**Healthcare cost based on Health Management Organization**

**Group: M004\_Anderson\_Group 1**

**Group Member: Vaibhav Chaudhari; Shan Jiang; Vedant Sanjay Murudkar;**

**Sai Roshan Varma Pandati**

**Introduction**

This report presents the finding of a customer survey conducted by Health Management Organization. After a comprehensive understanding of the data, we aim to analyze it to provide actionable recommendations to Health Management Organizations to lower their total healthcare costs. After data cleansing, our team plans to apply three data modeling techniques in this course to identify the attributes influencing the customer’s costs to recommend. So, we can focus on these attributes to lower customer healthcare costs next year. The recommendations will base on our modeling result.

**Business Question:**

1. What is the meaning of these variables in the survey?
2. Do the existing variables sufficient? Do any new variables need to be added in the future survey?
3. What are the key variables that influence customer health care costs most?
4. How these variables influence customer health care costs, higher or lower?
5. What are some possible actions Health Management Organization can take to improve?

**Data Cleansing**

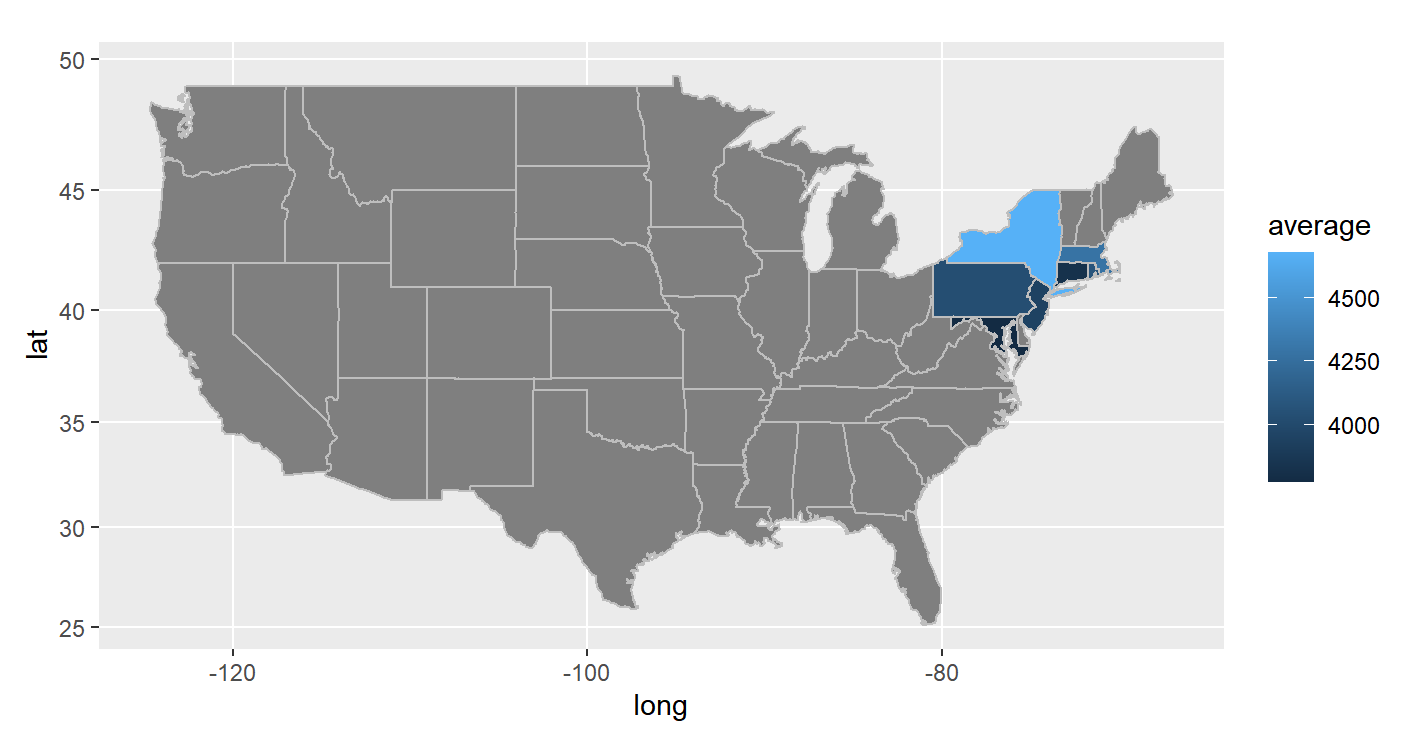
The source data has 7582 observations. There are 78 missing values in the ‘bmi’ and 80 in the ‘hypertension columns’. As hypertension is a binary data type column, we cannot replace missing values with mean or na\_interpolation. We will delete all rows with missing values for the ‘hypertention’ attribute. ‘bmi’ is a continuous value type column; we can use na\_interpolation to substitute the NAS in the column. It also means our analysis will not include the case that the value is missing.

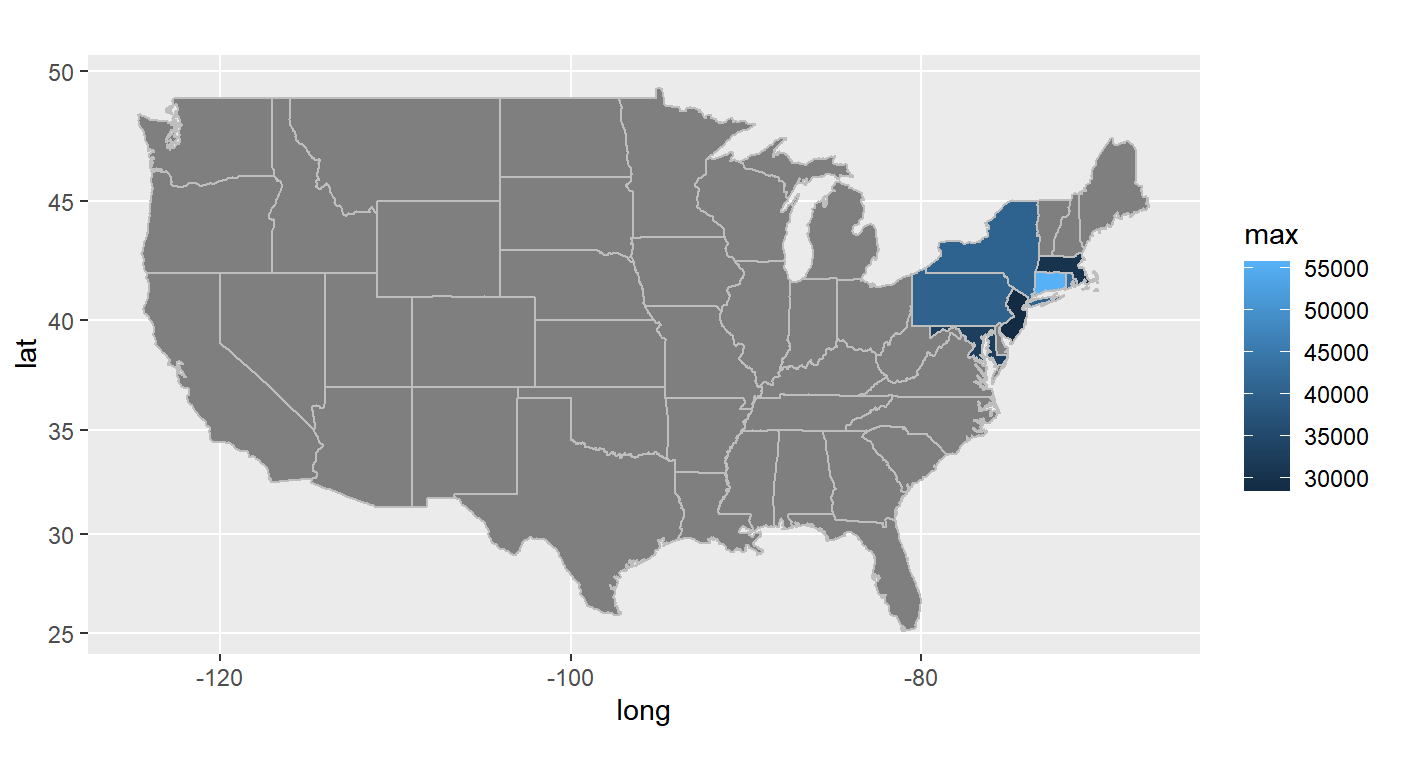
We keep the remaining 7502 observations as our original data to analyze. Moreover, we will not do any further modifications, for now, to remain intact. Any further modification of the data frame will be stated in each data model.

