# Team 3 Final Project

### **Introduction**

**Dataset**:

* Source: [Medical Cost Personal Datasets | Kaggle](https://www.kaggle.com/datasets/mirichoi0218/insurance)
* Variables:
  + Age
  + Sex
  + BMI
  + Children
  + Smoker
  + Region
  + Charges

**Business Questions:**

* What are the demographics of our patient population?
* What variables have the most effect on insurance charges?
* How can we predict an individual's insurance cost based on their variables/ how accurately can we forecast costs for each patient by looking at the data?
* What variables in the dataset can an individual control to potentially reduce their insurance charges, and by how much?
* Other than behavior of individuals such as smoking, what can cause increase of medical cost?
* What variables have the strongest correlation?
* What are the limitations of the dataset in relation to medical cost?
* What topics can be explored in the future?

### **Methodology**

We began by identifying the demographics of the data set and identifying whether any data cleaning was required. Overall, very little cleaning had to be done. We created dummy variables of the categorical data (smoker, sex, and region). With four regions included in the data, we created three dummy variables and the fourth was included in our base case.

We created a regression model to identify the statistically significant variables on our dependent variable of charges. With the information from this model, we then removed the non-statistically significant variables (region was removed even though two of three dummy variables were statistically significant because it was decided it would be inappropriate to continue to analyze other regions when one is not within the model).

After creating a regression model of key variables, we evaluated which variables had the most effect on the model by dropping each off. After that we created a prediction function to be used by an end-user that would predict an output of charges based on our input variables (age, BMI, smoker, and children). We then used that function to create a Shiny app for advanced user input.

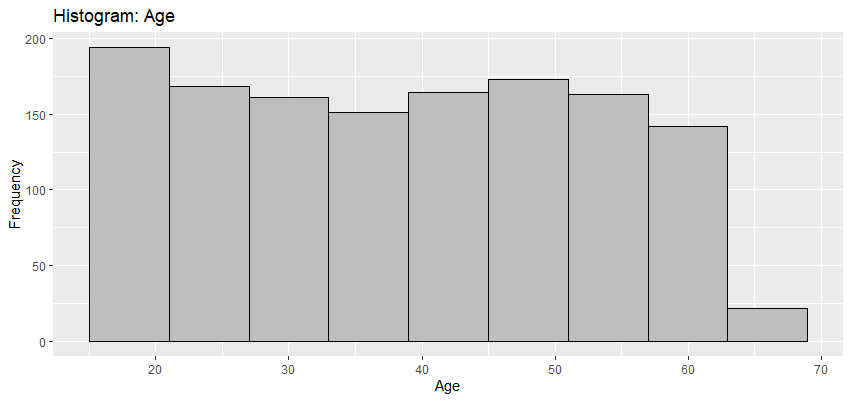
Initially we hypothesized that there were variables in the data set that an individual could or could not control. With the coefficients and other models we created, we identified that BMI, children, and smoking can be controlled (to a certain extent) to reduce an individuals’ insurance charges.

