

Health Management organization healthcare predictions

RASHMI UMA SHIVAPRAKASH

Contents

[Introduction 2](#_TOC_250010)

[Problem Statement 2](#_TOC_250009)

[Data Variables 2](#_TOC_250008)

[Goal 3](#_TOC_250007)

[Process 4](#_TOC_250006)

[Health Management Organization Healthcare Prediction 4](#_TOC_250005)

[Reading and Understanding the Data 4](#_TOC_250004)

[Data Cleaning 5](#_TOC_250003)

[Data Interpretation 6](#_TOC_250002)

[Predictive model and accuracy 18](#_TOC_250001)

[Recommendations 22](#_TOC_250000)

# Introduction:

Health is a very important factor that affects human life. But one thing that humans fear more than their health is the healthcare cost they have to bear in case of emergencies. To help customers understand their healthcare expenses, Health Maintenance Organizations (HMOs), a healthcare organization, believe in helping customers understand how they can save on healthcare. To determine what is expensive and affordable in HMOs, data must be analyzed. The analysis should consider various factors such as smoking, exercise, BMI, and hypertension, among others, that could impact the cost variable. By examining historical data that shows the cost based on consumers, it is essential to estimate whether the cost will remain significant. This information can be valuable to the health industry in making more precise predictions.

# Problem Statement

To accurately forecast which customers would be pricey based on the dataset and to deliver relevant information based on the HMO data available.

# Data Variables

* **X**: Integer, Unique identified for each person
* **age**: Integer, The age of the person (at the end of the year).
* **location**: Categorical, the name of the state (in the United States) where the person lived (at the end of the year)
* **location\_type**: Categorical, a description of the environment where the person lived (urban or country).
* **exercise**: Categorical, “Not-Active” if the person did not exercise regularly during the year, “Active” if the person did exercise regularly during the year.
* **smoker**: Categorical, “yes” if the person smoked during the past year, “no” if the person didn’t smoke during the year.
* **bmi**: Integer, the body mass index of the person. The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy.
* **yearly\_physical**: Categorical, “yes” if the person had a well visit (yearly physical) with their doctor during the year. “no” if the person did not have a well visit with their doctor.
* **Hypertension**: “0” if the person did not have hypertension.
* **gender**: Categorical, the gender of the person
* **education\_level**: Categorical, the amount of college education ("No College Degree", "Bachelor", "Master", "PhD")
* **married**: Categorical, describing if the person is “Married” or “Not\_Married”
* **num\_children**: Integer, Number of children
* **cost**: Integer, the total cost of health care for that person, during the past year.

# Goal

* Predict people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs).
* Provide actionable insight to the HMO, in terms of how to lower their total health care costs, by providing a specific recommendation on how to lower health care costs.

# Process:

# Health Management Organization Healthcare Prediction

## Reading and Understanding the Data

The HMO data is in the form of CSV and to read and understand the data we need to read the CSV file into a data frame. In this case, we read the CSV using the function **read.csv ()** in R and store it in a data frame created under the name Project and Project\_1. We store the data in 2 different data frames as we will be using one of them for data interpretation and the other for creating a predictive model.

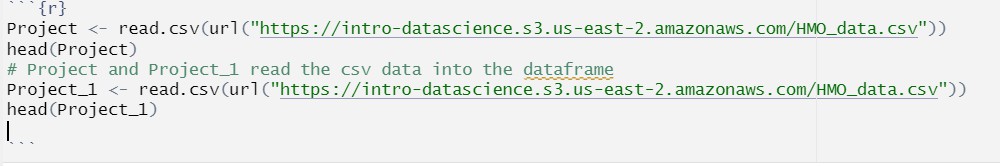


Fig 1: Read the CSV into data frame.

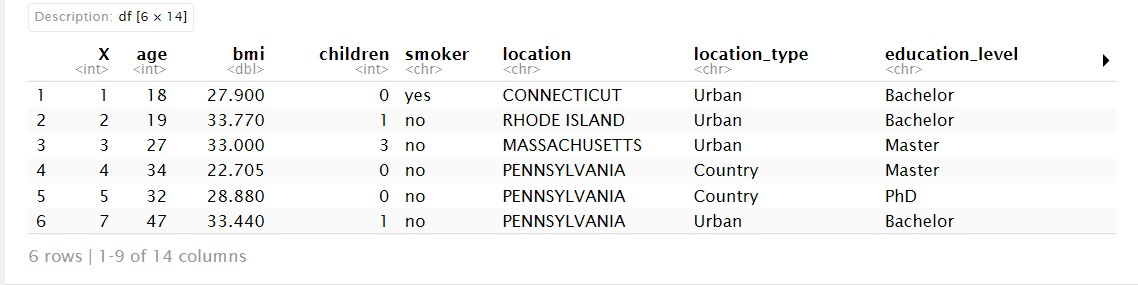


Fig 2: The first 6 values of the Project data frame

## Data Cleaning

To understand if the dataset has any null or NA values, we need to run code to check NAs in the data. In R, it can be done using the function **is. na ().** After identifying the NA values, we need to resolve the NA’s and in R we perform this action using the interpolation method. This is a function of the Impetus packet which fills in the NAs with values using linear interpolation. In the case of HMO data, we have NA values in BMI and hypertension which we fill in using **na\_interpolation().**

