IST 687

Introduction to Data Science

Submitted by:

Yuktha Lakshmi Arcot Badarinath

Health Care Management Organization

A close-up of a calculator and a calculator

Description automatically generated with medium confidence

**Abstract**

In the field of healthcare, costs are a significant factor, and some patients may present a greater financial strain than others. There are a lot of factors that go into making one customer more expensive than another, and if a company can figure out why some customers are more expensive than others, they will see a significant increase in their bottom line. The HMO is able to set the premium plans up correctly, which allows for more effective management of the firm. This document presents our research into leveraging data science (exploratory data analysis as well as prediction models) to anticipate how and why some people have higher bills and deliver actionable insights to Health Management Organizations to help them make better decisions. The research was carried out using exploratory data analysis and prediction models.

**Dataset & Variables**



The data set we are using consists of 7582 rows and 14 columns, details about the people were given in each of these columns such as age, location, gender, etc.

Text, letter

Description automatically generated

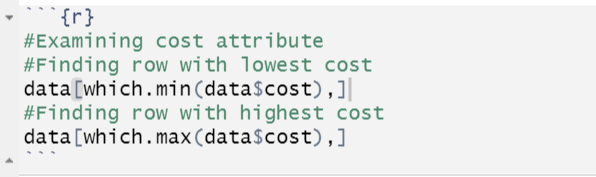
**Process**

We began by retrieving the information from the specified URL, and then we put it into the variable denoted by the letter 'DF.' This is the dataset in its raw form. The first thing that we did was look through the database to see if there was any mention of a NA for it anyplace. In order to fill in the blanks left by the N/A values in the BMI, Hypertension, and Cost columns, we made use of the interpolation function, which is included in the imputeTS package. In order to finish our exploratory data analysis, we made use of all of the variables (EDA).

A picture containing text

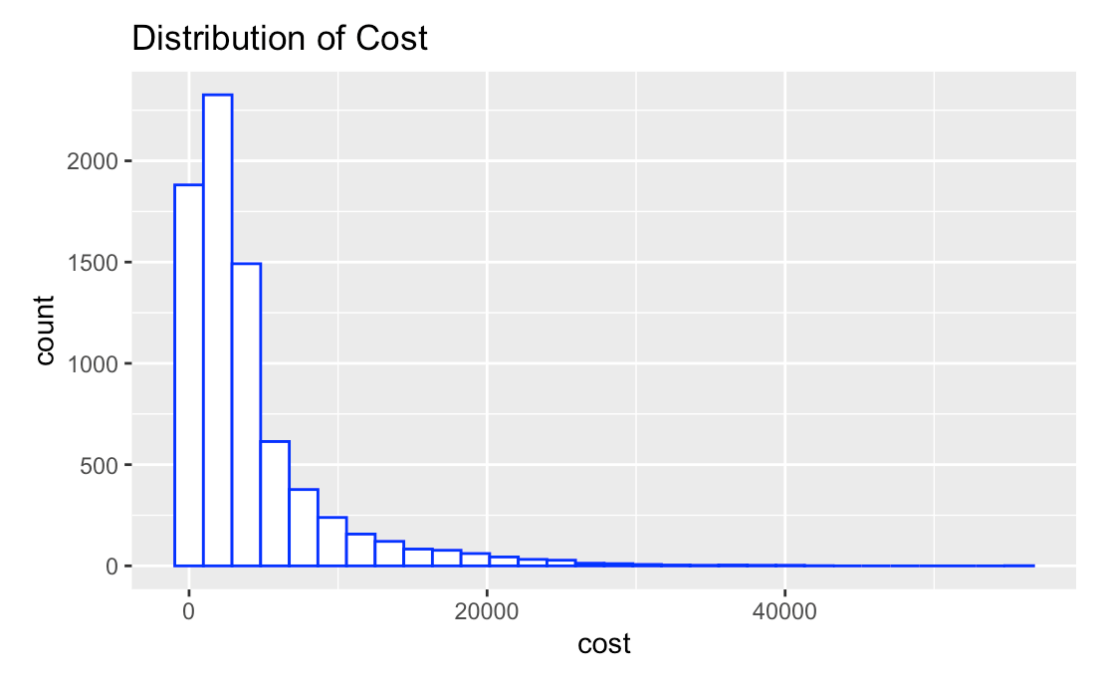
Description automatically generated

As we are trying to find why some people are more spending more than others in the medical costs the cost variable becomes absolutely important and that is the reason we have examined the cost attribute. NEW JERSEY is the state with the lowest cost where CONNECTICUT is the state with the highest cost.



**Exploratory Data Analysis with Data Visualization:**

1. We tried to see the trend of cost over the and the frequency of the cost variable varied from 0 and above 2000. We used ggplot to see the trend of cost variable. It seemed to be skewed towards the right.



1. The trend of the cost was analyzed over BMI by plotting the histogram. The histogram seems to normalize curve reaching its peak at 28 with the frequency of about 700. The BMI has a normalized effect on cost.

