained a slightly larger sample of males than females, the "distribution of smoker by sex" bar graph demonstrates that distribution. It can be seen in the box plot of smoking status with sex compared to charges as well. There exists a similar average healthcare cost between males and females who smoke and in those who do not. However, the median cost is higher in males who smoke compared to females who smoke. This is indicative of the higher male sample size and distribution of males who smoke.

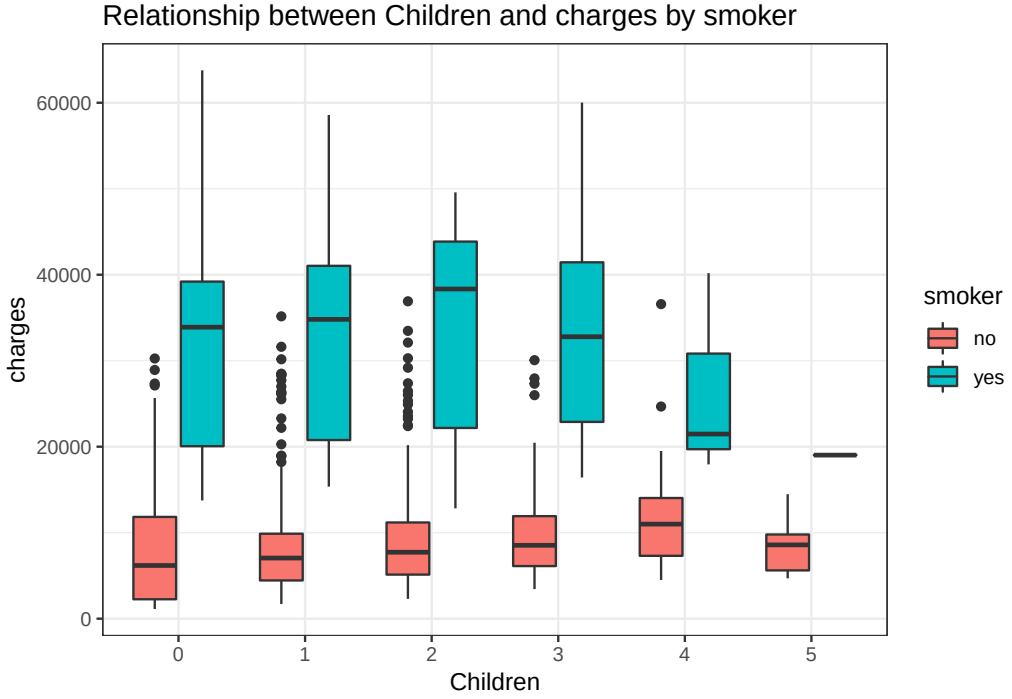


As for smoking status by region. There is no significant difference in charges by region. The only difference in charges by region is the southeast region, which is due to the larger sample size. However, there is no real difference in distribution of charges in nonsmokers by region when there is a slight difference in smokers. It can be observed that there is a higher health care cost in the southeast region for smokers compared to the

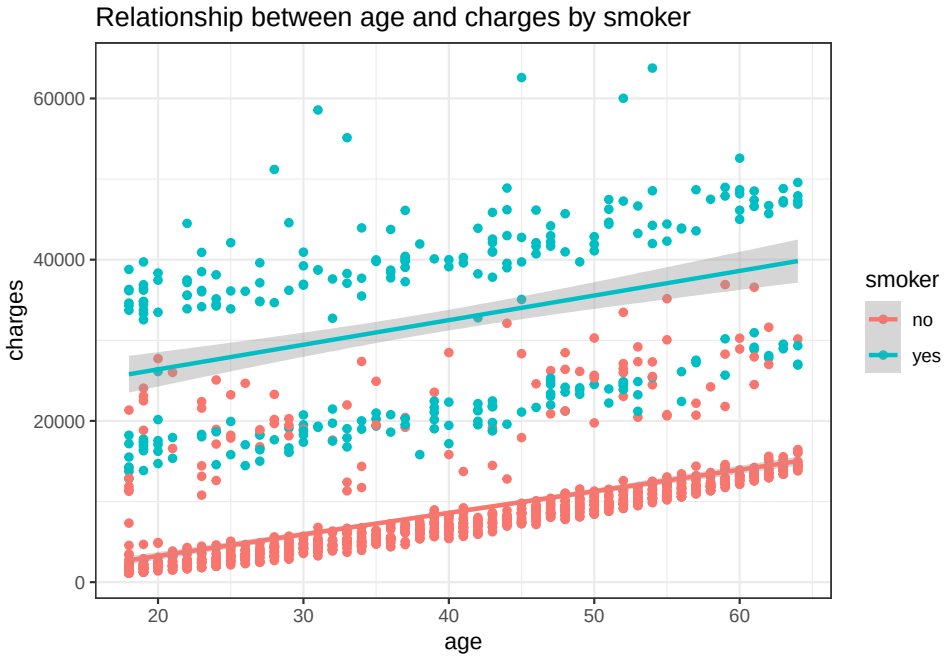
other regions thus signifying there are either more smokers in the southeast or higher costs for smokers in the southeast. Another explanation may be that the difference in sample size could contain only outliers, which could increase the average and range as well as slightly increasing the median. However, this is unlikely.