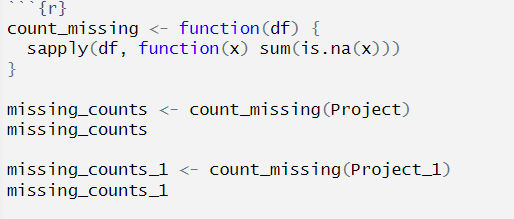


Fig 3: Missing values in a data frame

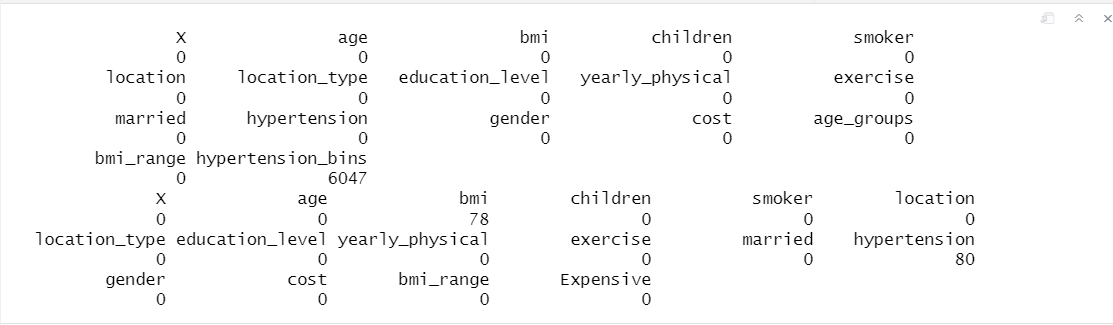


Fig 4: Output for missing values

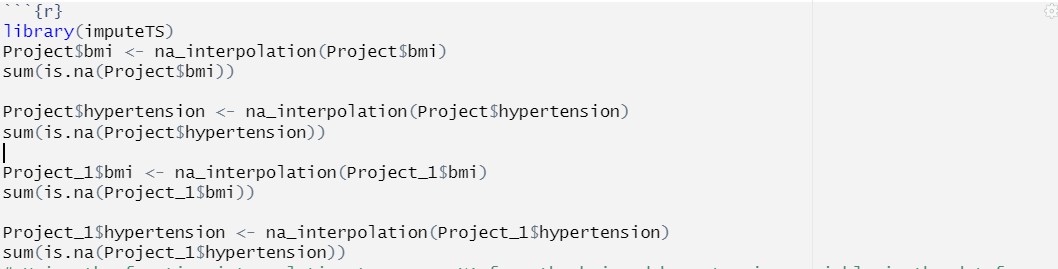


Fig 5: Remove the NA values using interpolation.

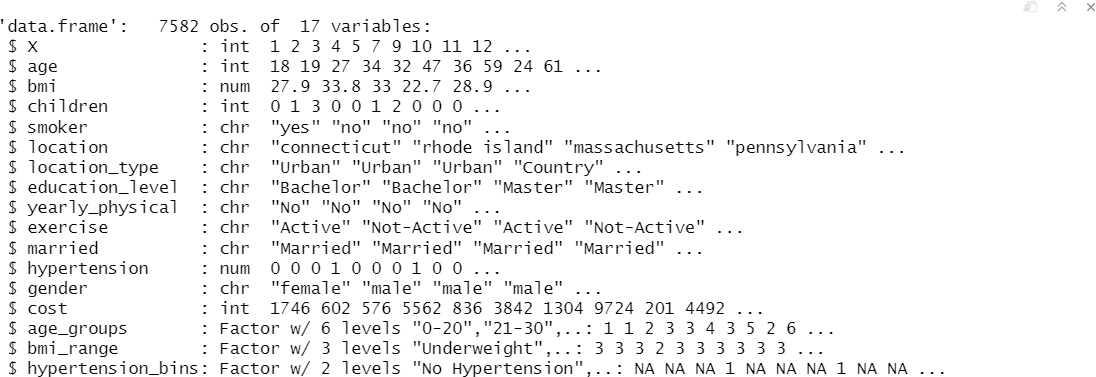
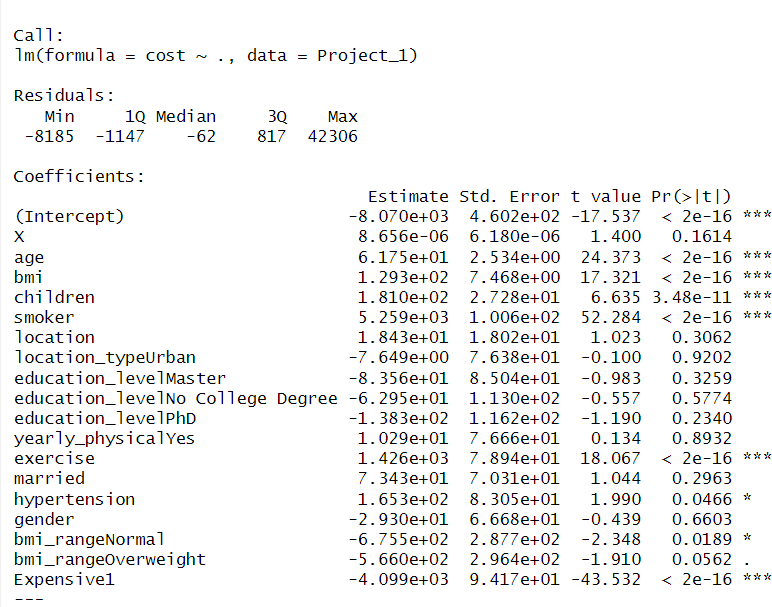