Chart, histogram

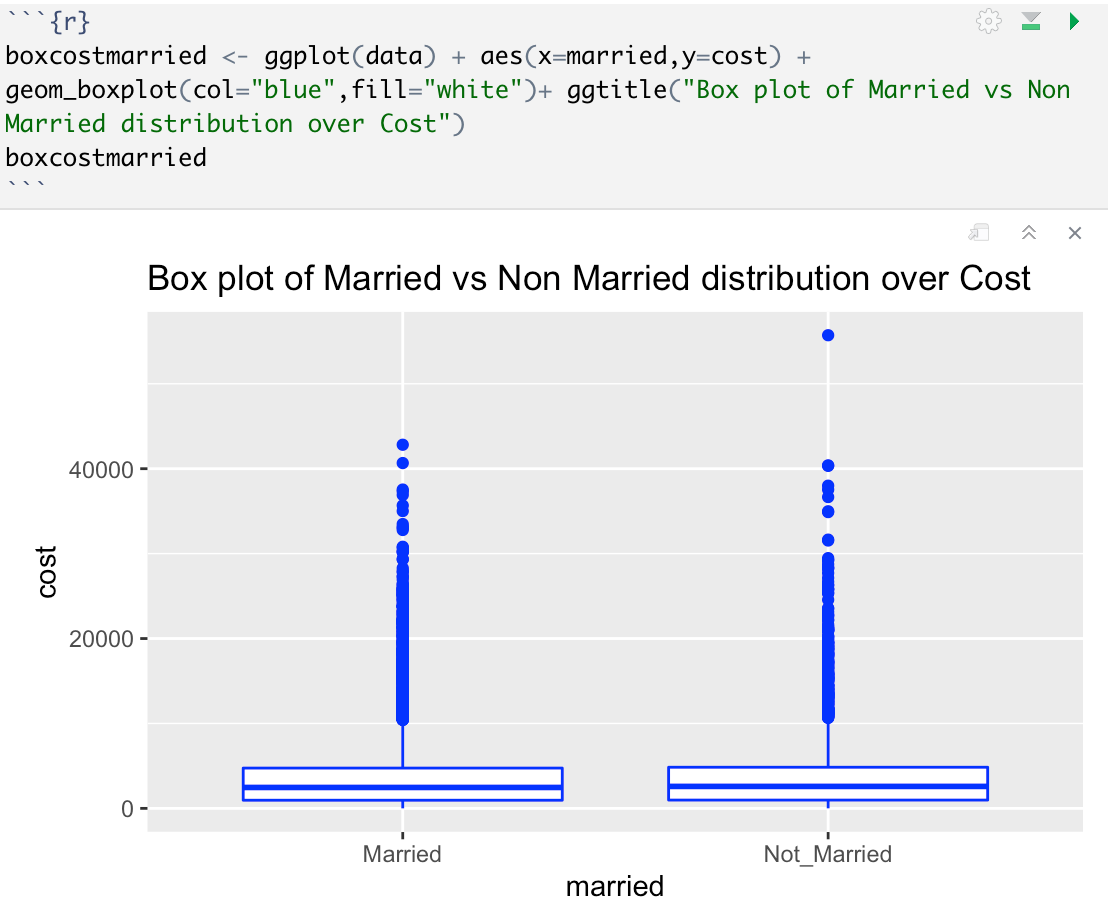
Description automatically generated

1. As the age is categorical variable, we tried to plot a scattered plot using ggplot2 package. Almost all the age group people cost is below 20000. This indicates that cost and age do not have much of correlation between them. The same thing is represented in the graph, where there is correlation between age and cost.

Chart, scatter chart

Description automatically generated

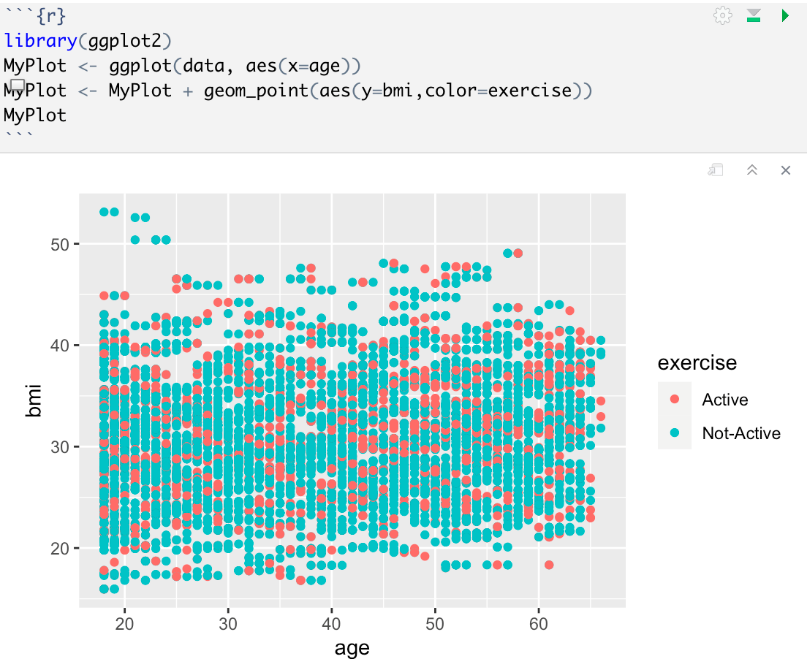
1. To check the dependency of cost whether a person is married or not married, we implemented box plots and we saw that there are many outliers in data, the cost means of both married and not married is overlapping, which further indicates that both category variable accounts the same cost.