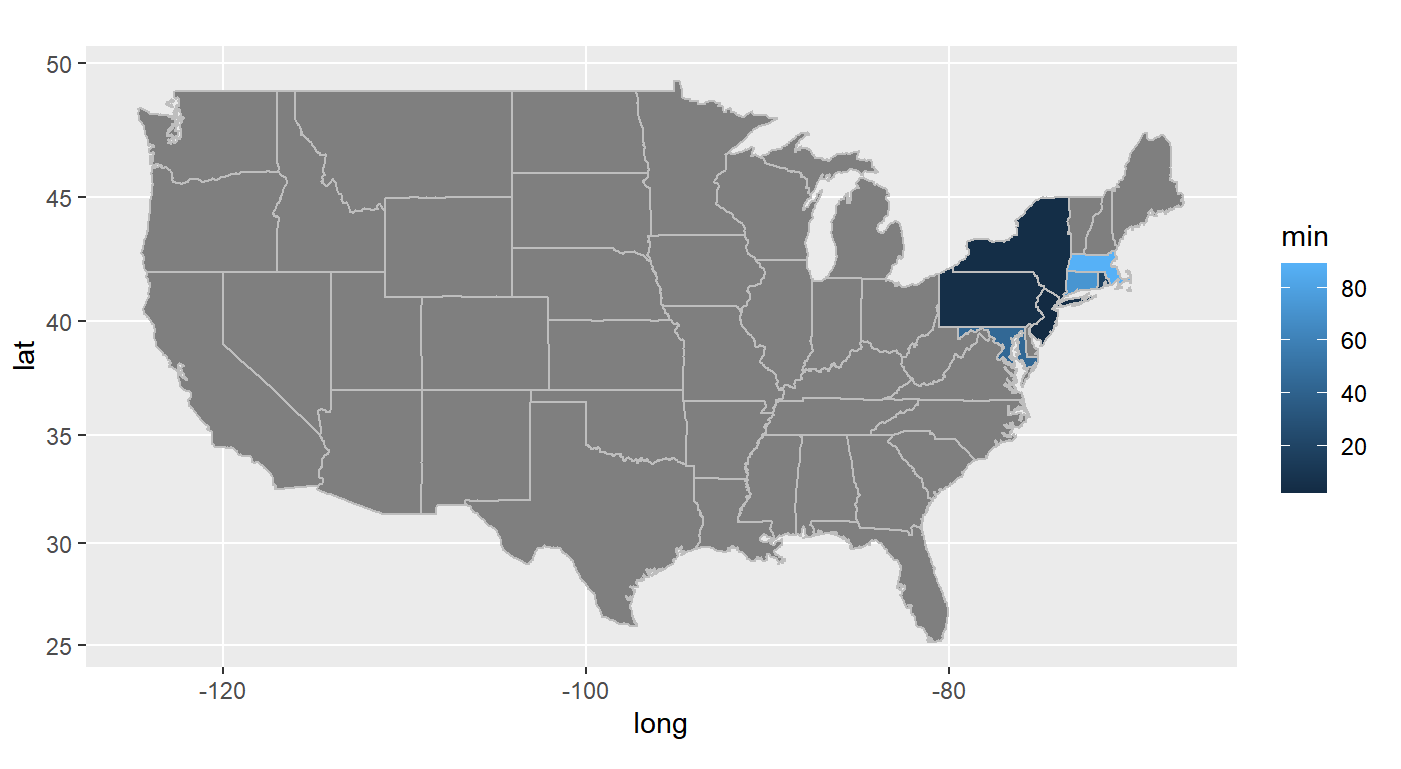
**Data Exploration**

From the following map, we can know that this range of dataset is only between 7 states (Connecticut, Rhode Island, Massachusetts, Pennsylvania, Maryland, New Jersey, and New York)

We can see that New York State has more litter color than others, which represents New York State has a higher average cost than other states. The average cost in Maryland State and The Constitution State is below 4000.

.

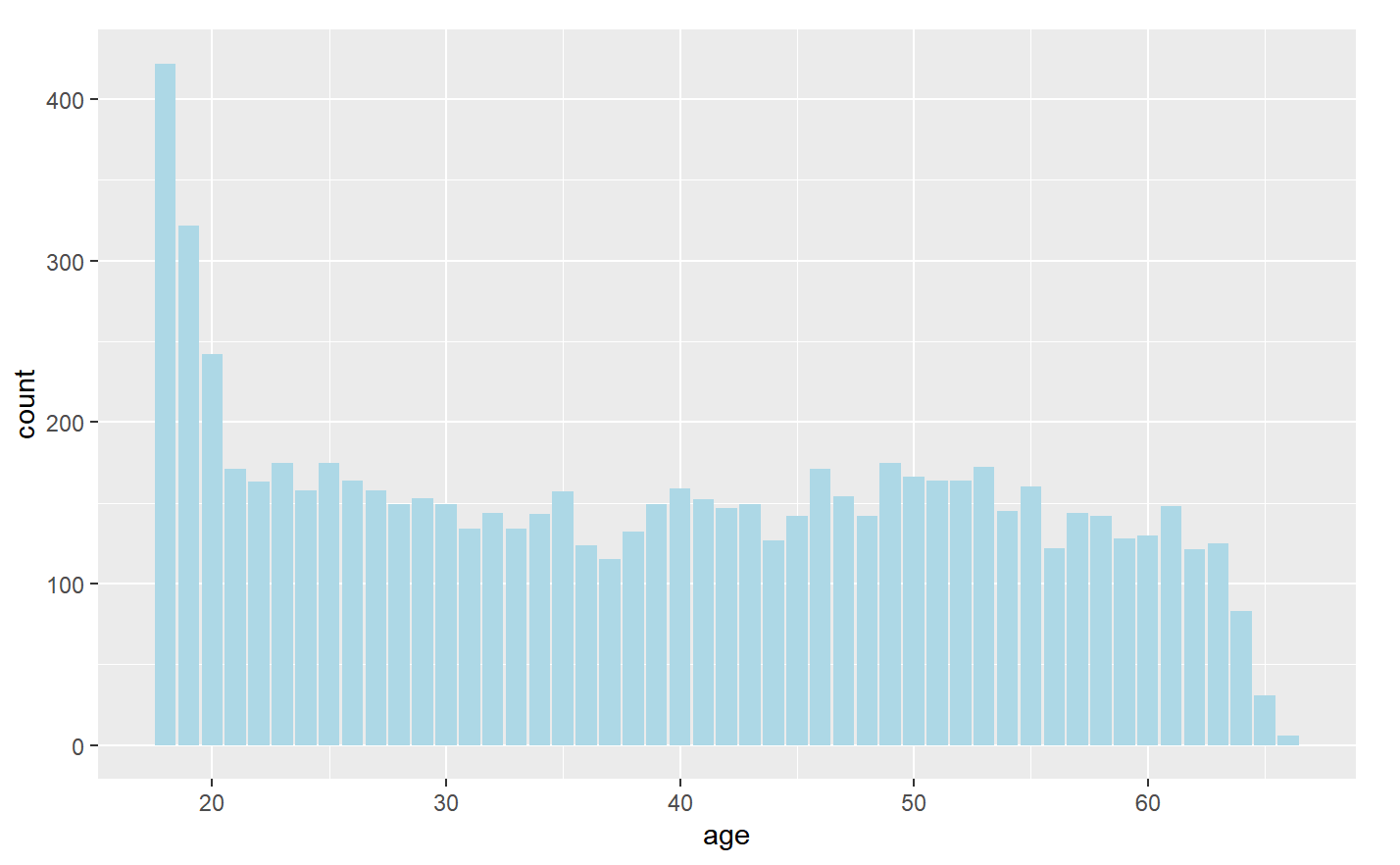




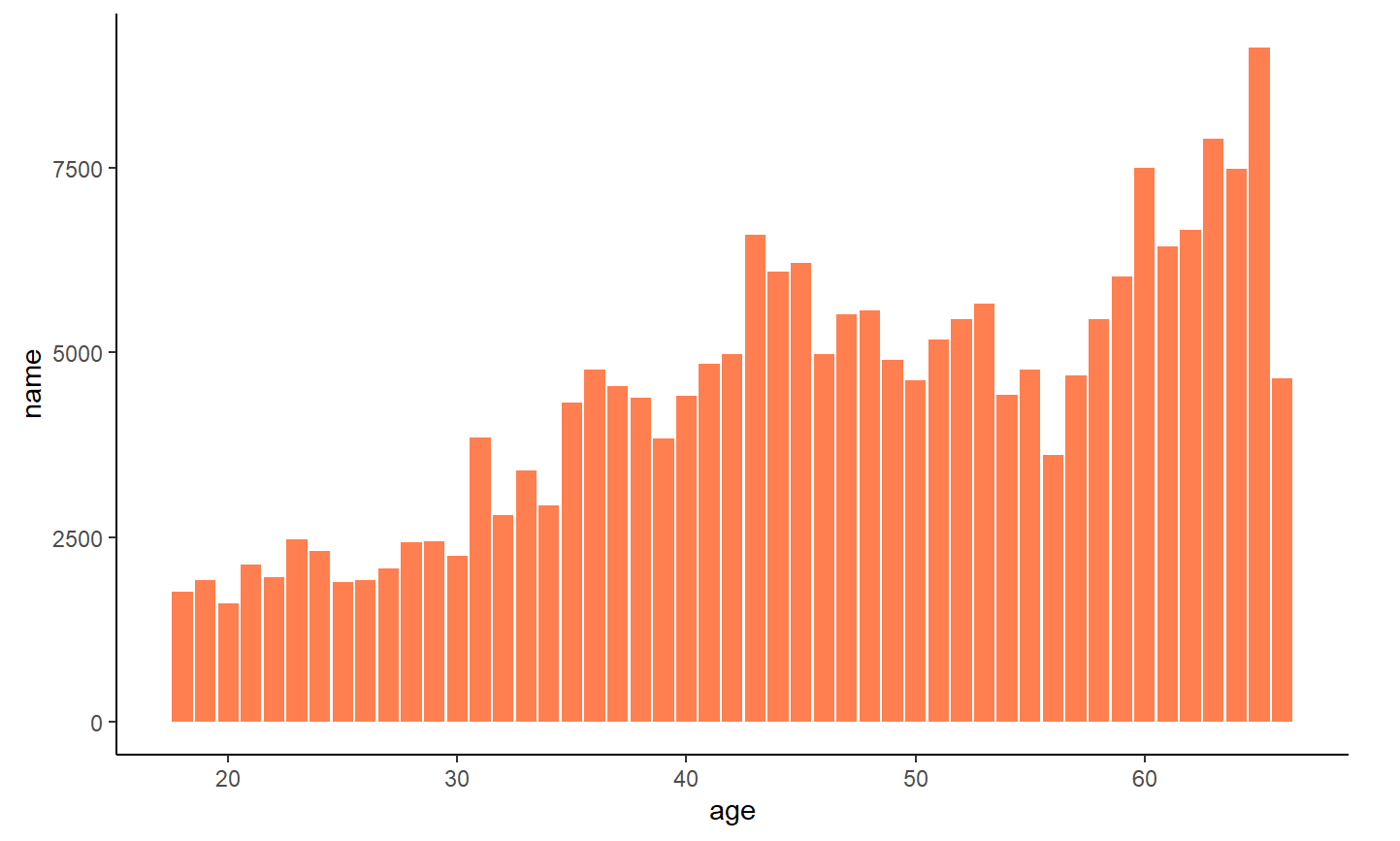
The map above shows that the maximum cost in The Constitution State is over 55000. Pennsylvania and New York State have the minimum costs than other states.