Fig 6: Summary of the data frame Project.

## Data Interpretation

To have a visual representation and interpretation of the variables from the HMO dataset, we have to find the correlation between the variables against the cost of the healthcare of the customers. The trends for visual representation can be understood by analyzing the underlying positive and negative correction between the attributes of the dataset. We saw a number of patterns, including age-related increases in health care costs, a favorable relationship between smoking and health, a yearly doctor visit, etc. Despite the fact that these trends were visible in the visuals we created, we could not entirely rely on them because visualizations cannot be the only useful tool for determining potential correlations between variables. As a result, we made the decision to use additional analytical methods to better gauge these possibilities. Information gain ratio analysis, which seeks to provide answers to the following questions by utilizing the

idea of entropy, which may be described as the degree of uncertainty, impurity, and disorder in a data collection, is one method for estimating the value of certain qualities as prediction tools.



To understand the relationship between location and cost, we plotted a box plot that gives a max, min, and median value of the cost based on the location. But from the below graph, it is difficult to interpret the relation between the two variables. We decided to plot a heat map which helped us know the distribution of cost.

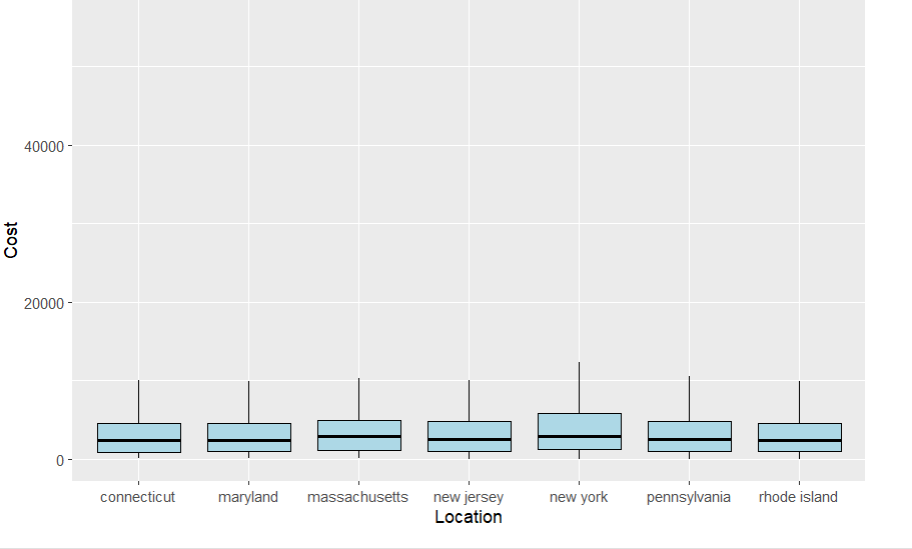


Fig: Box Plot representing the minimums, median, and maximum along with the quartiles of the cost by location. Here New York has slightly higher minimums than all other locations