### **Results**

**Demographics**

Age

* Average: 39.21
* Range: 18-64



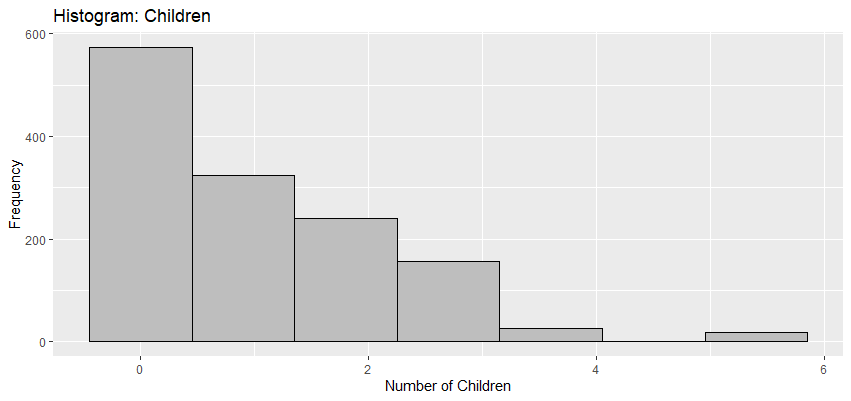
BMI

* Average: 30.66
* Range: 15.96-53.13



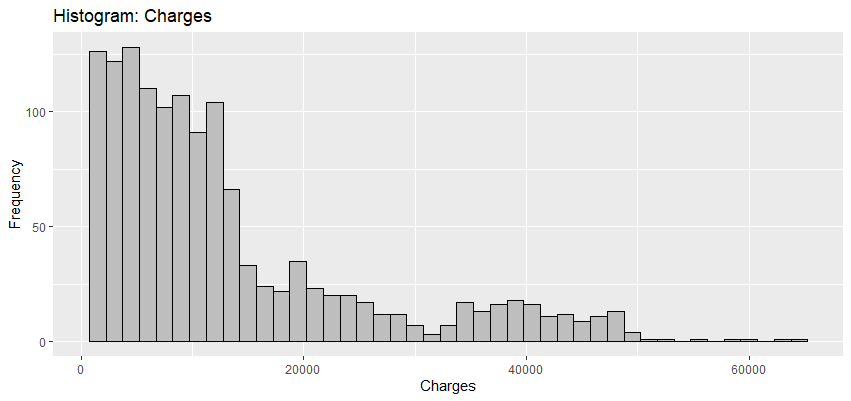
Children

* Average: 1.09
* Range: 0-5



Charges

* Average: $13,270.42
* Range: $1,121.87-$63,770.43



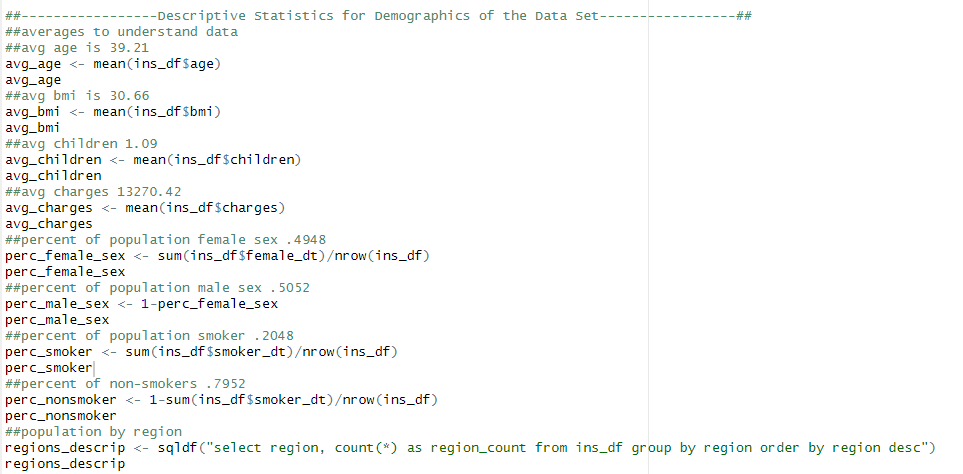
Percent of Population by Sex: 4

* 9.48% Female
* 50.52% Male

Percent of Population by Smoking Habits

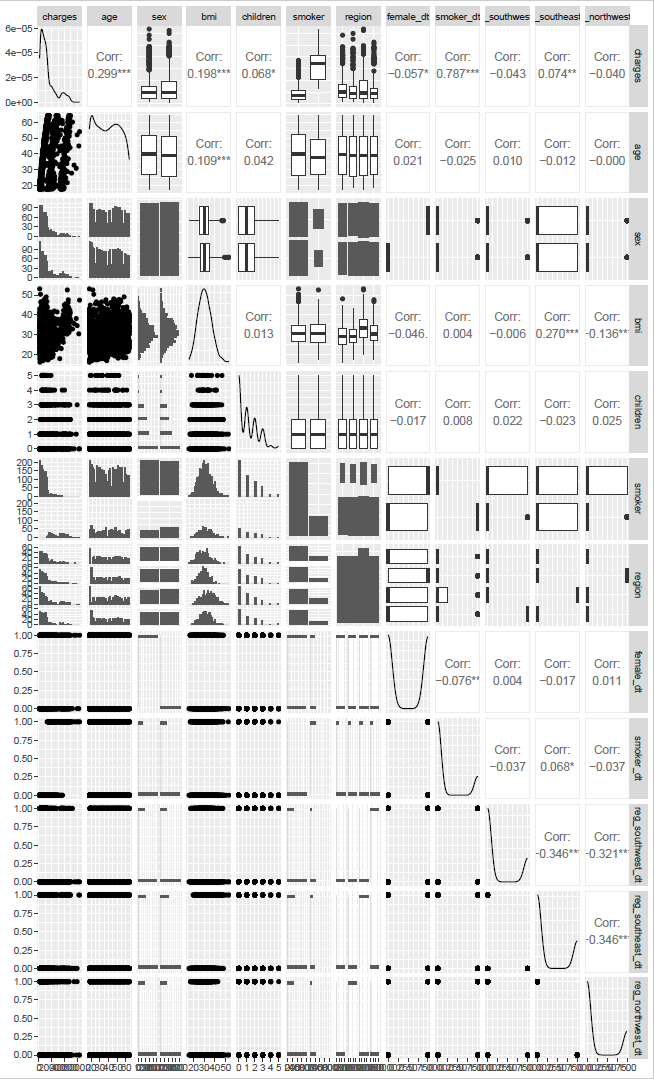
* 20.48% Smoker
* 79.52% Non-smoker

R-Code to generate demographics



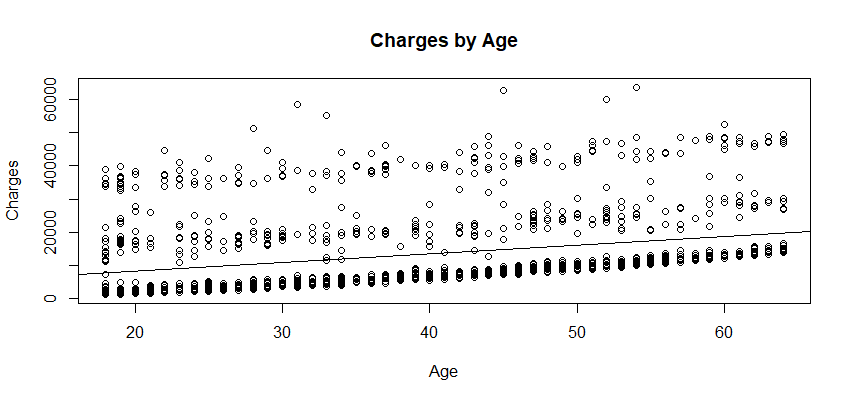
**Correlation of Variables**

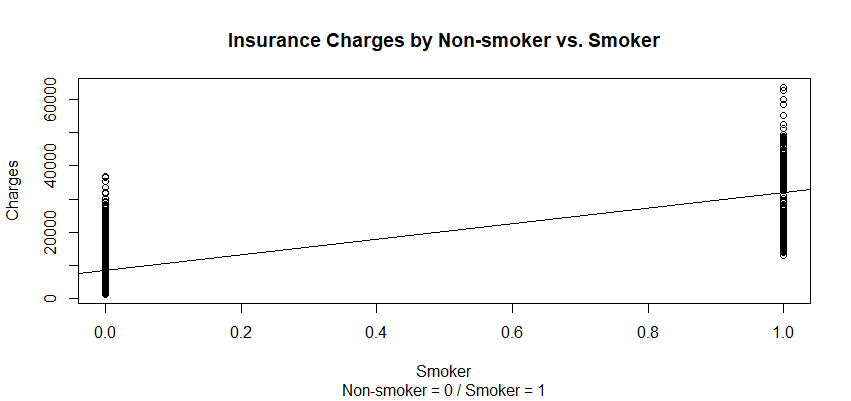
The GGally package was utilized to get a visual of negative and positive correlations with insurance charges. The most significant correlation is between charges and smokers. Meaning, if you are a smoker, you are likely to see a 78% higher insurance rate than a non-smoker. Age proved to be the next highest correlation. The older you are, then you are more likely to pay 30% more than those younger than you. Interestingly, there is a 27% higher insurance rate to individuals who reside in the southeast region with a high body mass index (BMI), where there is little to no correlation in the other regions (southwest and northwest).