According to this bar chart, this dataset has people of all ages, and they are average. Nevertheless, the age under 20 has higher than other ages. This is an excellent dataset to analyze the correlation

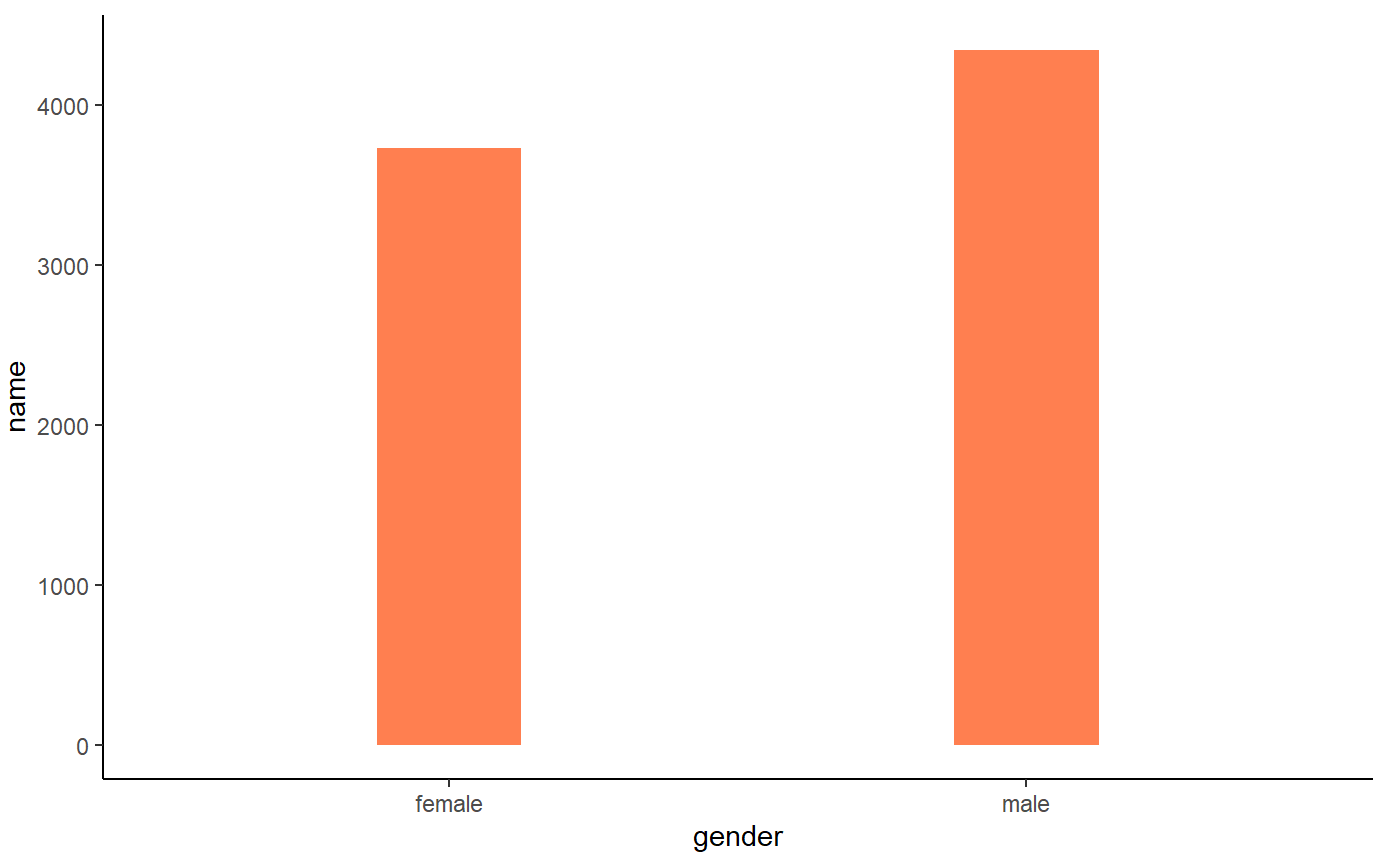
between age and cost.



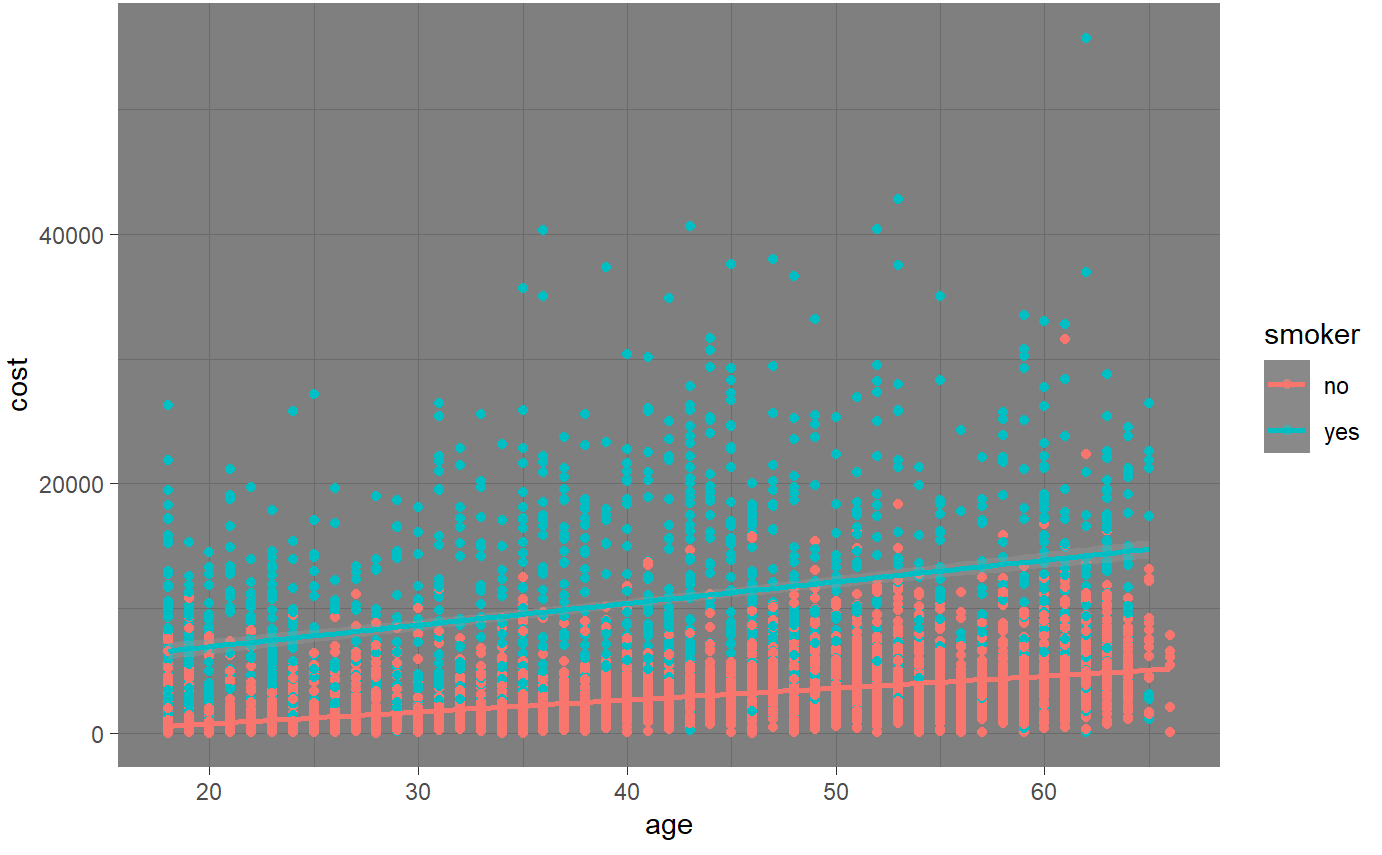
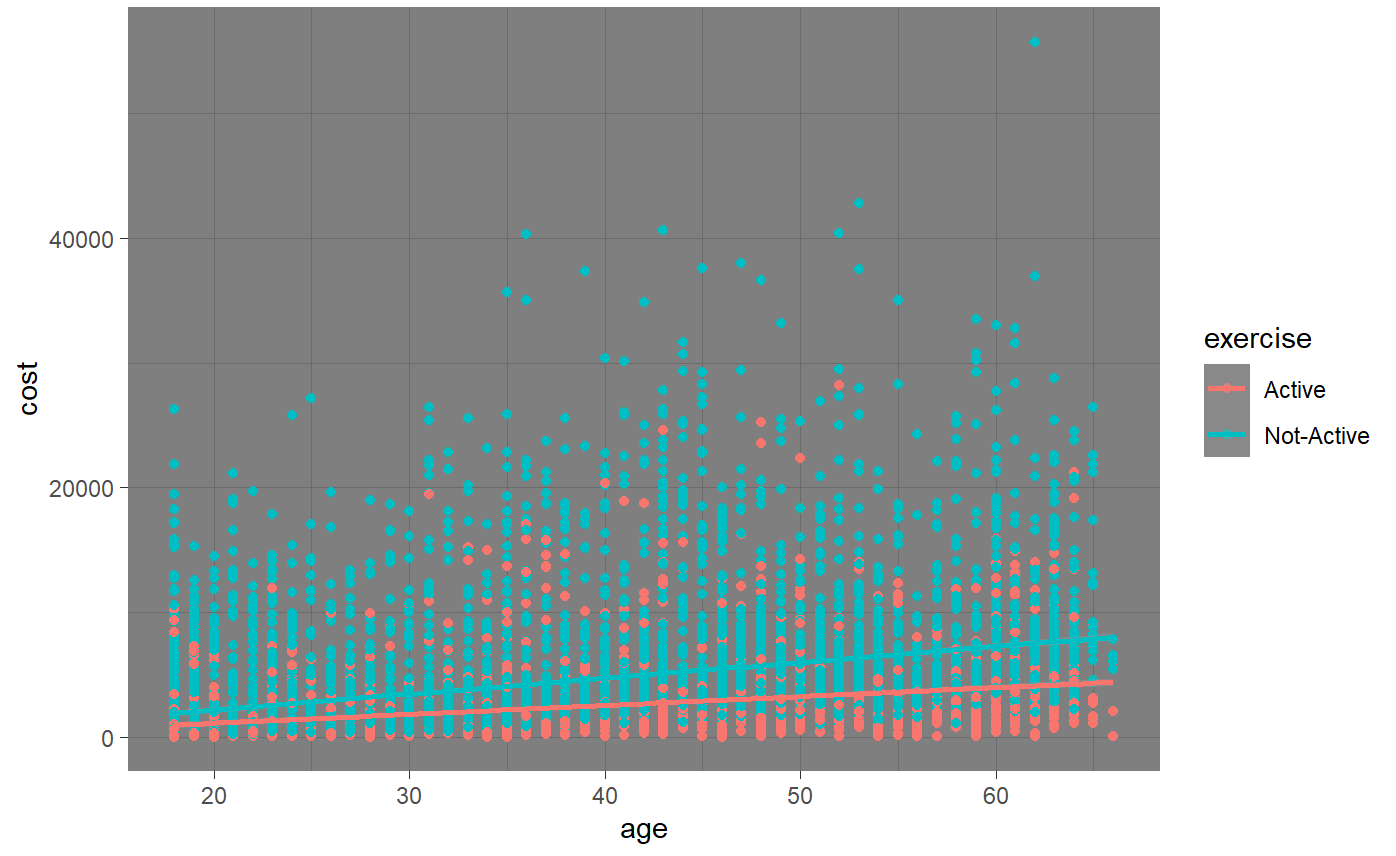
From the below graph, we can see that the more the average age of different ages, the higher the cost.