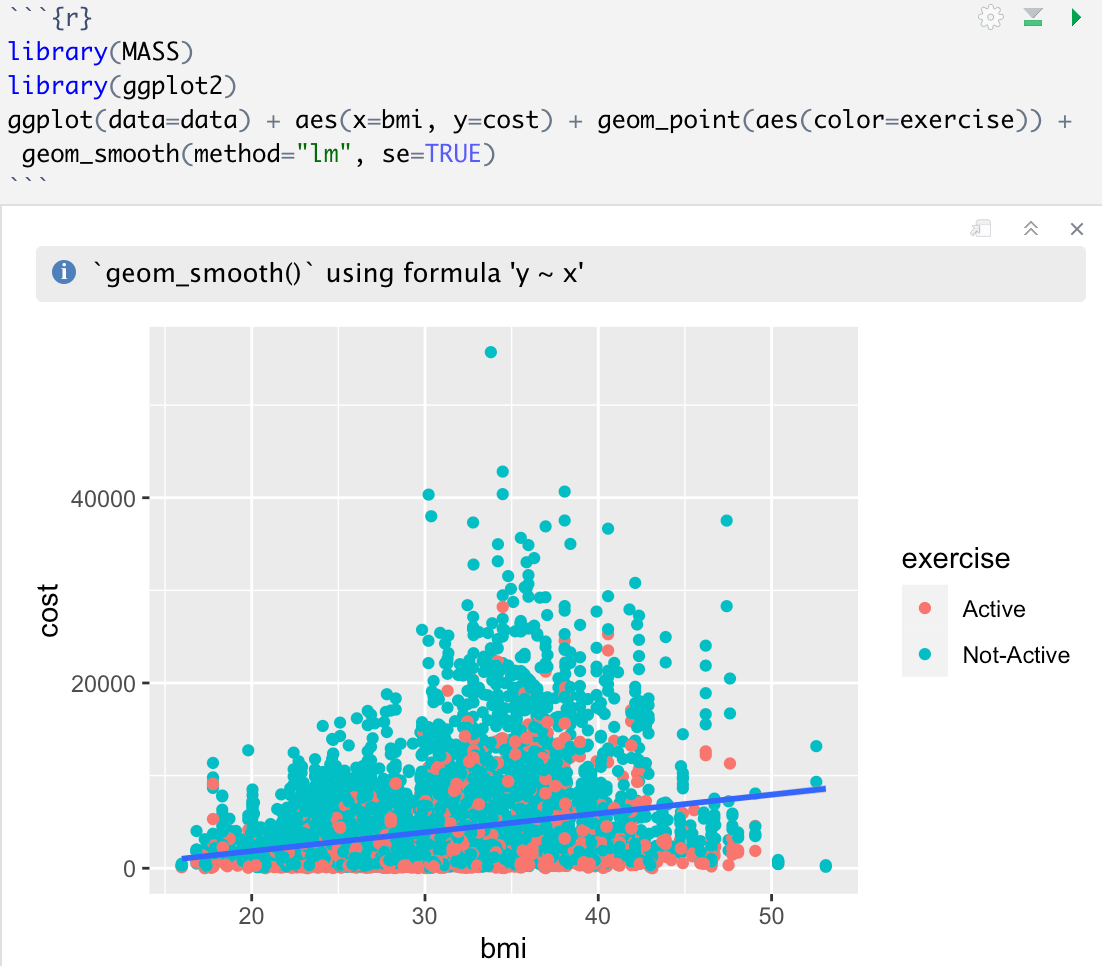
1. To predict the dependency of age and BMI would be expensive or not expensive can be predicted by plotting the scattered plot. By looking at the graph we can predict the as the age increases and the BMI increases, this increases the expenses of the person. The young adults with less BMI do not have any expenses. The plot can be represented in the following way:



1. To predict the if the person is active or not active, we consider two variables age and BMI, where we plotted the scattered plot to determine the correlation between the variables, we could determine that most of the people are not active even though the BMI is. This indicates that there is no correlation between BMI, age and exercise.



1. To predict if the person is active or not- active we consider variables of cost and BMI. There is a positive correlation between BMI and cost.

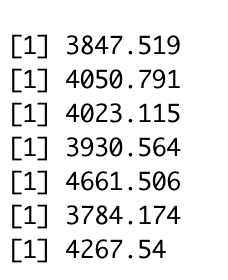


**Data filtering**

Data is being filtered by location, in the data set we have data from 7 states, using the filter function from the Tidyverse package we filtered the data on the basis of each state and then found out the average cost for each state. The state with the highest average cost is New York and the state with the lowest cost is Maryland.

Text

Description automatically generated with medium confidence



**Defining Expensive**

Based on the average cost in each state, we added a new column to the database that we titled "expensive\_type." In this section, we assessed if a person is expensive or not by contrasting the amount of money that a person spends on their medical bills with the amount that is considered to be the norm in the state in which the person resides (average cost of the state). If a person's costs are higher than the average cost across the state, then that individual will have higher healthcare costs. If the individual's costs are lower than the average for the state, then such costs are considered to be lower for that individual.

Table

Description automatically generated with low confidence

The code for the Expensive column is given below:



* When we plotted age against hypertension, we discovered that a substantial percentage of people are at the 0.50 mark, indicating that they are at risk of developing hypertension in the future.

Text

Description automatically generated

Chart, scatter chart

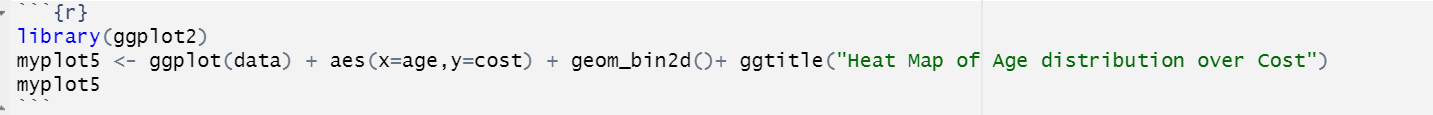
Description automatically generated

* We attempted to plot smokers and nonsmokers against the cost, and it is evident that smokers pay a much higher price than nonsmokers.

Chart, box and whisker chart

Description automatically generated

* We can clearly observe a relationship between cost and age when we plotted the cost versus age. The older the individual gets, the more expensive his medical bills become. Despite the fact that there are some outliers.

Chart, scatter chart

Description automatically generated

We tried to find if there is any relation between the cost variable and the children, based on the scatterplot we can see that people who have children up to 3 are incurring high costs when compared to other people who have more than 3 children. Text

Description automatically generated with low confidence

Chart, line chart

Description automatically generated

* We have plotted BMI against cost and found out that people whose BMI is higher have incurred high costs and are expensive.