**Most Predictive Variables**

Two variables that explained the change in our dependent variable by our x-variable are age and smoking.





**Variables that Patient Could Potentially Change**

Based on the GGPlot for BMI, you can see that when the cost is 30K and above, the BMI is between 30 and 50. Looking at this you can assume that if you lower your BMI closer to 30 and below then you would be able to lower your cost. The second plot shows the average charges compared to BMI and the higher the BMI the more the cost increases.

**Chart, scatter chart

Description automatically generated**

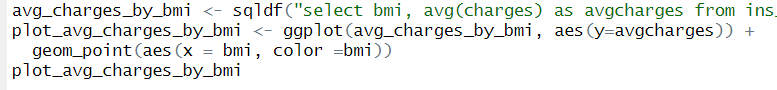
**Code used:**

**Picture 9**

**Chart, scatter chart

Description automatically generated**

**Code used:**

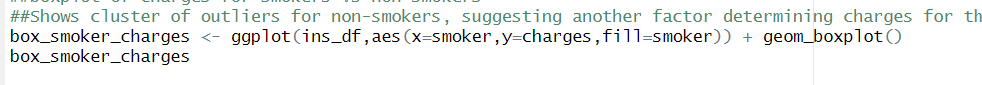
****

A second variable that could affect cost would be smoking status. Based on the box plot numbers for smoking vs. non-smoking, if the patient were to be a non-smoker, they could significantly decrease their costs. While there are outliers in the non-smoker data, most of the non-smokers have lower costs.

**Chart, box and whisker chart

Description automatically generated**

**Code used:**

****

**Variables that Patient Could Not Reasonably Change**

While there are variables that can be controlled by the patient, there are some that cannot be controlled. For example, age is uncontrollable – in the line chart below you can see that as the age of the patient increases the charges also increase. While the patient might be able to lower their costs by controlling the previously mentioned variables, the cost will still increase based on age.

**Chart, line chart

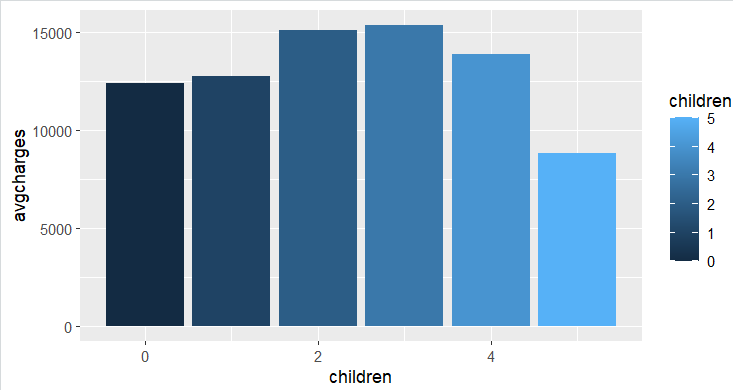
Description automatically generated**

**Code used:**

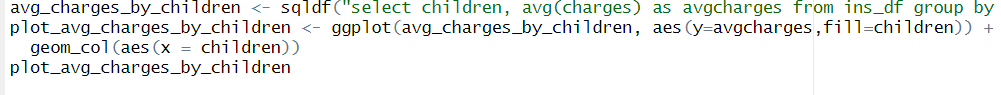
**Text

Description automatically generated**

Number of children is another variable that cannot reasonably change. The patient cannot choose to not have their existing children. In the bar chart below, you can see that the highest cost is when the patient has 2-4 children.