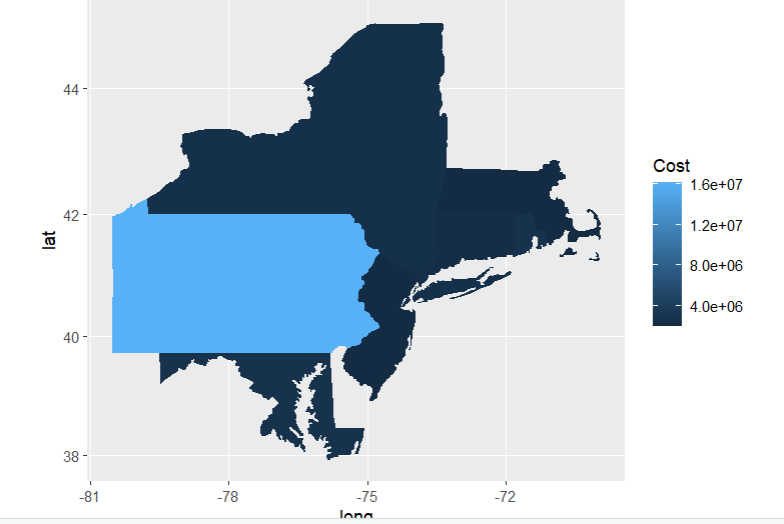


Fig: Location-based infographic measuring cost, and representing the state of Pennsylvania having the highest healthcare cost

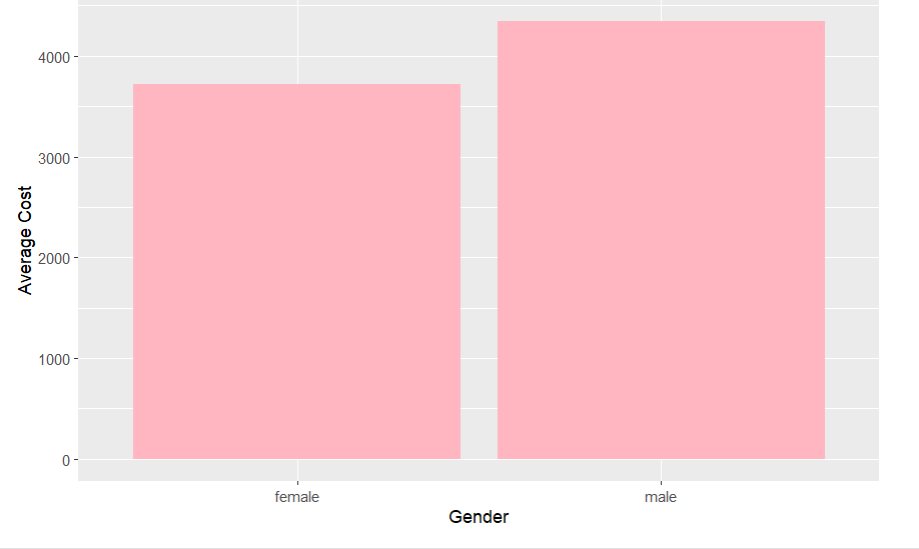


Fig: Average healthcare cost by gender (female, male) of the individuals. It is observed that females have slightly lesser healthcare costs than male by around 500$ at average.

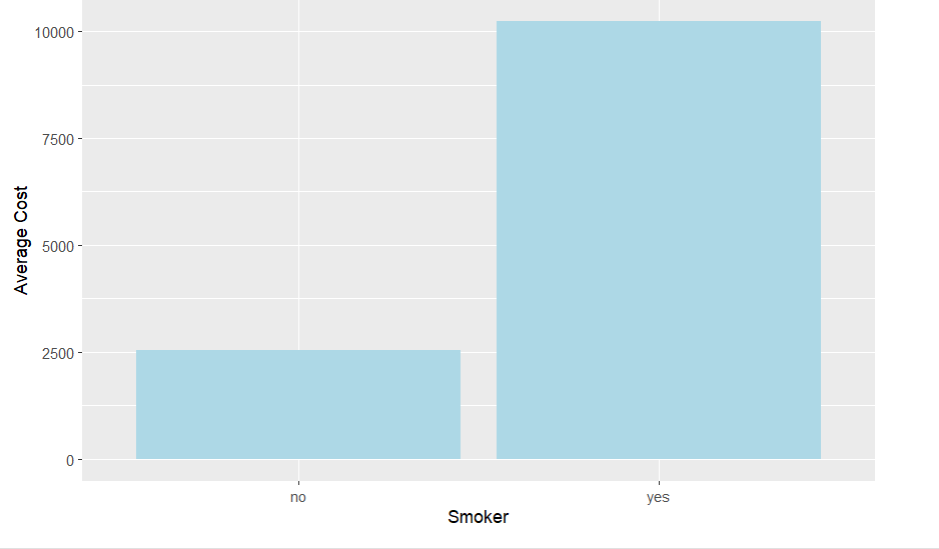


Fig: Bar plot showing the average healthcare cost if the individual smokes or not. It can be clearly observed that smokers have greatly higher healthcare costs than non-smoking individuals by around 7500$