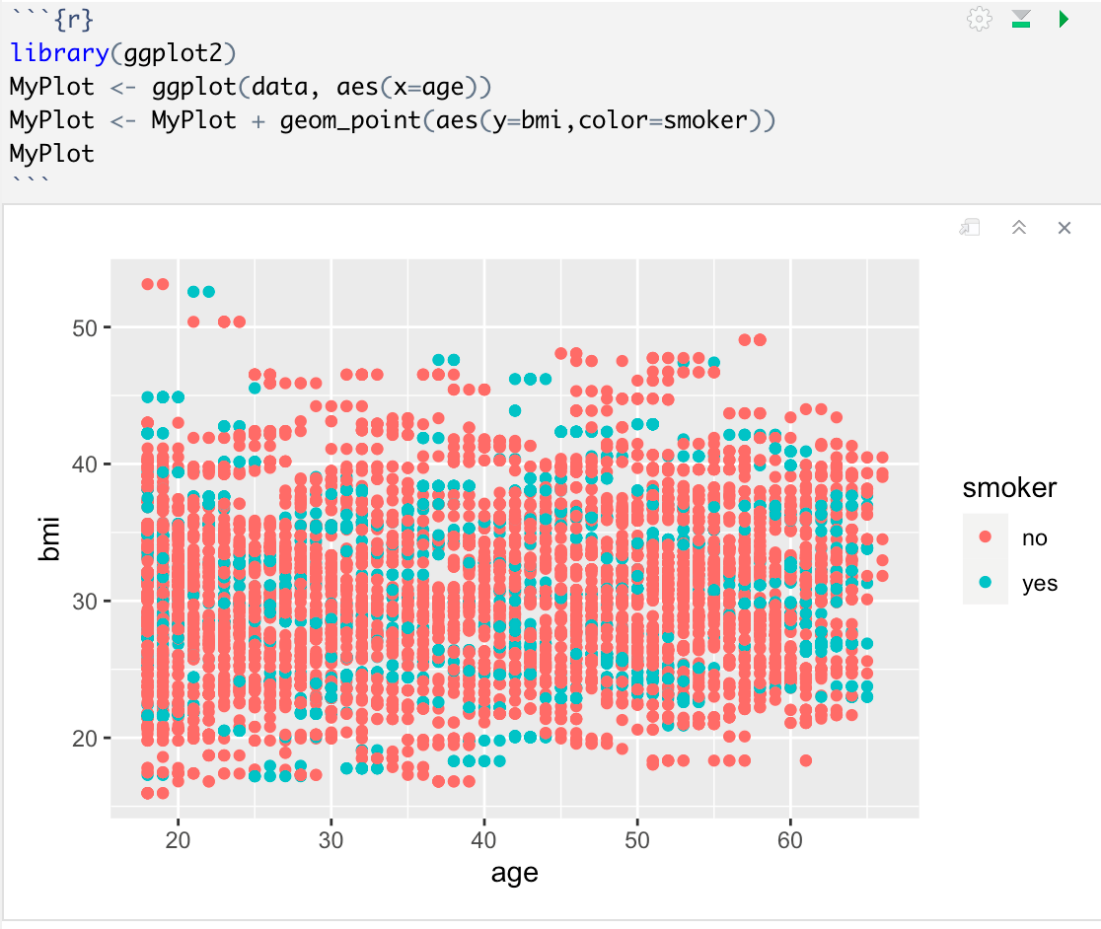
Text

Description automatically generated with low confidence

Chart, scatter chart

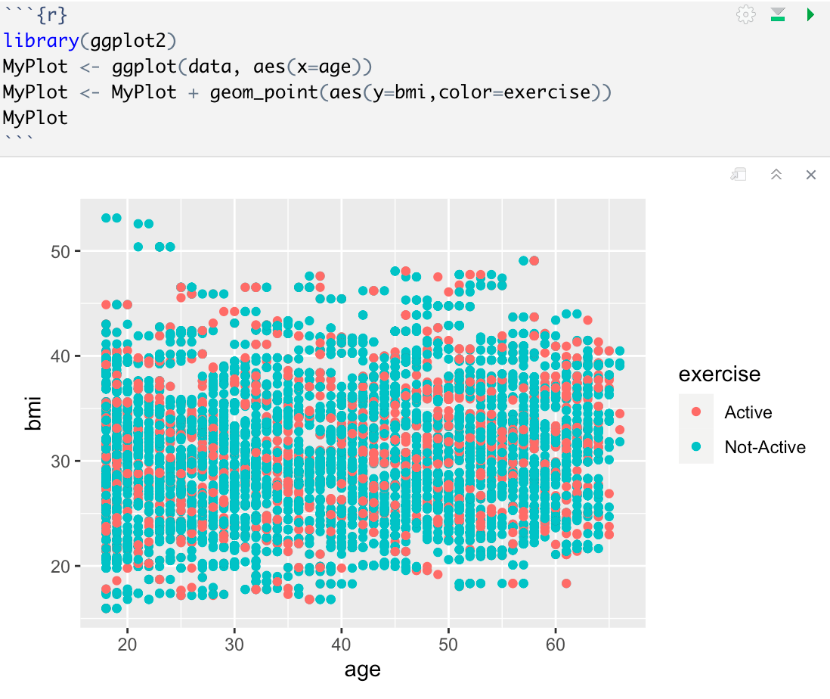
Description automatically generated

**Visualisations based on expensive variables:**

1. We tried to plot a scattered plot based on BMI and age, we colored the smoker variable. We came up with an assumption that as the age increases if the person is having more bmi proves that they are smoking 
2. We considered age and bmi if that is affecting the expenses of a person, and we found that people with age >47 with higher BMI has more expenses, than people with people with less BMI.



1. We plotted a graph to compare the age and BMI with active lifestyle. If the people with age>50 who are not having an active lifestyle. The code is given by :



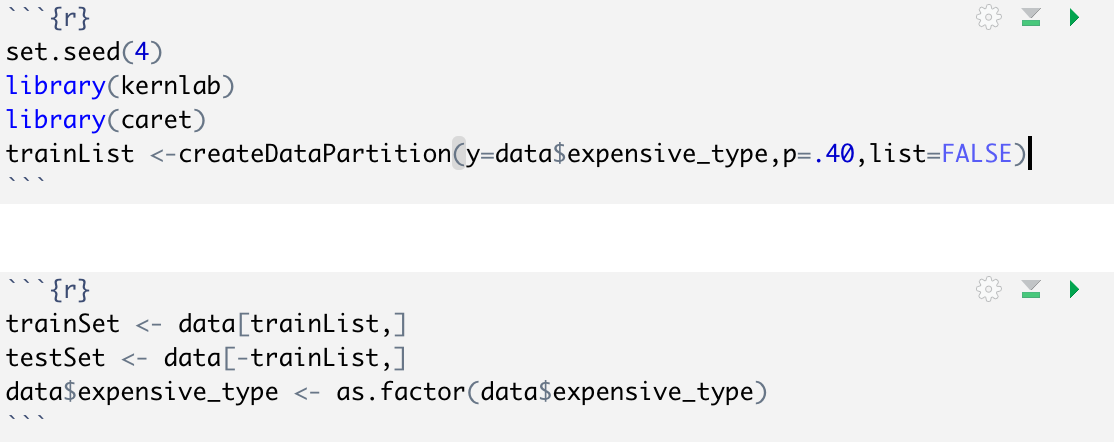
1. We plotted BMI against cost to check the linearity of people whether they are active or inactive. We were able to see that there is a positive correlation between the cost and BMI with exercise. The R code is given below:

Chart, scatter chart

Description automatically generated

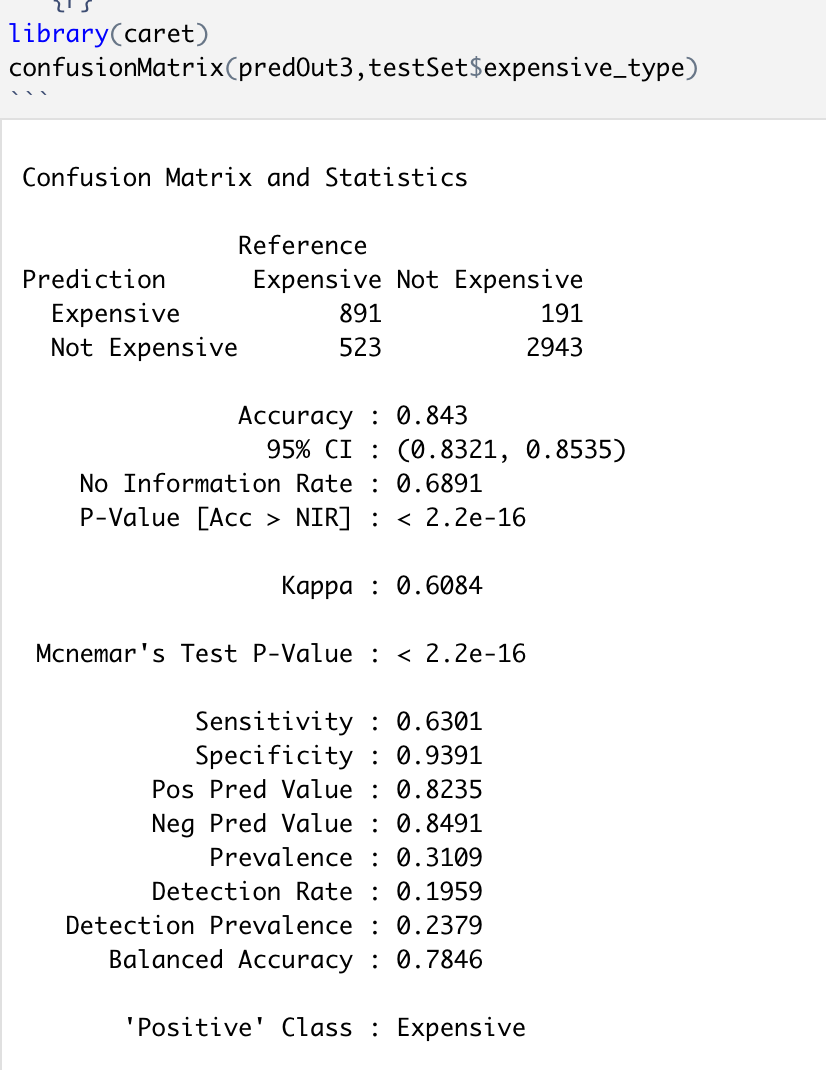
**SVM Modelling:**

Modelling is done by creating the training and testing data set. The training data set is formed by considering the expensive\_type variable. The data partition function is used to partition the data frame by expensive variable. The testing set is the partition created by the data partition function. The code for the training and the testing set is given below:



The svm model is created upon the age, BMI, smoker, exercise, hypertension to predict the expensive\_type from the train data set. Then predict function is used to predict the expoensive using the test data set. Confusion Matrix is created with the output of the predict function. The R code is given below:

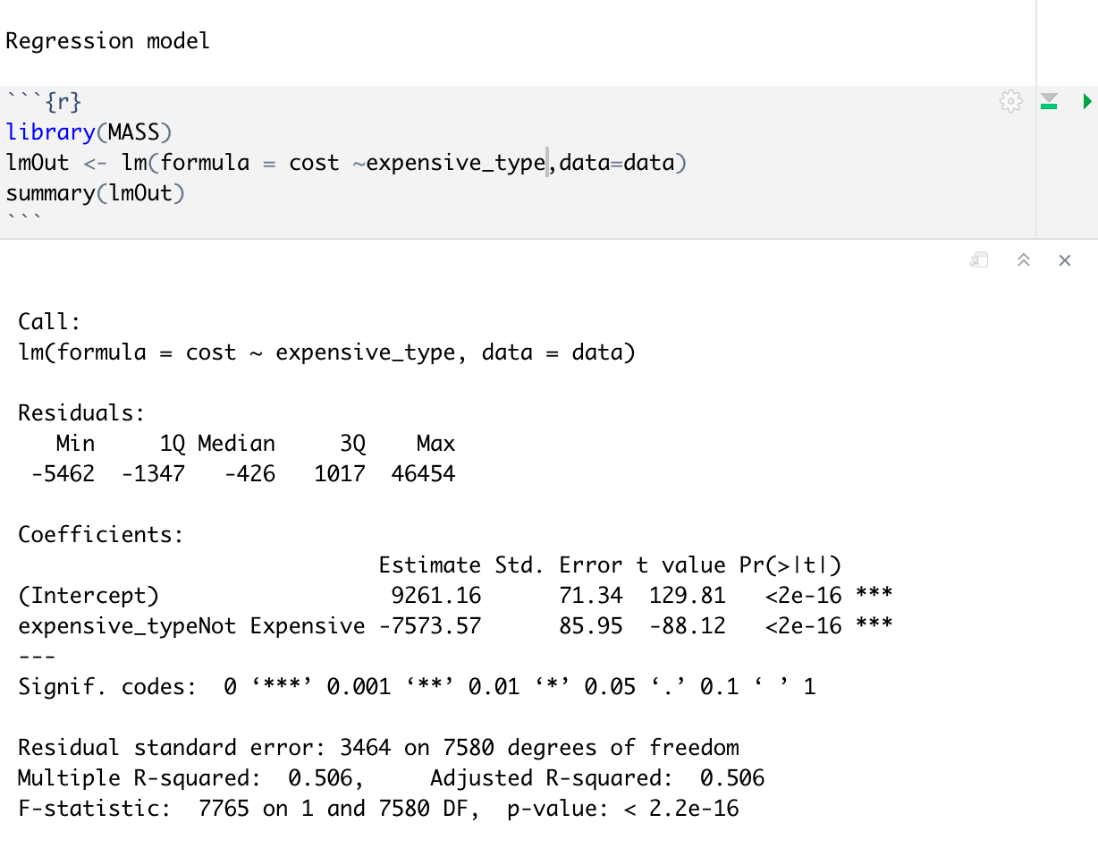




The output of the confusion matrix shows the accuracy of 84% with confidence interval of 0.83 to 0.853. The sensitivity that we obtained here is 63.01%.