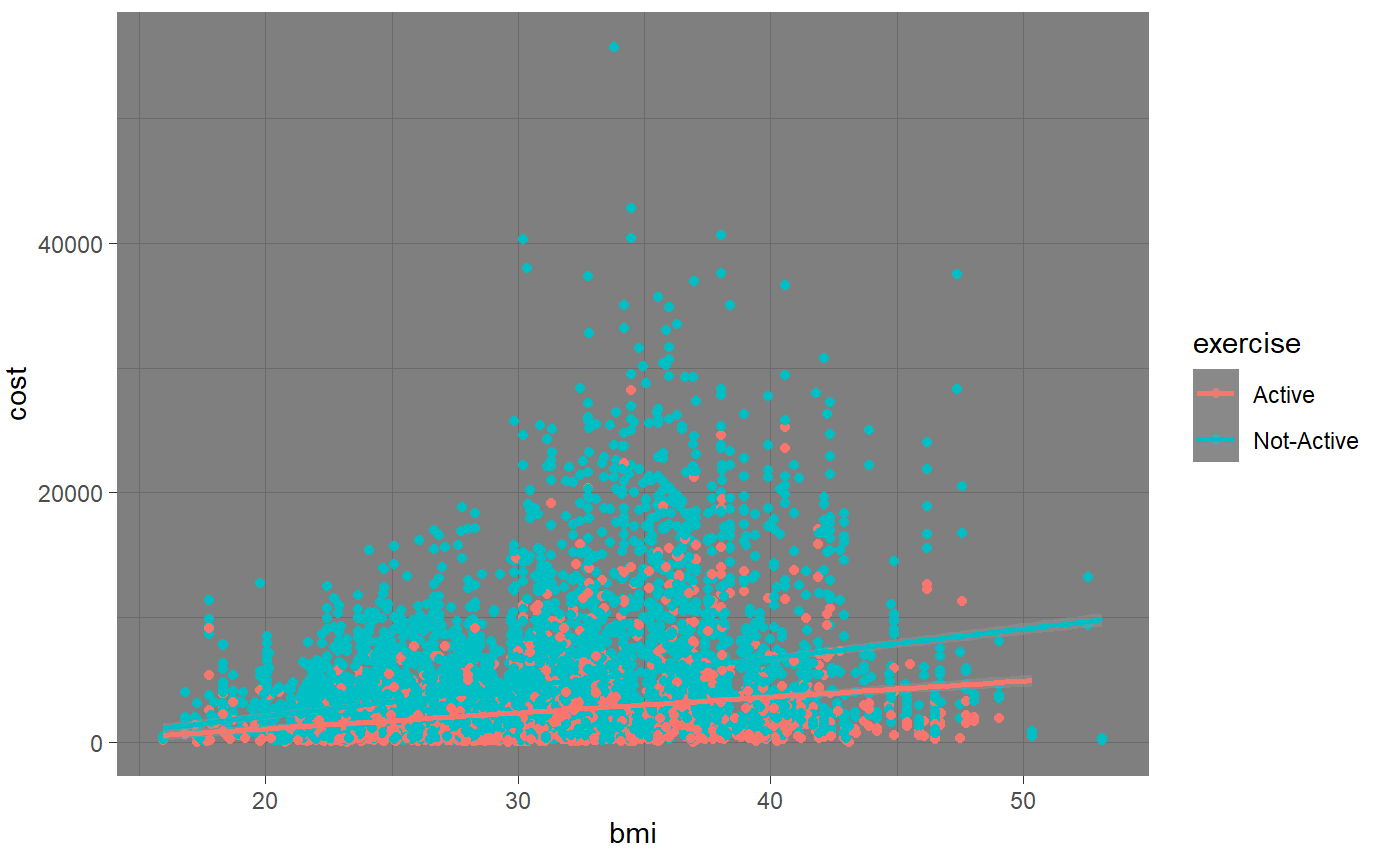
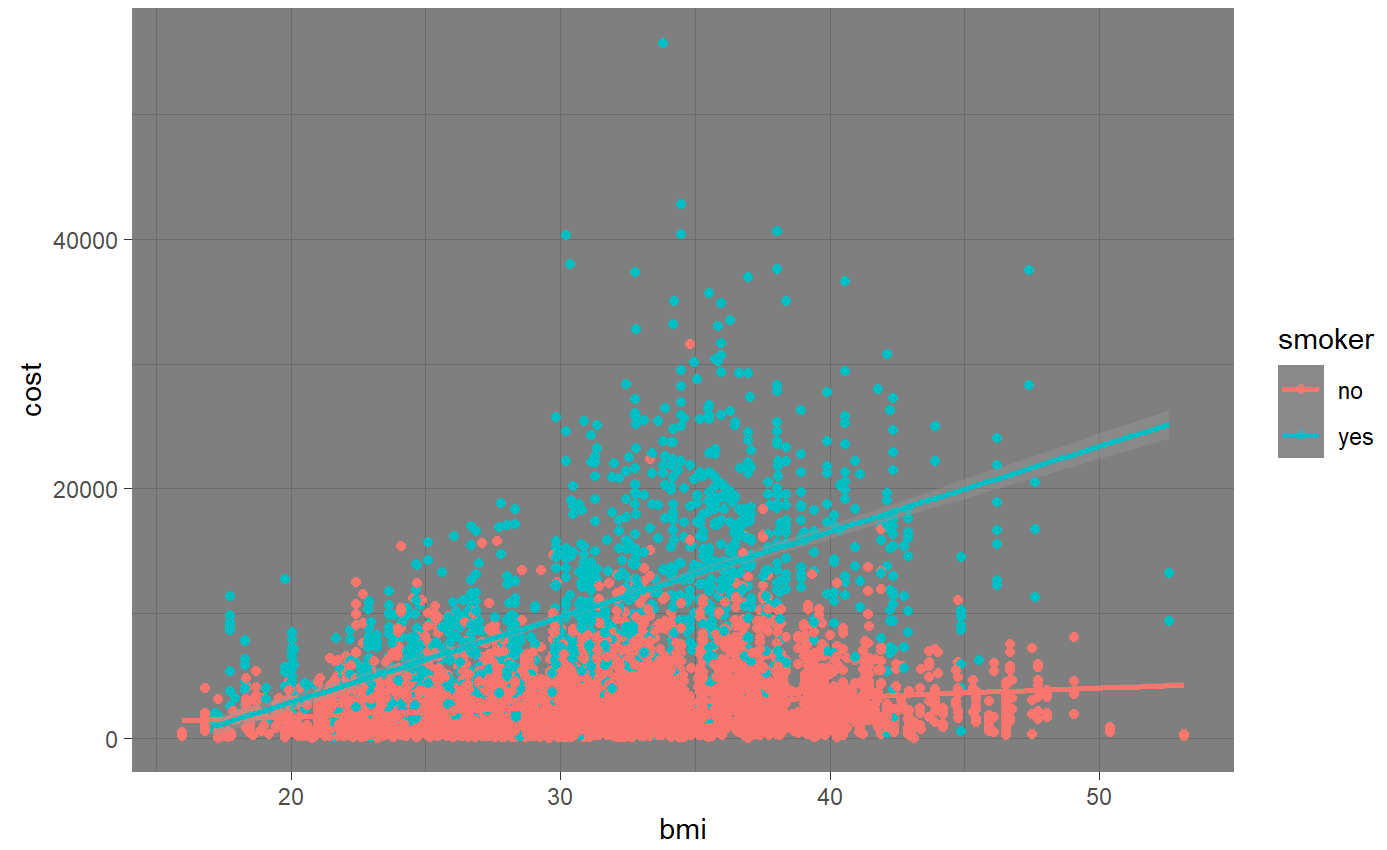


From the gender, the male has more cost healthcare.



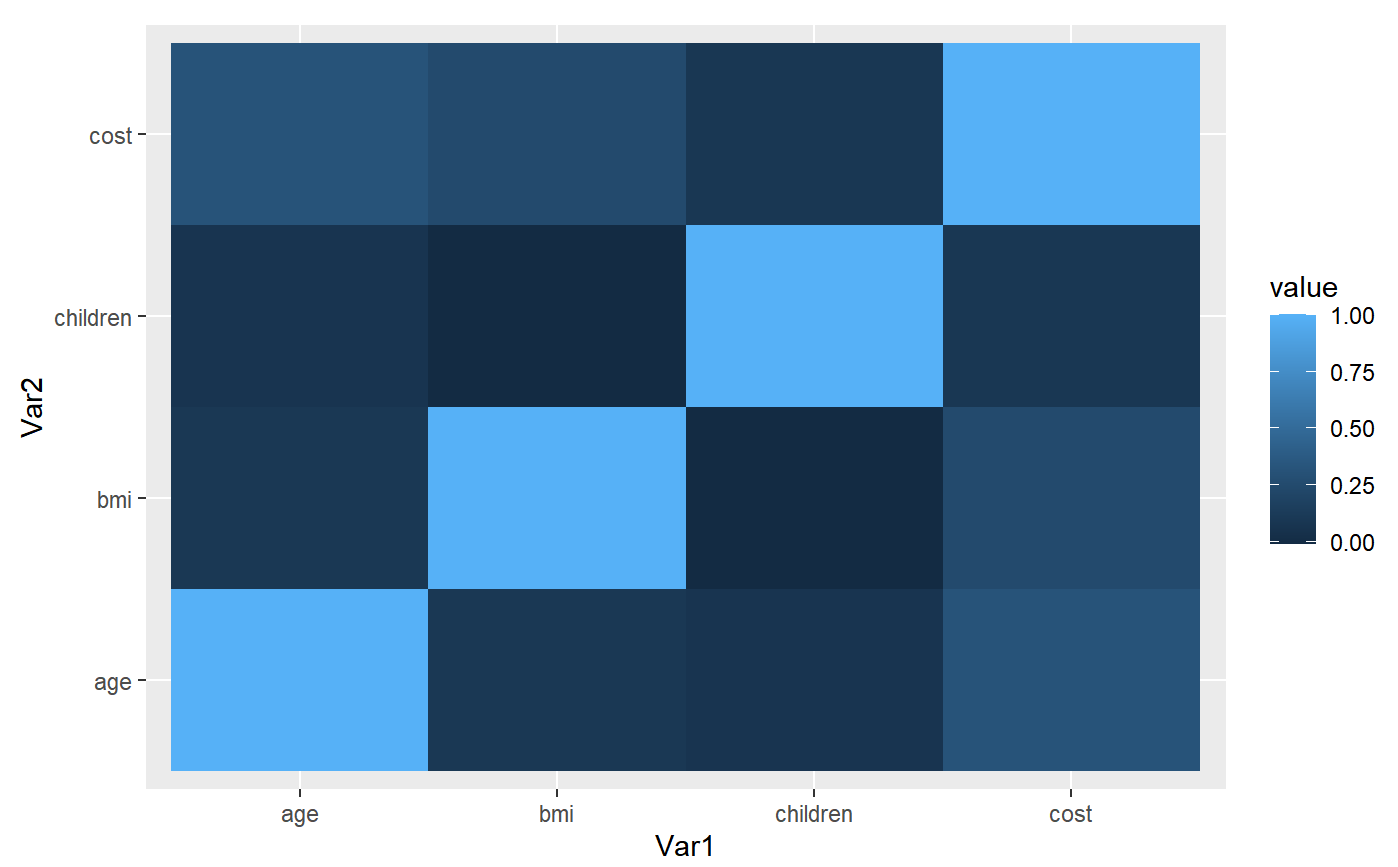
Then, we visualize the 'age' and 'bmi' attributes with cost based on 'smoker' and 'exercise' as a scatter plot to understand the correlation of cost with each attribute.