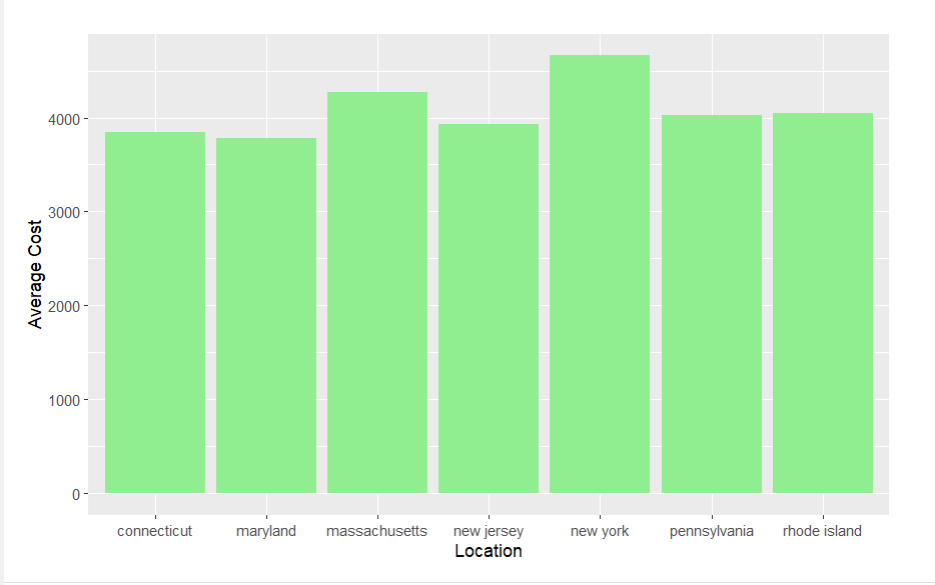


Fig: Bar plot showing the average healthcare cost by location, showing New York having slightly higher healthcare costs than other locations

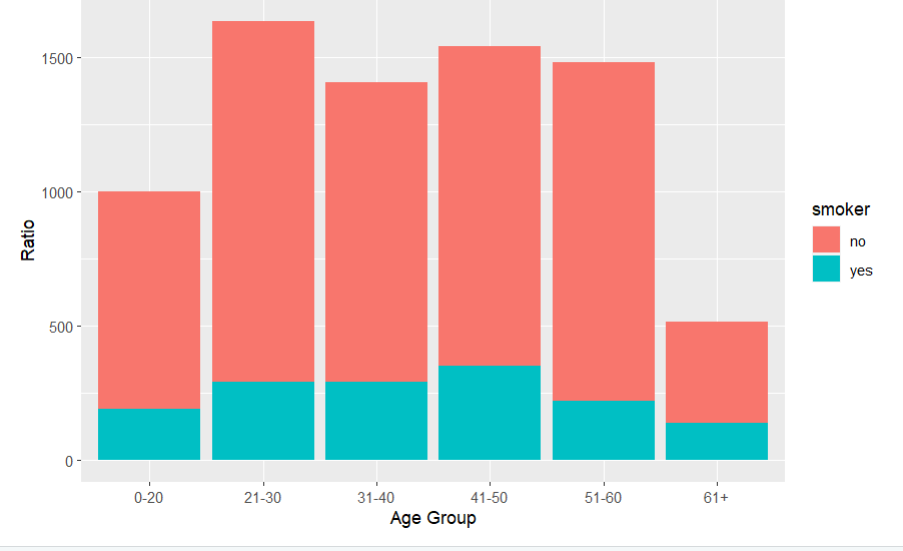


Fig: Plot depicting the ratio of individuals by smoking (yes/no) for each age group (0-20, 21-30

…). It can be inferred that the ratio is higher for the middle-aged for the number of people smoking in comparison to number of people not smoking and lesser for 0-20 and 60+ age groups.