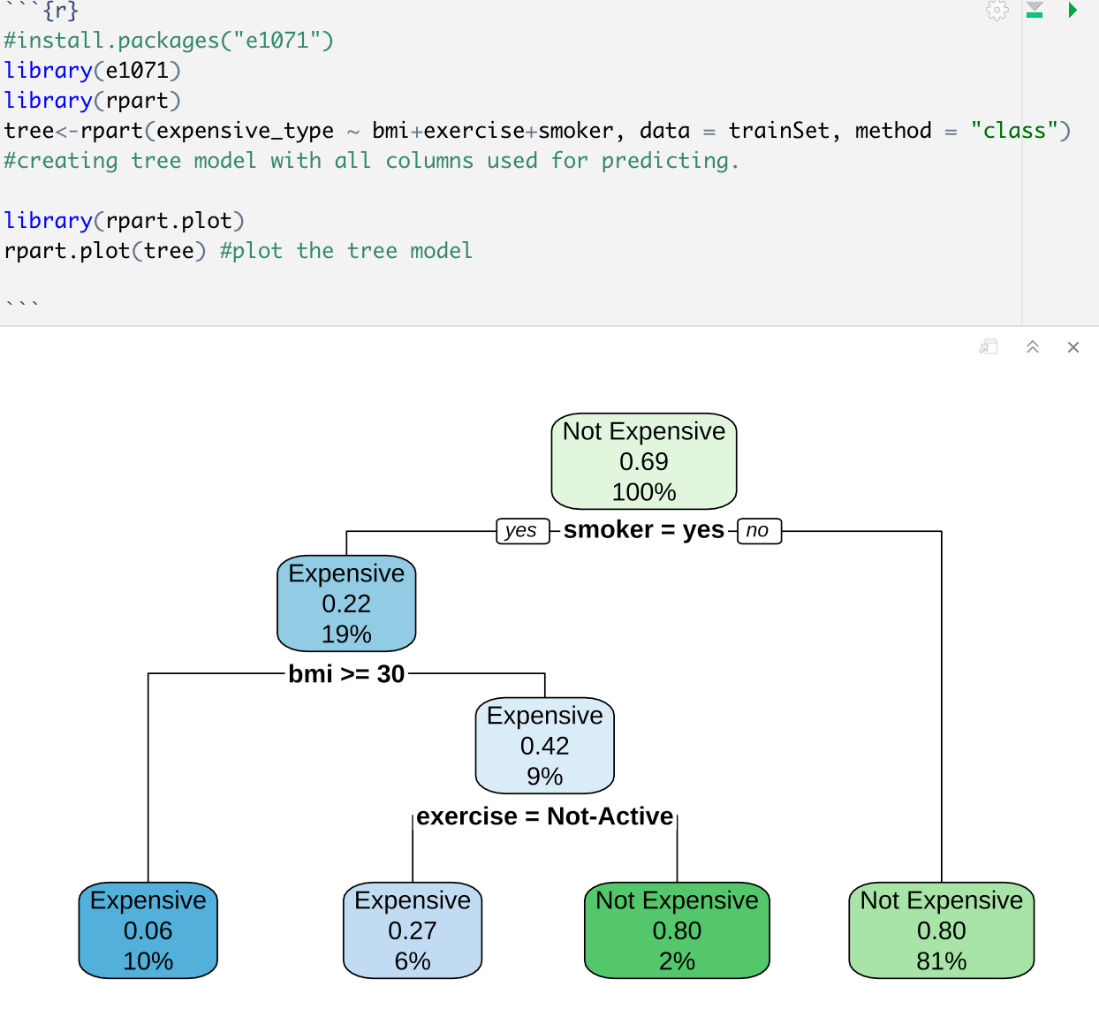
Regression Model:

We created a linear multiple regression model to predict the cost with the independent variable called Expensive variable. The summary of the linear regression shows that 50.6% of the expensive variable is accounting for cost. The adjusted r square value is 50.6%. The residual standard error is 3463 with 7580 degrees of freedom.