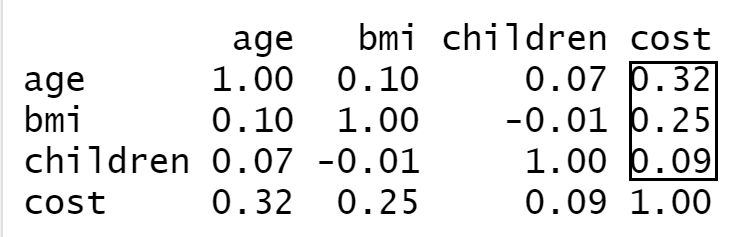
 



It is quite evident that 'age' and 'bmi' positively correlated with a customer's healthcare cost.

We want to quantify the correlation in terms of a correlation value. This can be achieved with a correlation table and a heatmap of attributes.



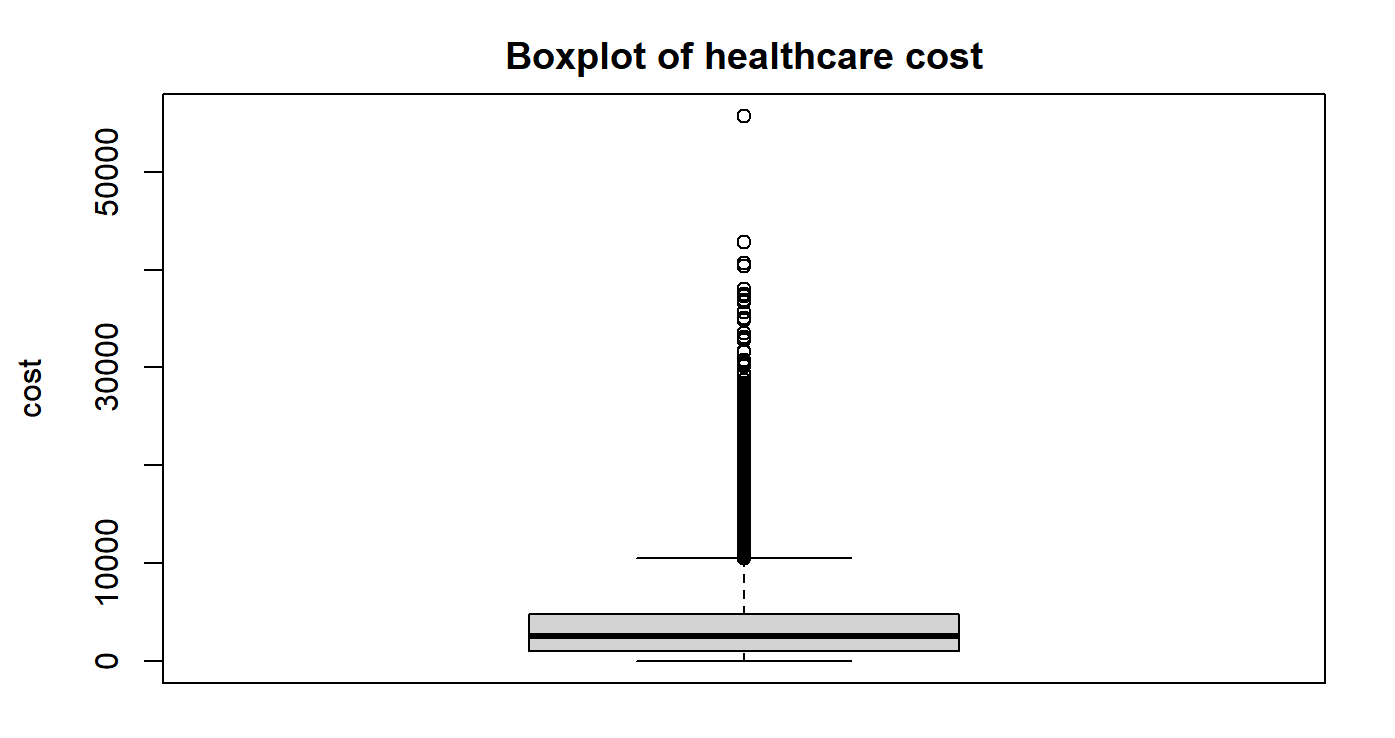


From this correlation heatmap and correlation matrix of numeric parameters, we could see that cost with 'age' and 'bmi' have a lighter color, the correlation coefficient of age is 0.32, and the correlation coefficient of 'bmi' is 0.25, which means age and 'bmi' have a positive correlation with cost. Moreover, the cost with children has a dark color, and the correlation coefficient is 0.09, which means children have a weak positive correlation with cost.

**Definite Health Care Cost : Categorizing Expensive / Not Expensive**

To decide the cap cost where it becomes expensive for the health care company to cover health insurance, we need to look at the cost spread for the data available to us.

- distribution of cost



From the boxplot, we found that the range of this dataset is between 2 and 55715. The 25% cost of this dataset is 966.50, 75% cost of this dataset is 4778.75. The median cost of this dataset is 4778.75. Furthermore, the average is 4049.492.

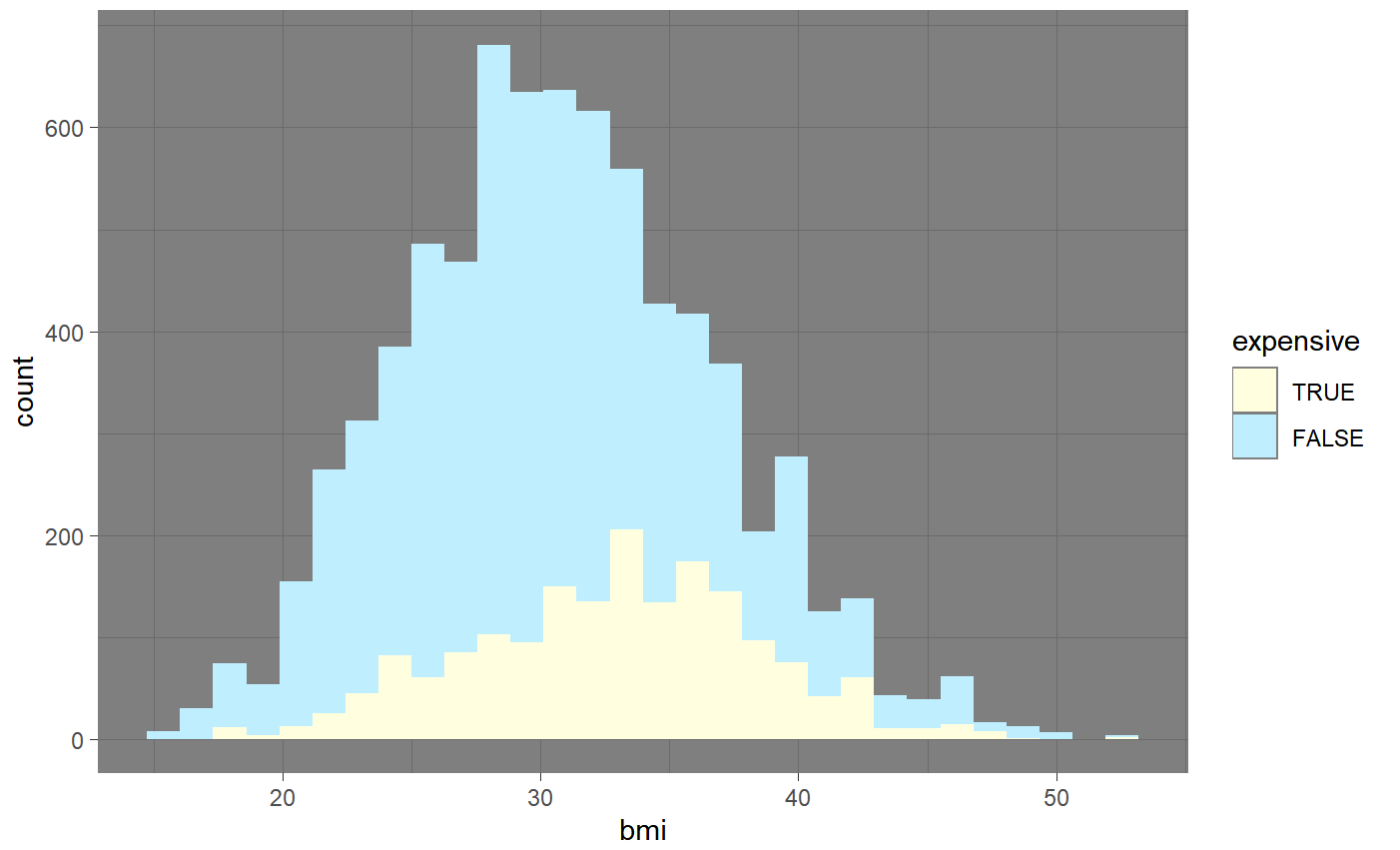
After considering the quantile study and statistical values of the cost columns, we want to decide the cap cost at the 75th quantile or $4778. The cost values increase exponentially after the 75th quantile, and the mean cost is 4049. We could have decided on the cost cap at the 70th quantile, but considering little flexibility for borderline customers and the high quantile jumps post the 75th quantile, we made this decision.