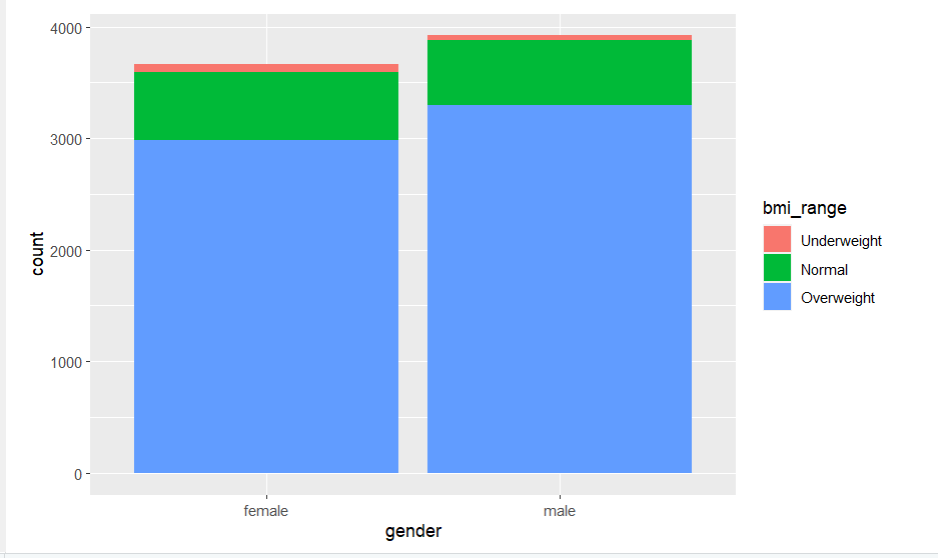


Fig: Plot showing count of individuals in each BMI range (underweight, normal, and overweight) by their gender. It can be inferred from the plot that most people fall in the overweight range for males and females, while underweight individuals are slightly more for females than males.

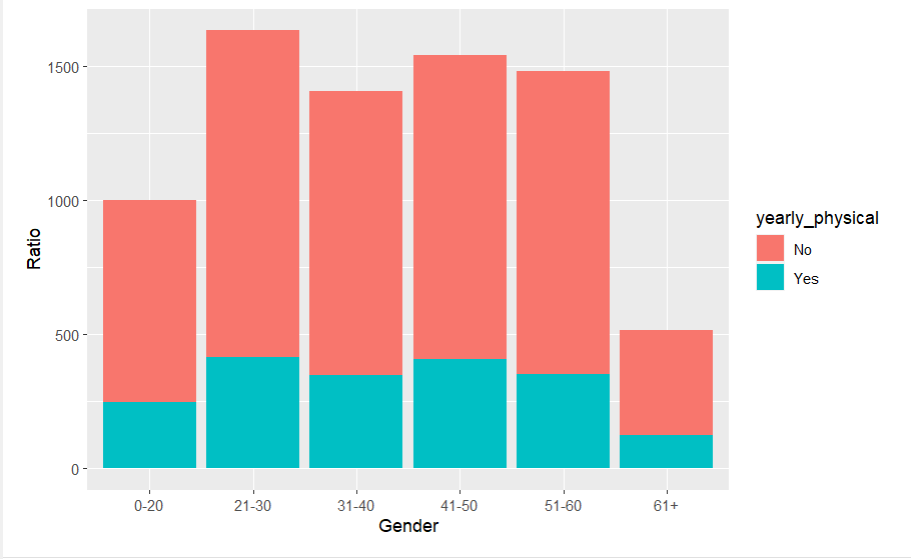


Fig: Plot depicting ratio of people performing physical exercise yearly for every age group (0-20, 21-30 …). It can be inferred that ratio becomes higher for middle-aged while in the age group 0- 20 and 61+ the number of individuals performing physical activity than number of individuals not performing is lesser.

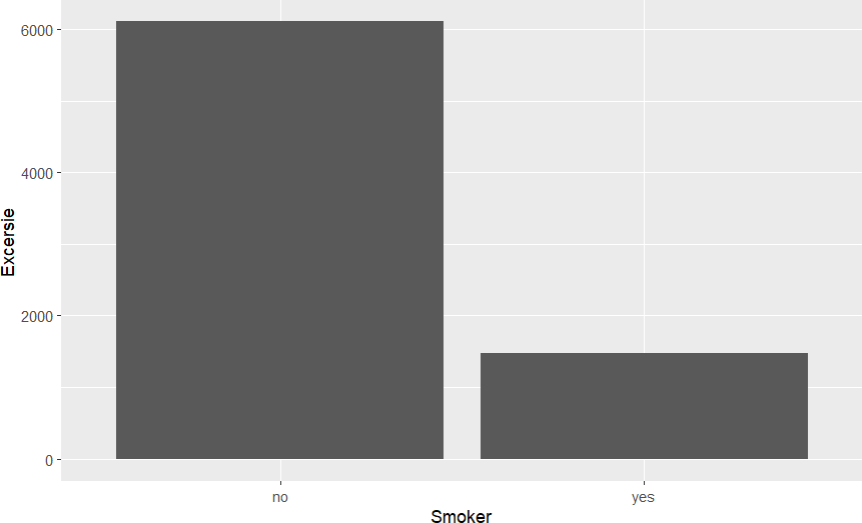


Fig: Bar plot depicting the number of individuals exercising by if they smoke or not. It can be clearly seen that number of people smoke don’t exercise as well.