****

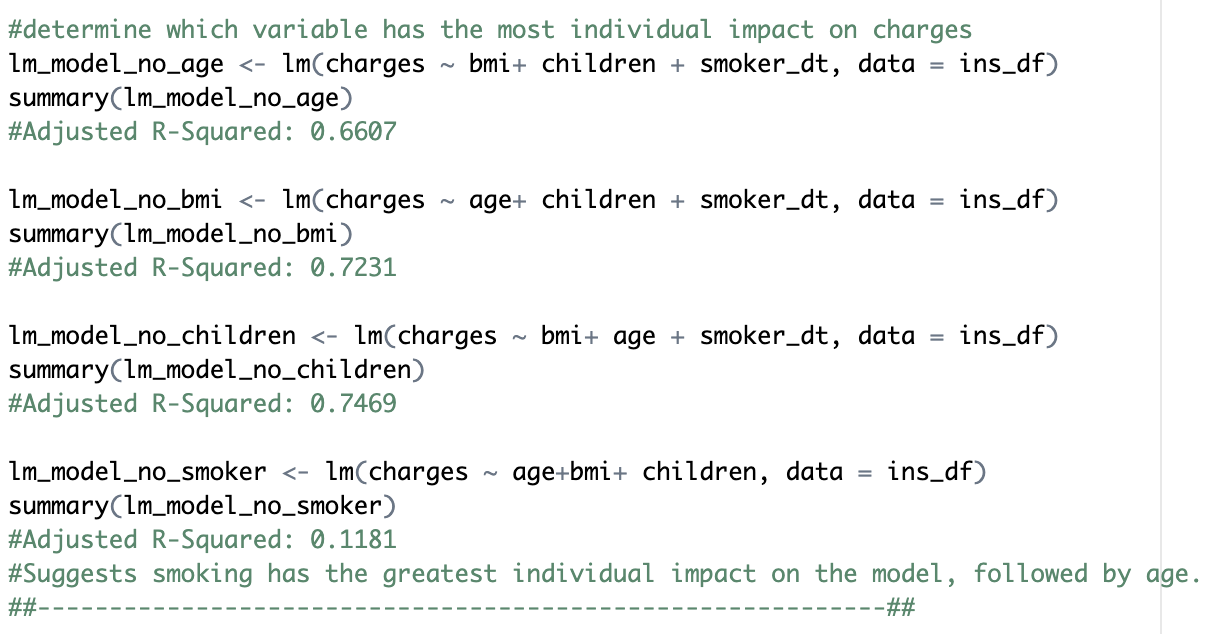
**Code used:**

****

**Predictions and Forecasting**

Though there are variables not listed in this dataset which could impact insurance charges a patient would incur, based on the data and several regression models, we determined that the most predictive variable was smoking habits, followed by age.