Therefore, we defined that if the cost is more than 4778, it belongs to expensive. If the cost is less than or equal to 4778, it belongs to not expensive.

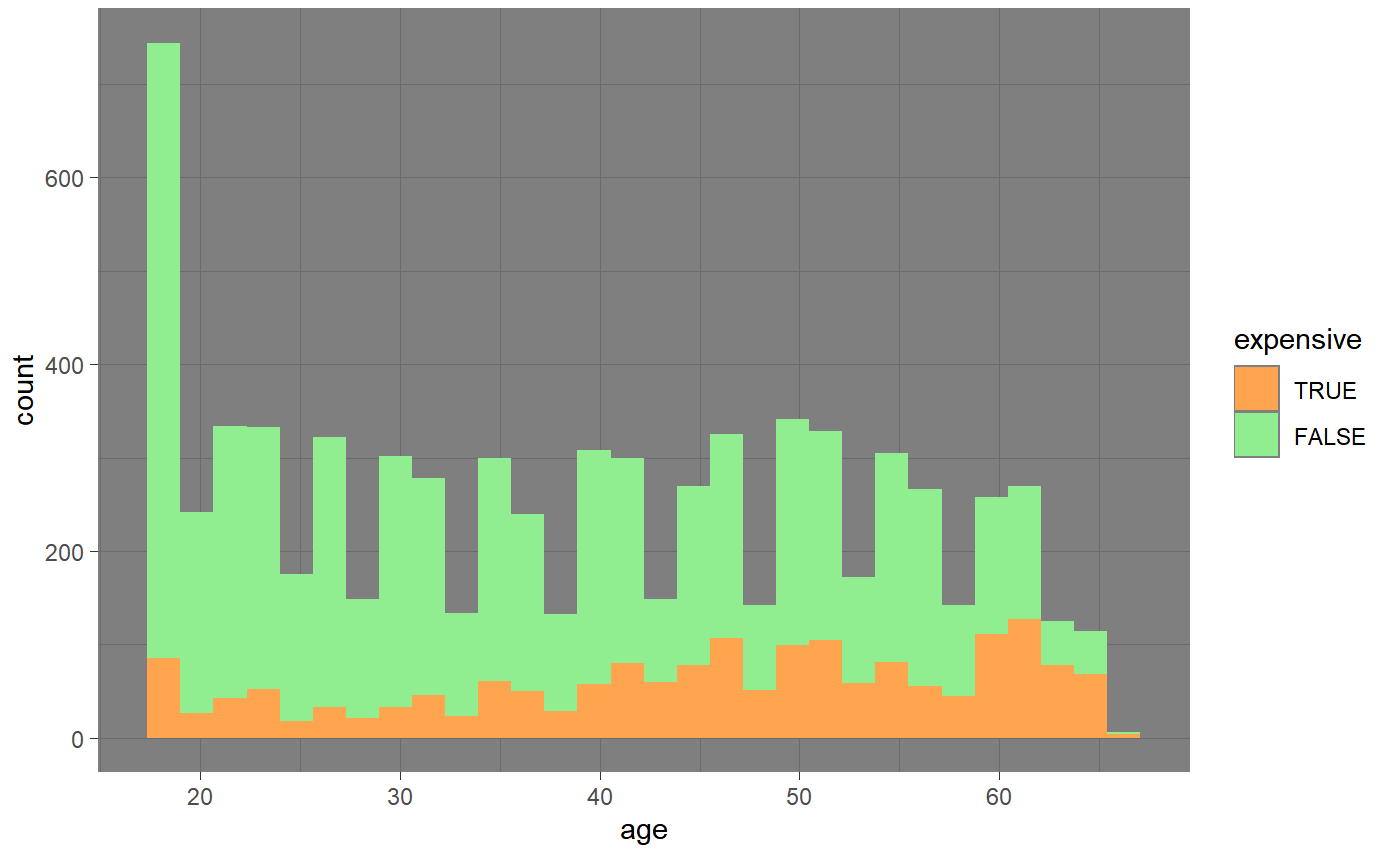
**Overview of important variables**

Let us look at the two groups of expensive and non-expensive customers. We will look at plots for the two types of customers:

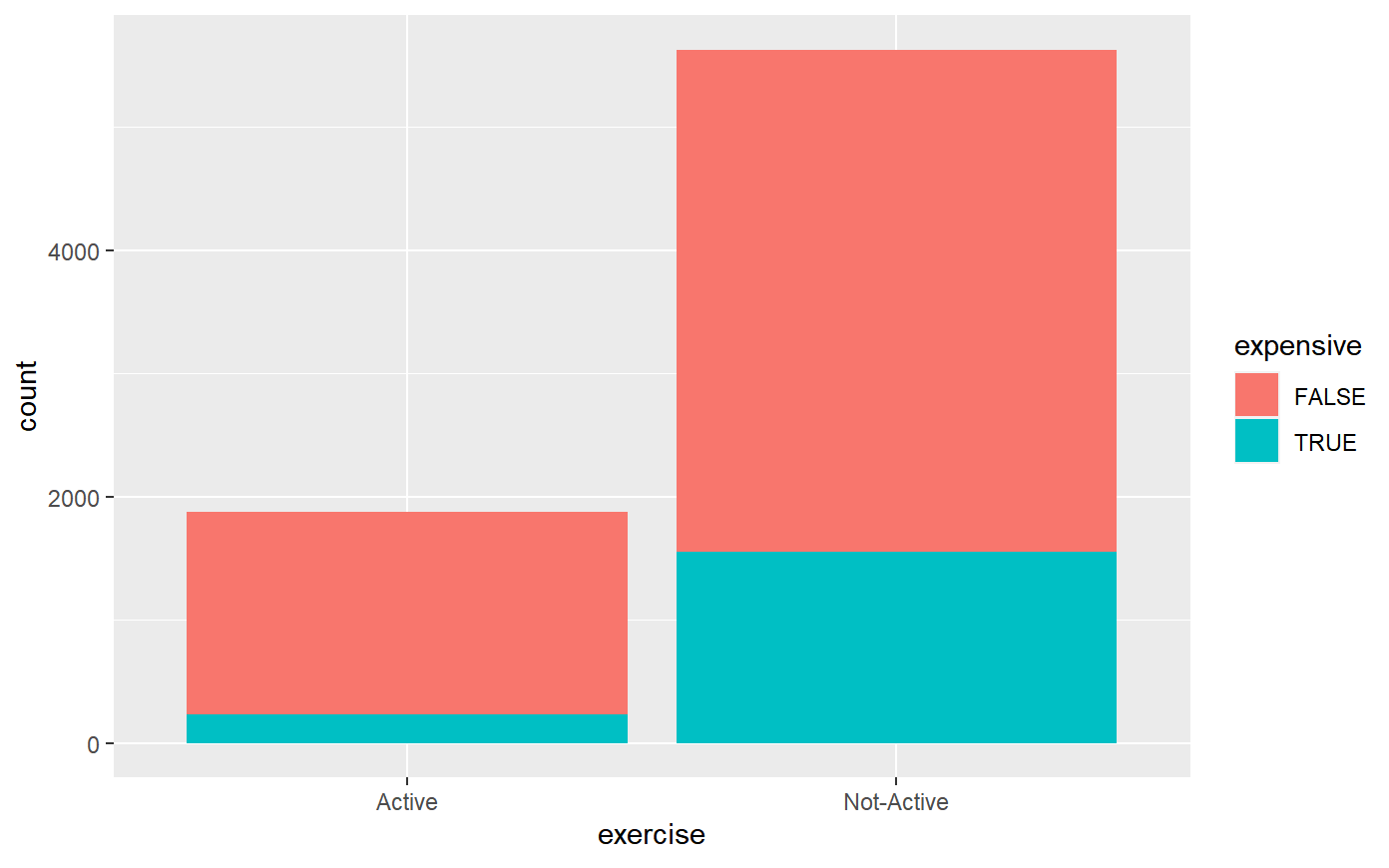
- bmi: We can find that no matter whether it belongs to expensive or non-expensive, they all exhibit a normal distribution. Since this graph is like a “bell”.



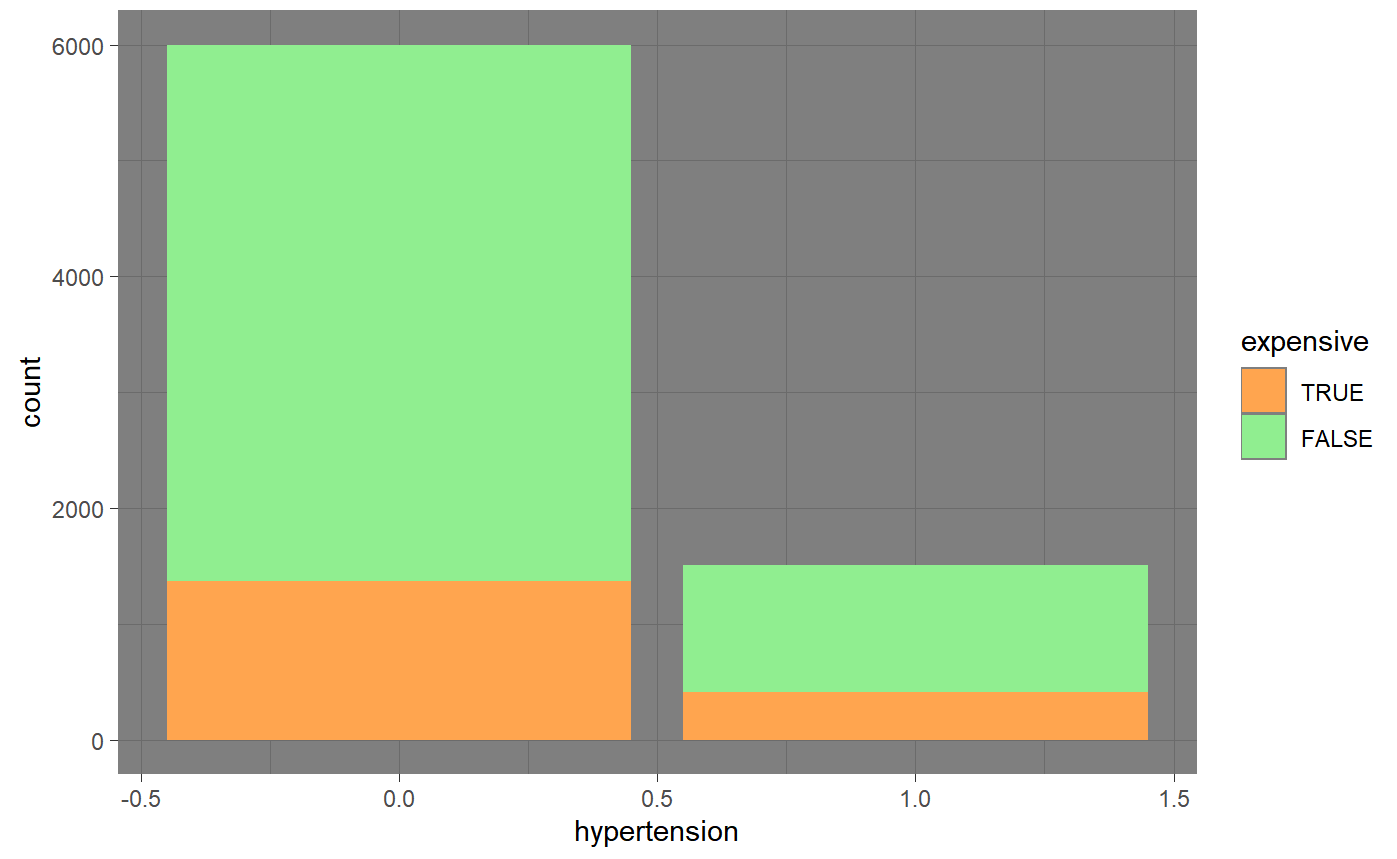
- age: Evenly distributed in each age group.