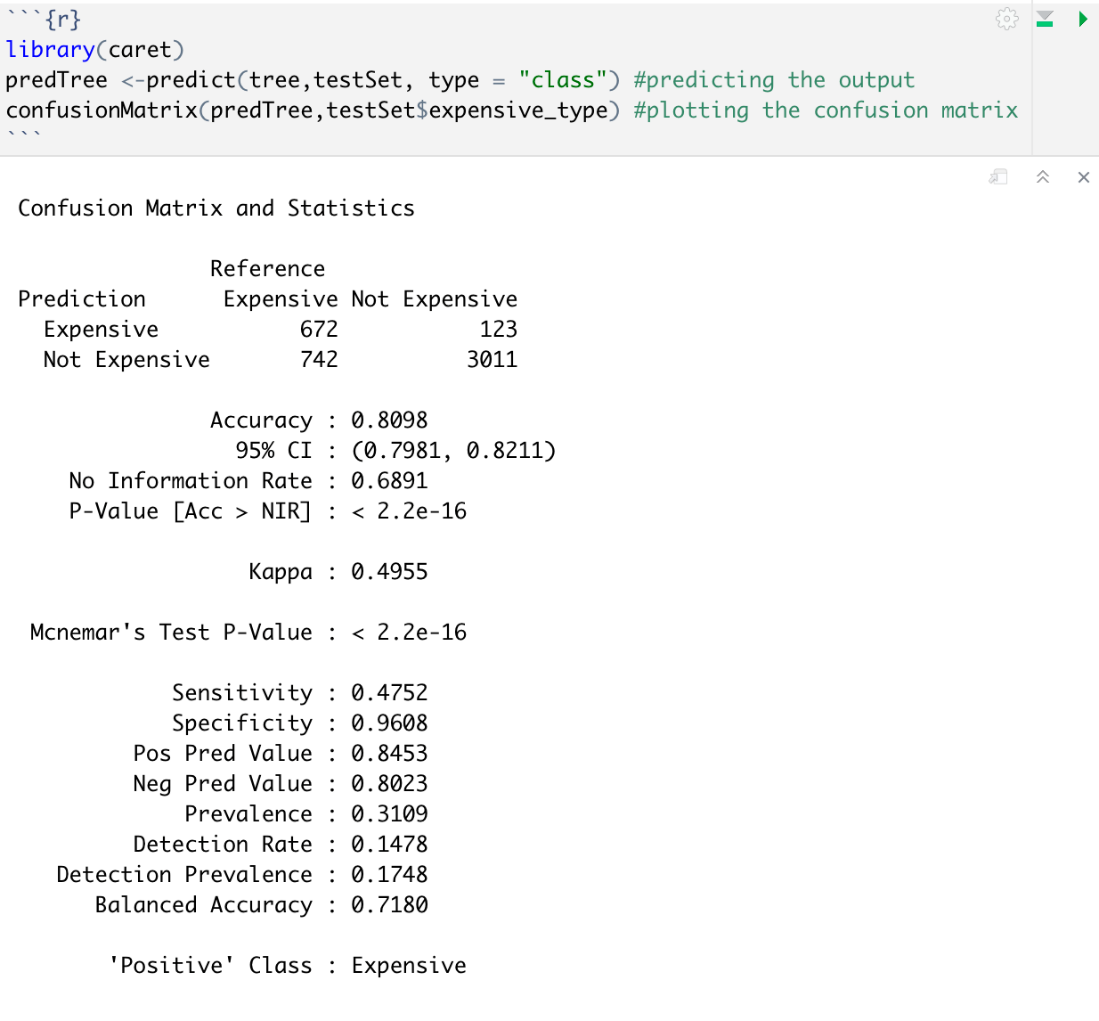
**Tree Model:**

We are trying to implement the tree model using e1071, rpart libraries. Over here we are trying to predict the rate of independent varible expensive type variable using BMI, exercise, smoker from the trainset. The R code for the tree model is given below:



The output of the tree model indicates that smoker variable is one of the main variables for the prediction, the next one would be BMI with 19%, the third variable is exercise with 9%.

The confusion matrix was created on the predicted tree variable, and we obtained the accuracy of 80.98% and sensitivity of 47%. This indicates that is model is not providing the results as expected.



**Shiny app -**

We created a shiny app that can be used by other analysts as well as laymen.

Our group's best version of Support Vector Machine of highest accuracy and sensitivities was written as a function. This function would return the predictions and confusion matrix.

The code was divided into front end and back end (ui and server).

In the UI, we selected a theme for the whole app.

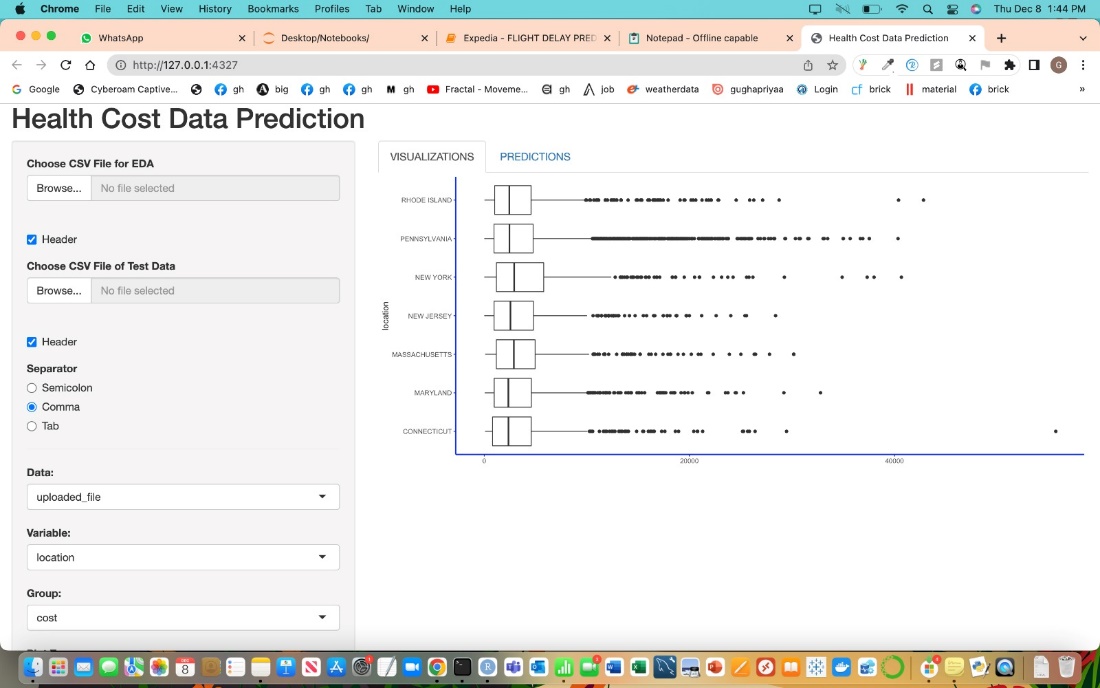
The interface was divided into a side panel bar and a main panel. The main panel was further divided into two tabs - one for exploratory data analysis and another for predictions.