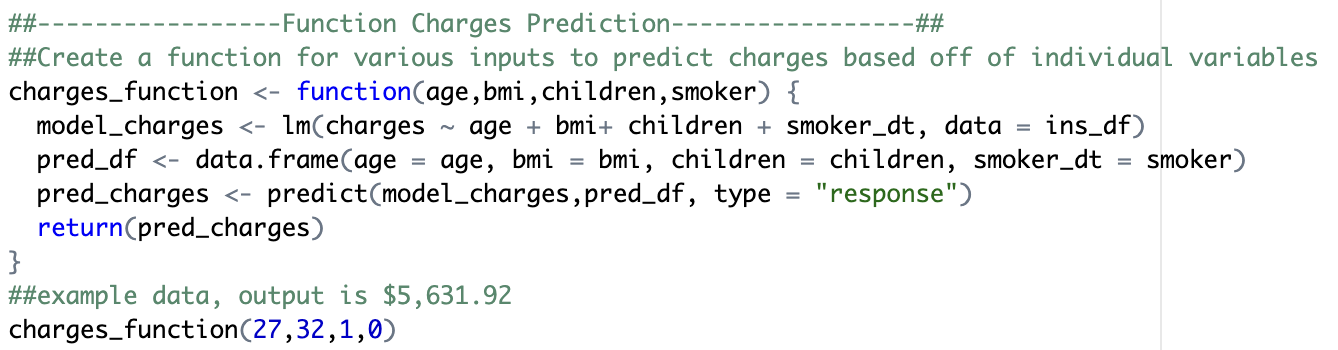
**Code:**



While the regression models we created showed that these variables were very statistically significant, the model with the highest accuracy yielded an adjusted R-squared of around 75% — suggesting that the model is accurate, but not very accurate. We believe that if more data was available (for example type of coverage, pre-existing conditions, economic status, etc.), the model would be more accurate. However, as we were dealing with medical/insurance data, this information was likely and appropriately withheld for privacy reasons.

We also created a predictive function to estimate charges a patient would incur based on their age, BMI, number of children, and smoking habits.

**Code:**



And in turn created a Shiny app that would allow the client to estimate charges without an understanding of the code.

**Code:**



**Limitations of the Dataset and Future Topics to be Explored**

Healthcare cost is complicated and there are factors beyond variables that can impact cost including accessibility and equity in healthcare1. Although the dataset was obtained from claims billed by a hospital to an insurance company, it is unable to predict why the average healthcare cost is high or higher than other developed nations2.

One of the things that complicates healthcare cost is lack of transparency in the bill sent by the hospital including administrative cost and fees which varies from hospital to hospital and region to region which makes the cost hard to measure. It is also another reason why we removed regions as variable even though some of the region variables were statistically significant.

The dataset was also obtained from one insurance company, we know there are more than one insurance companies including government health insurance. A cost comparison is needed across healthcare systems to predict average cost.

The role of uninsured patients in the healthcare cost is one of the topics that needs to be explored in the future and a cross comparison of different regions and sexes in relation to the cost might give us an insight on the limitations of the dataset. For example, we know the Southeast has higher average cost compared to other regions due to fewer hospitals and lack of accessibility.

We also have several large outliers in the cost in some of the regions which may affect assumptions made in our linear regression analysis because it skews the cost to a higher average.

Lastly, models cannot predict future events that appear suddenly and affect the entire healthcare system. For example, the COVID-19 pandemic caused disruption in the healthcare systems and drove costs.

**Conclusions**

Overall, this dataset and the questions we asked of it provided a big picture view of what variables impact insurance charges. However, the dataset did not include details like diagnosed conditions, whether a patient was married, what sort of insurance coverage they had, etc., all of which might have significantly impacted their insurance charges. The dataset also had very little data from senior patients (I.e. 60+), who are generally expected to incur more healthcare costs (though these costs are often absorbed by Medicare). In addition, though the dataset provided the region in which the patient lived, this data was vague and not of use during the project. With these factors in mind, our analysis may only be significant in the context of this dataset.

On the data analysis side, this dataset was useful for testing out many of the techniques we learned in this class. We were able to evaluate which variables were most predictive of charges with regression models, create visualizations that could explain this relationship easily, and create a function and a Shiny app that a non-data scientist could use to predict charges.

While processing the data, it became clear that some of our original business questions overlapped, and some were difficult to answer with the tools and data available to us. We made some adjustments to our original questions to better reflect what was important in the dataset, focusing on which variables are most predictive, where to go from here, and, in general, the information that would be most useful for the client to know.

**References**

1. Institute of Medicine (US) Committee on Monitoring Access to Personal Health Care Services; Millman M, editor. Access to Health Care in America. Washington (DC): National Academies Press (US); 1993. 1, Introduction. Available from: https://www.ncbi.nlm.nih.gov/books/NBK235885/
2. [Health Serv Insights.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9240587/) 2022; 15: 11786329221109755.Published online 2022 Jun 23. doi: [10.1177/11786329221109755](https://doi.org/10.1177/11786329221109755)