- exercise: In the Active, expensive occupies a smaller proportion than Not-Active.



- hypertension: Expensive almost occupies the same proportion in the non-hypertension and hypertension.



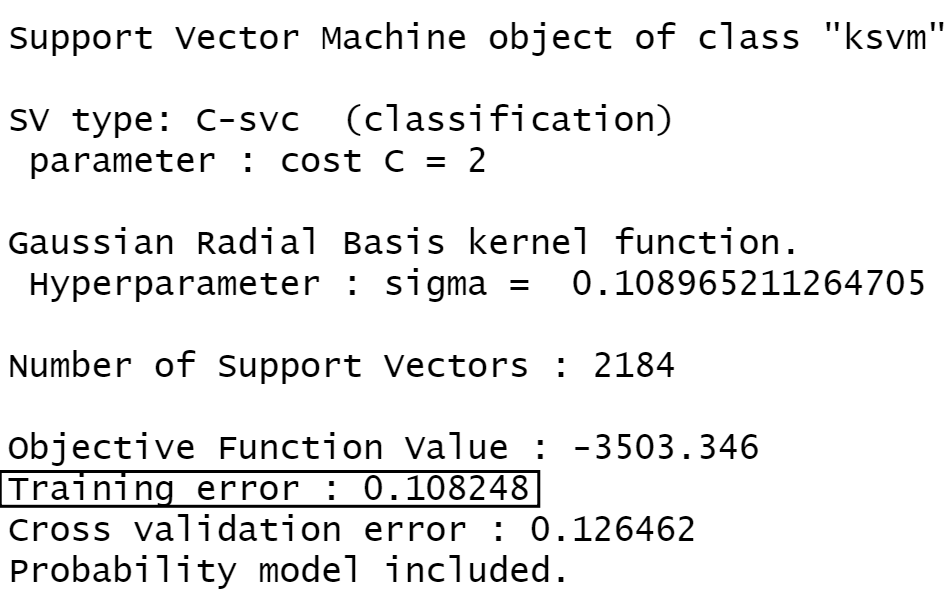
**Data Modeling and Trends**

**SVM Model:**

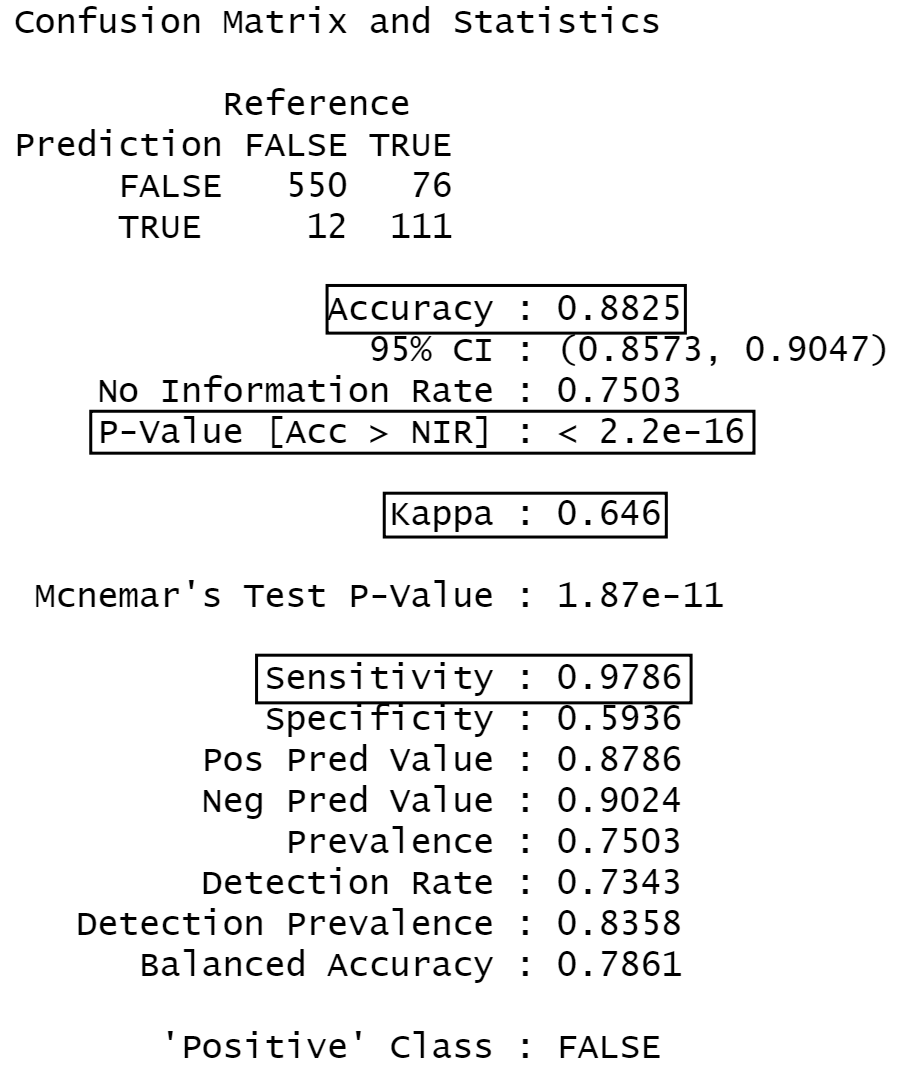
After cleaning the original data and exploring the data variable, we will create a Support Vector Machine to predict whether a customer is Expensive or Not Expensive for the healthcare company.

Hence, the model's dependent value will be "cost," and independent values or predictors will be the rest of the columns. We randomly select 90% of data from this dataset as train data and 10% as test data.

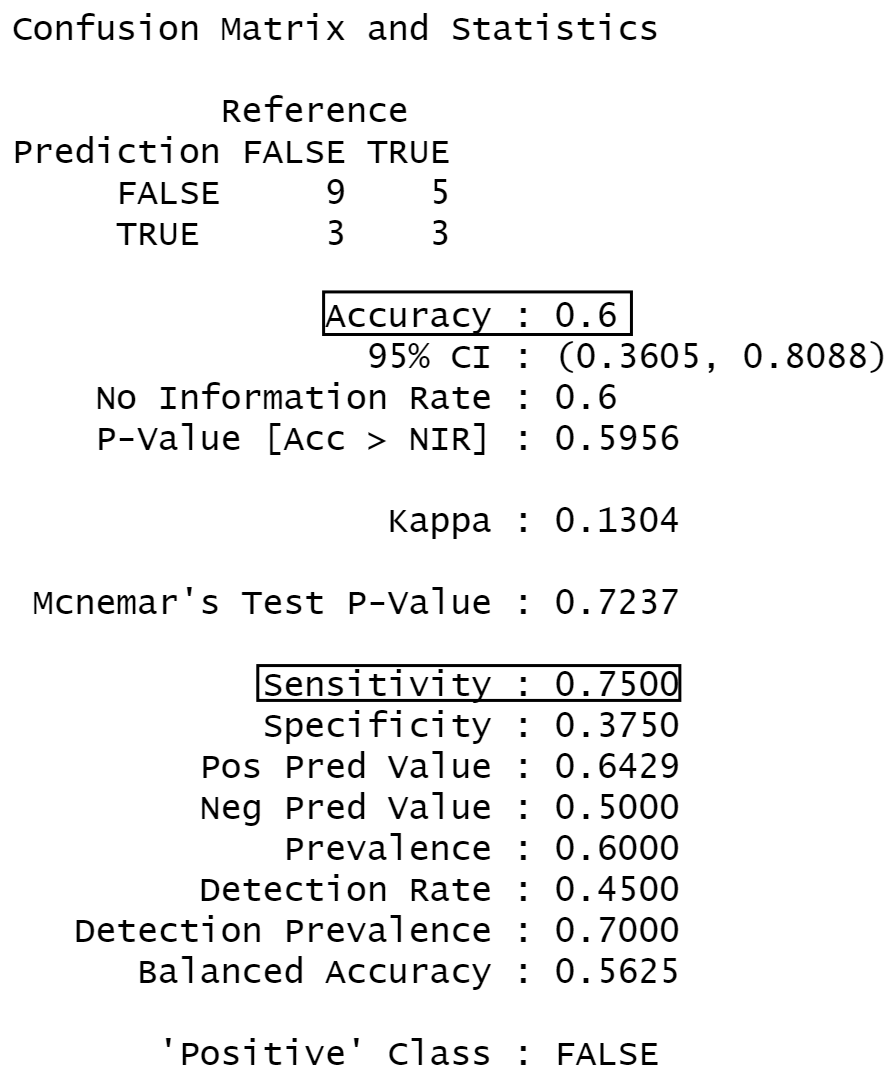
From the below picture, after declaring an SVM model with defined cost and cross-validation parameters, the training error of 10.08 % proves the training accuracy of 89.92%.