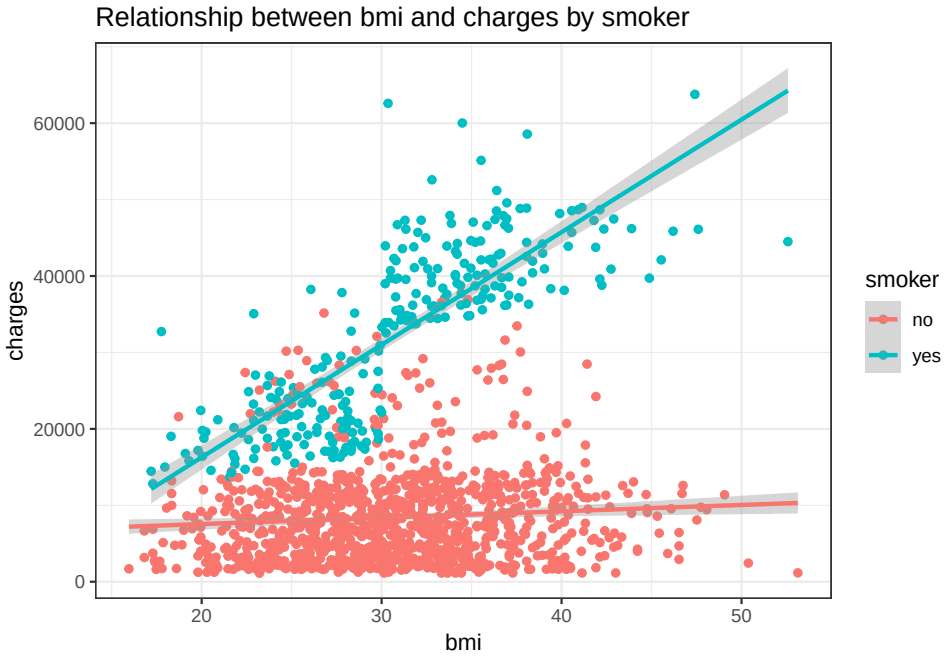
Smoking status with number of children was analyzed to determine the relationship with charges. There exists no real significance in difference of charges for smokers and children as well as nonsmokers and children. The slight differences in range is resembled by the number of observations of the variable within the dataset, as shown above. For parents with 4 children or more, there exists some difference worth mentioning, however the sample size is too low to make a conclusion on that variable with smoking status and cost



Smoking status with age was analyzed to determine the relationship with charges. Based on the plot, it can be observed that heath care costs increase with age (as represented by the increase for non-smokers). This plot also demonstrates that heath care cost is higher in those who are older and smoke than those who are younger and smoke, demonstrating that increase in age in people that smoke, will have some of the highest costs in the dataset.



Smoking status with BMI was analyzed to determine the relationship with charges. In the plot below, for nonsmokers, there is no significant correlation between increase in BMI and increase in cost (or vise versa). However, when looking at smokers and increase in BMI, there is direct correlation between BMI and charges. Comparing observations of smokers with BMI over 30 to smokers with BMI under 30, there is about a two-fold increase in cost. Thus demonstrating smokers with BMI over 30 should have some of the highest healthcare costs in the dataset.



4. Model

Model 1: Multiple linear regression

```
model <- lm(charges ~ age + bmi + children + sex + region + smoker, data = data)
```

Response variable: charges
Number of effects: 7
R-square: 0.750913
ANOVA is used to test this model

Model 2: Polynomial Regression

```
model2 <- lm(charges ~ age + I(age^2) + I(age^3) + bmi + I(bmi^2) + I(bmi^3) +
               children + sex + region + smoker, data = data)
```

Response variable: charges
Number of effects: 11
R-square: 0.7555693
ANOVA is used to test this model

Polynomial Regression gives us a similar R-square value although a little higher, in such case, we prefer multiple linear regression because it is more straightforward.

5. Discussion

Limitations

This study is only limited to the dataset obtained on Kaggle.com. It is unknown what specific American population this dataset includes and may not be representative of the commercially or publicly insured. Outliers were not taken out of this study and were used in this analysis; certain statistical methods can account for such outliers and can be seen in the charts and graphs.

Conclusion

The objective of this study was to determine trends and predictors of medical costs. As shown in the data analysis there are a number of variables that health insurance companies could utilize to help determine premiums and cost of coverage. Our hypothesis was correct in that increase in age is shown to be associated with an increase of cost. This is expected because as people age, the harder time the body has in fighting off disease, thus increasing medical cost. Those that smoke were also associated with higher medical costs, which was explained in the introduction why this is. Sex, number of children and region did not demonstrate any significant correlation with costs and should not be used for cost prediction (with the minor exception of the southeast region as explained in the data analysis). Smokers in any variable were shown to be associated with higher costs than those who do not smoke within the same variable being analyzed. We were correct on all our hypotheses except for our hypothesis of BMI. Increase in BMI is only associated with an increase in health care costs for those that smoke. Non-smokers with a relatively high BMI were shown to have similar costs to those with average and even low BMI. Smokers with high BMI, specifically over 30 BMI, were associated with very high costs compared to those who do not smoke (about four-times) and were much higher than smokers under 30 BMI (about two-times). Reasoning behind this finding cannot be determined, however, these findings should definitely be taken into consideration by health care providers and warrants further research.