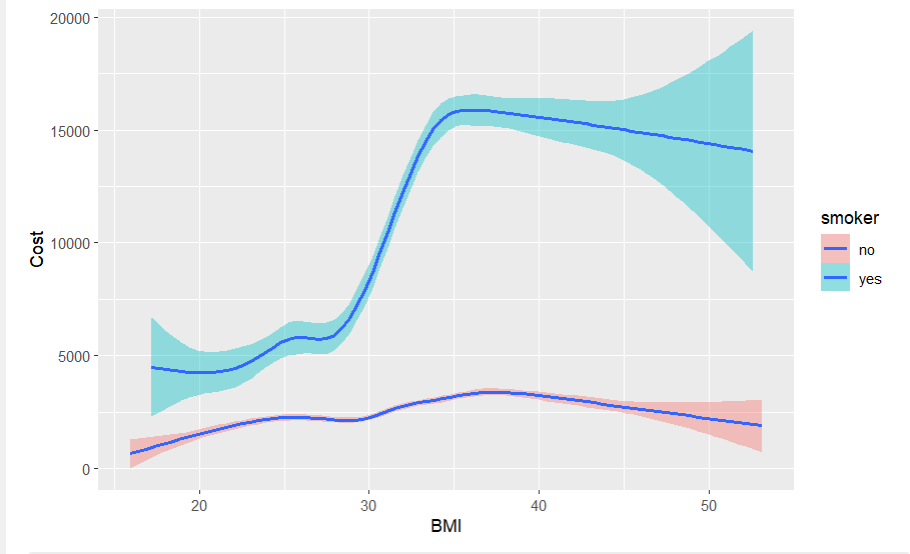
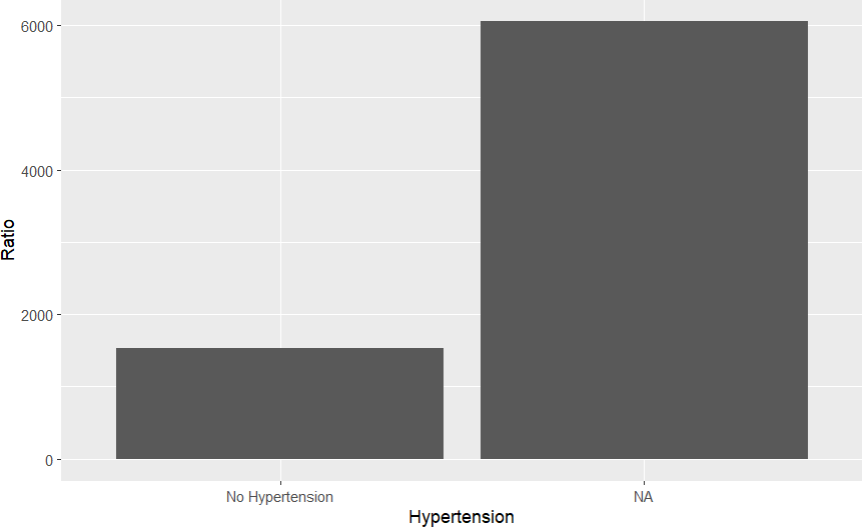
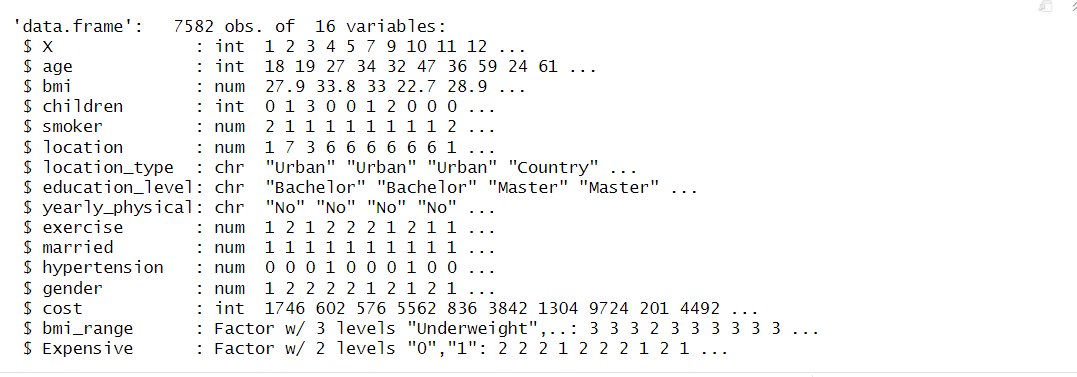


Fig: : Line plot of BMI vs healthcare cost based on if the individual smokes or not. It can be clearly seen from the plot that with a lesser BMI the difference between healthcare cost of smoker vs not smoker is a lot less than individuals with a higher BMI. It clearly indicates if the individual is smoking and has a higher BMI than the healthcare costs would towards around 15000$.