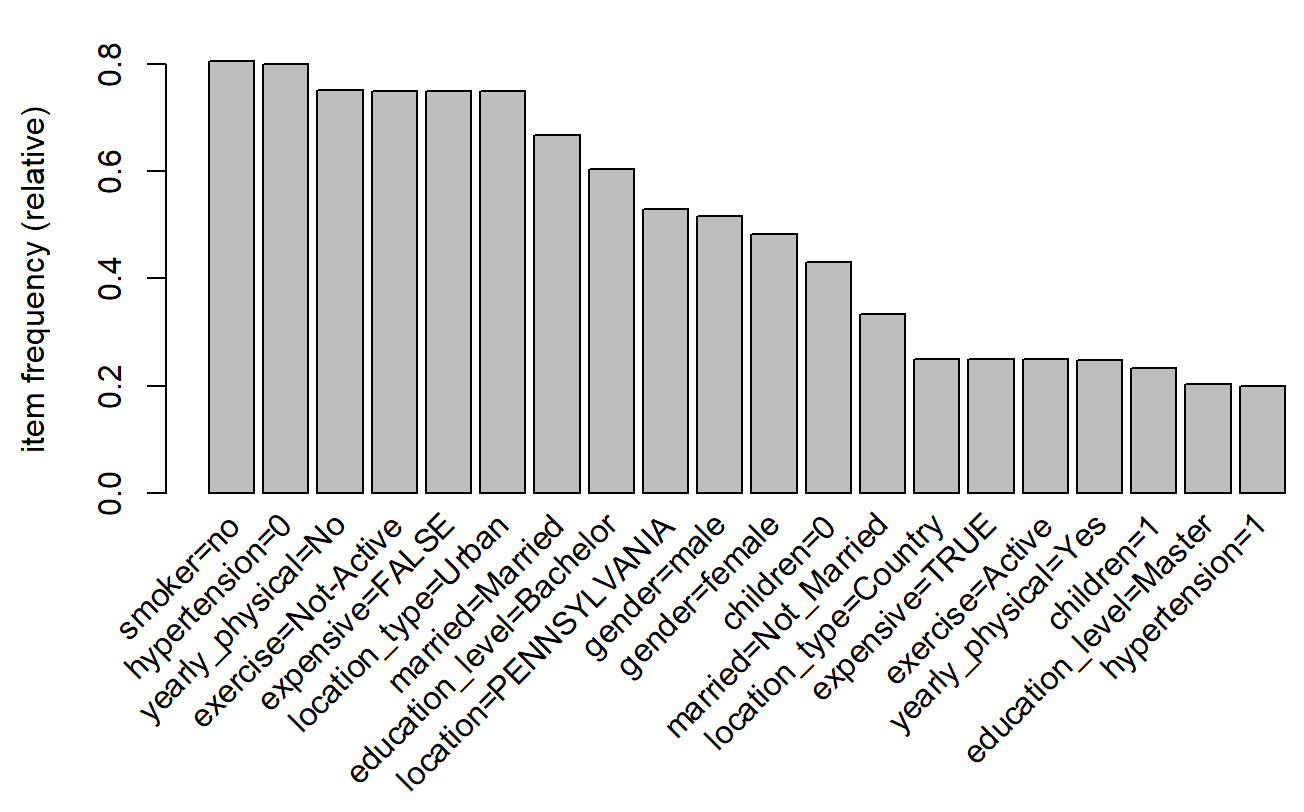
We use test data to predict cost classification. The result shows that the predicted accuracy is 88.25%, the sensitivity is 97.86 %, and the kappa is 0.646. The p-value is less than 0.05, proving that this model is convincing.



We also use HMO\_TEST\_data\_sample to have a second test. The accuracy is 0.6, and the sensitivity is 0.75. It is enough to prove that this model is convincing.



We also find trends in why health care costs of a subset of customers are expensive. For this purpose, we will explore Association Rules to define rules leading to expensive costs. The below graph is the best rule. We can see that the first rule is no smoking and no hypertension.

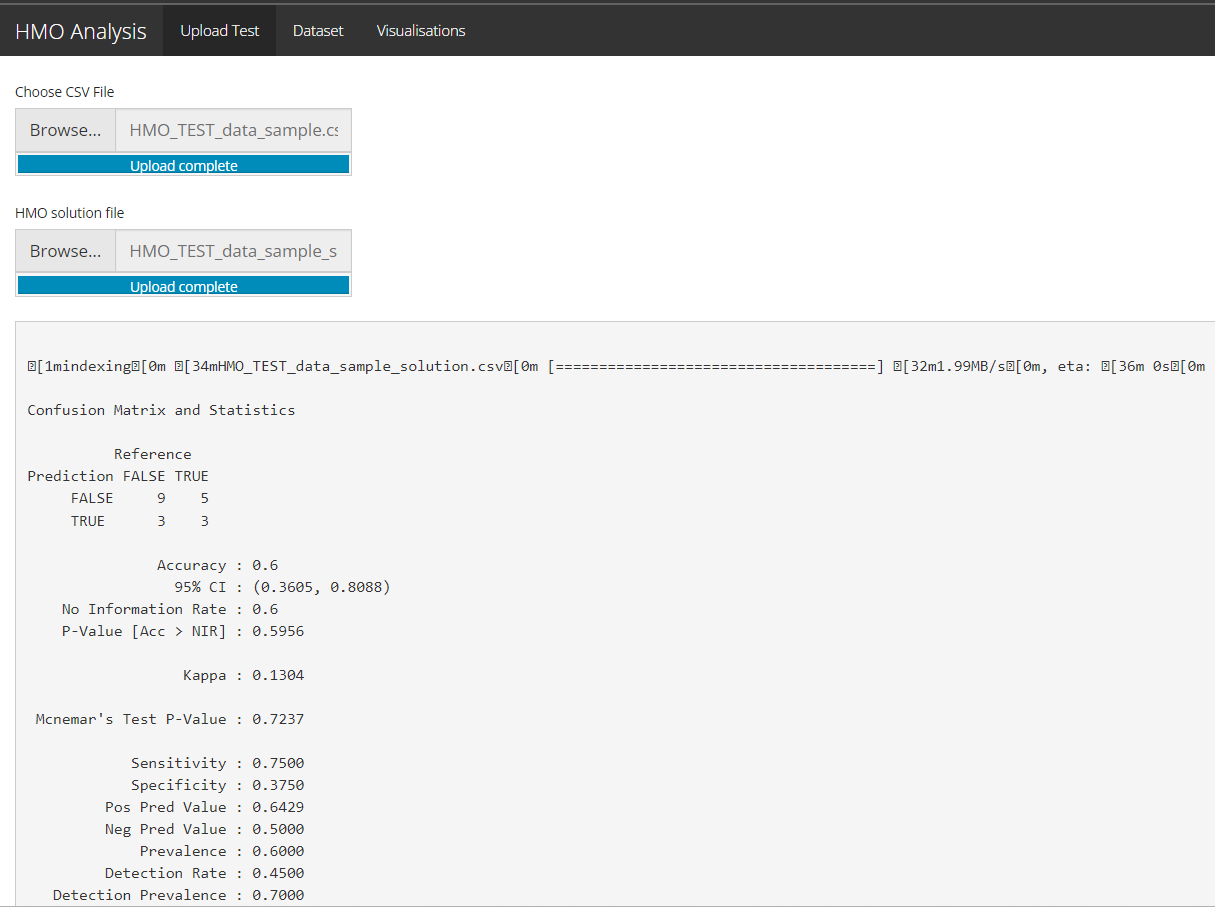


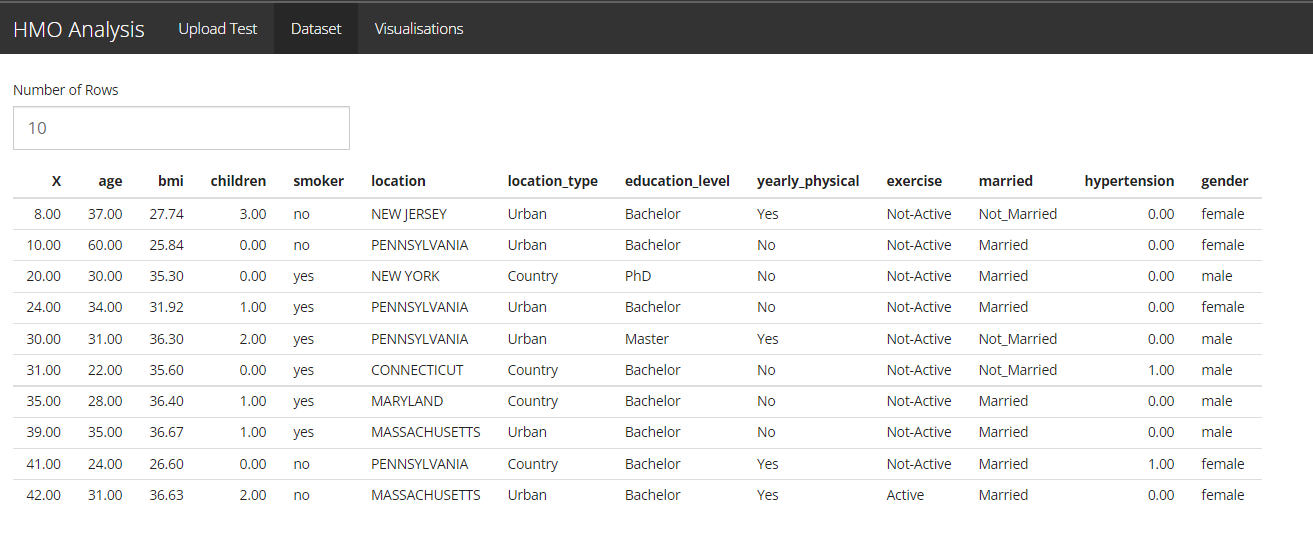
**Shiny App**

Shiny is an R package that makes it easy to build interactive web apps straight from R.

In this project we have made a single R file which contains both the server and the UI code.

The shiny app consists of 3 tabs and the theme is set to “yeti”



1. Upload test- In this section we are taking 2 files as input which are test file and solutions file. After that we are displaying the model statistics which includes the accuracy and sensitivity of the file
2. Data Set- 

This portion displays the test file in a table and we can adjust the number of rows to be displayed

1. Visualizations-

In this tab we are displaying various EDA charts such as scatter plots, box plots and a map

**Conclusion and Recommendation**

According to our model, A person will be expensive when he or she has the following characteristics: does not smoke this year, without hypertension, had a bad visit with his or her doctor, does not exercise regularly, live in an urban, married, the education level is bachelors and so on. According to the rule, we give the below recommendation:

* Anti Smoking programs.
* Promoting regular health checks up for senior citizens
* Promoting a fit lifestyle and controlling weight (marathons organized regularly in counties).
* Developing communities that support safe and accessible places for people to walk, bike, and be physically active.