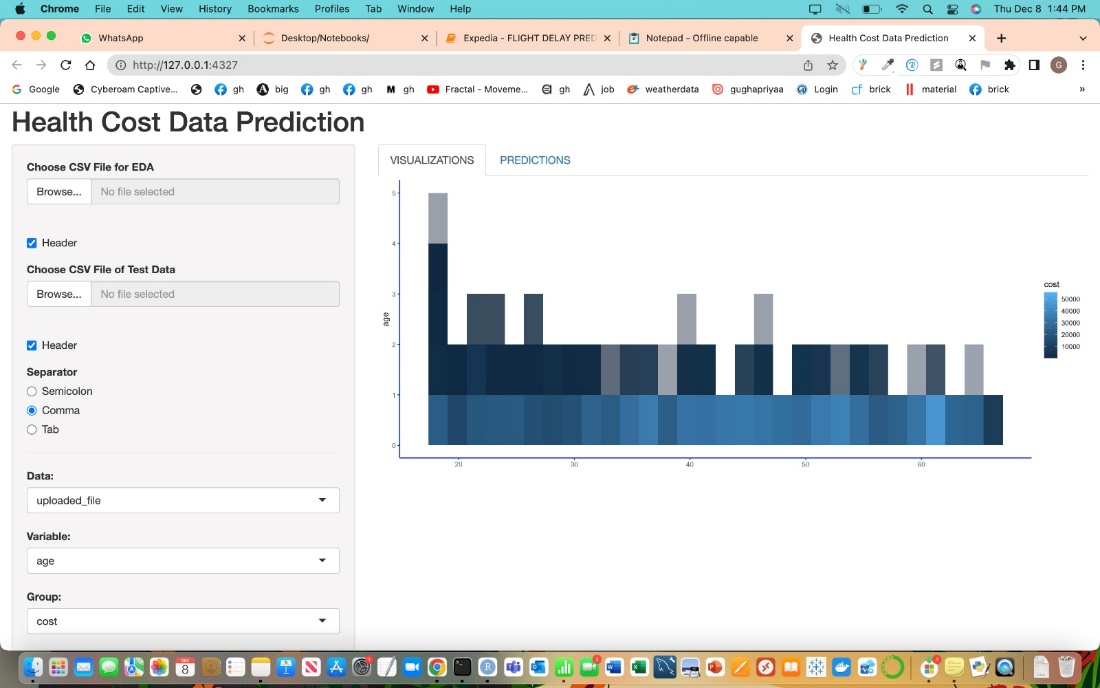
The side panel bar can take input of upload file of csv and read the file path and store in a variable file1. According to the file uploaded, checkboxes were created to select the delimiter of the data. As we usually work with csv, comma was set as default delimiter. Horizontal lines were used to divide portions of the interface and give clarity to the user.

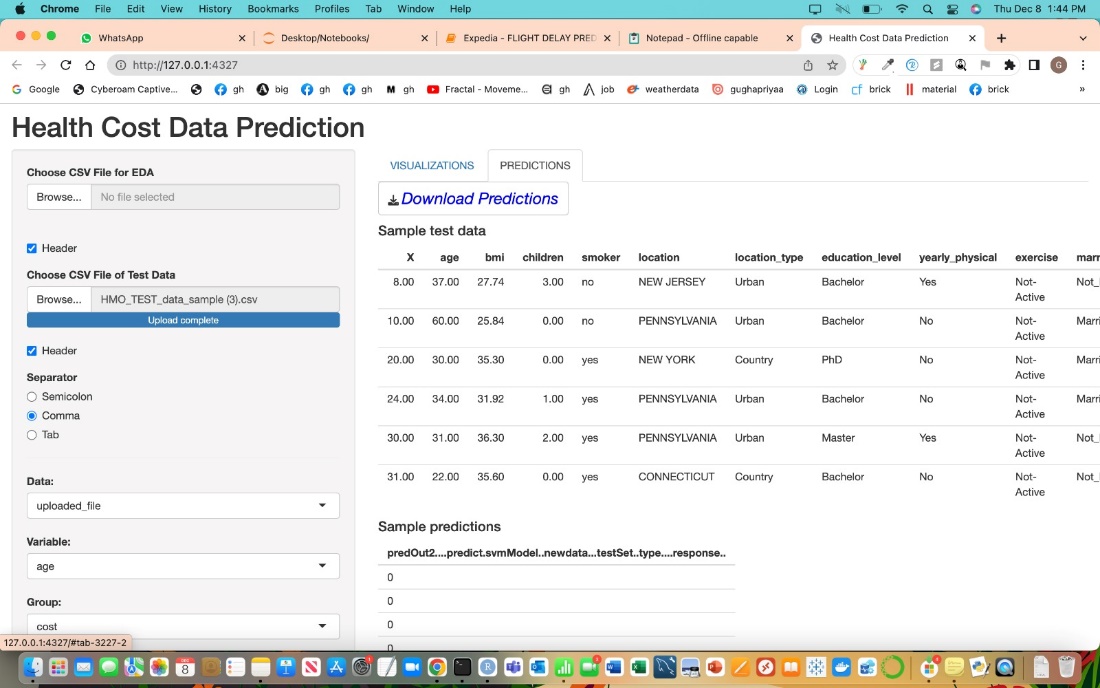
The side panel bar also had drop down menus to select the different attributes to compare the statistics and draw graphs. The main panel of graphs displays the graphs according to the variables selected. The tab for predictions provides a download button to download the predictions. It also displays sample test data as well as sample predictions.

In the server side, the backend code was written. An observe event was written to check if the csv file was uploaded or not. Once uploaded, a reactive function would detect the variables to the dropdown menu. render Text functions were used to write headlines and captions. Once the variables and groups were detected, we were able to plot graphs. Four different graph types were used - Boxplot, Histograms, Density and Bar graphs with an option of scatterplot in all the graphs. A checkbox to turn on or off the scatter points was provided. Using the render plots were displayed according to user selection. Separate functions for each graph were written.

Using Render Table, we displayed sample test data and passed the test data to the svm function. renderUI function was used to display sample predictions. In order to download the csv file, we used write.csv function to save it in .csv format.







**Recommendations –**  after analysing the data we came up with a few recommendations which can make it less expensive for the customers and help the HMO.

1. We offer health care programs for persons who are fat; with the help of these programs, they can not only become healthier but also reduce their medical bills.

2.As we can see, smokers are more expensive consumers for HMOs, thus launching a de-addiction program to aid them will be beneficial.

3. As a family with three children pays more in medical costs than children with more than three, we would want to offer a family plan that can significantly lower their medical expenditures.

4.Although there are not many people with hypertension at 0.50, there are quite a few who are in that range and may develop hypertension in the future, therefore we would like them to be checked on a frequent basis.