## Predictive model and accuracy

To have the correct predictive model and high amount of accuracy in the HMO data, the variables which impacted the most should be converted into categorical values using the function **as.factor ()**



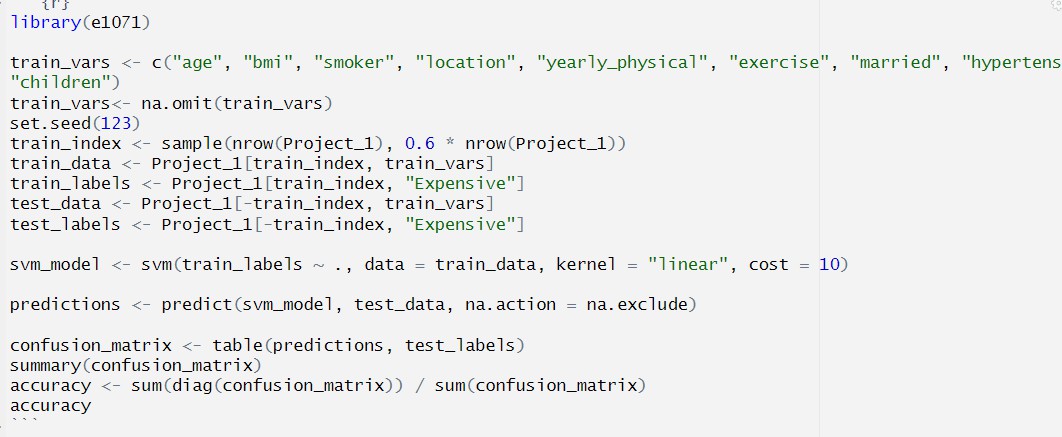
Post conversion of the variable we declare a variable ‘Expensive’ which has the comparison of the customers who have healthcare cost higher than the average or not. It gives an elaborate knowledge of the dataset using machine learning models. In this case we have implemented Support Vector Machine (SVM) and Decision Tree

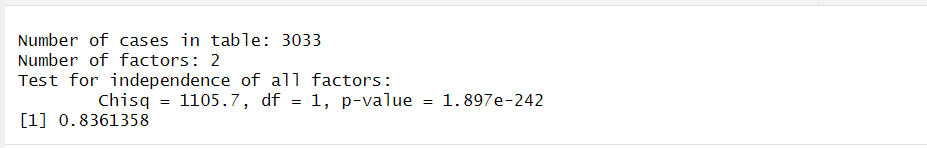
**Support Vector Machine**

A well-liked supervised machine learning algorithm for classification and regression analysis is the SVM (Support Vector Machine). SVM seeks to establish a boundary (hyperplane) that best divides the data into various classes. The technique seeks to maximize the distance between the support vectors and the border, which are the data points closest to the boundary.

In this case, we have implemented two kernel SVM, Radial and Linear

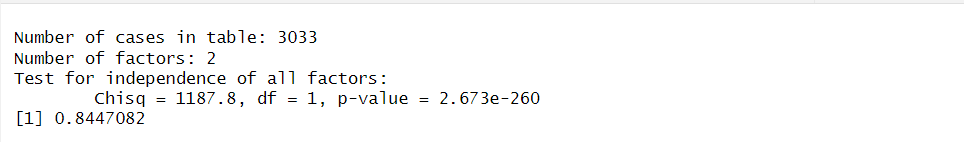
1. **Linear Model**





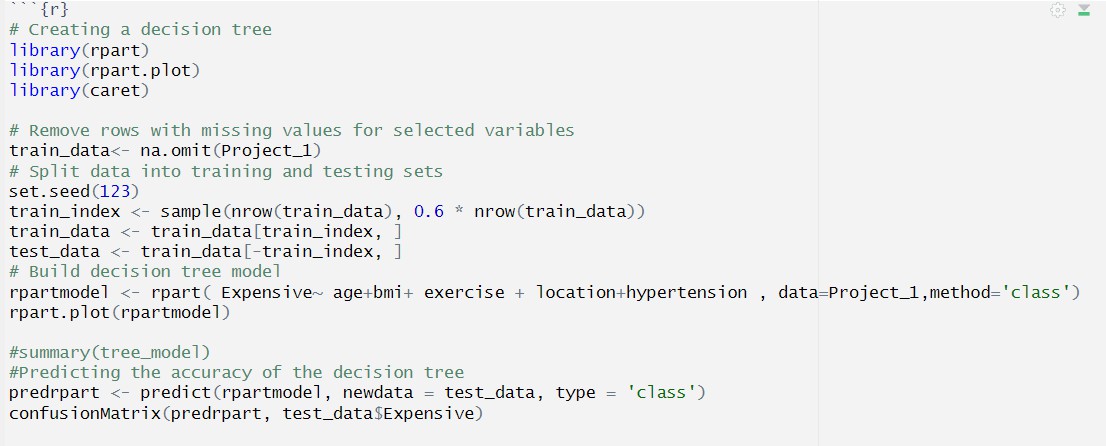
1. **Radial Model**

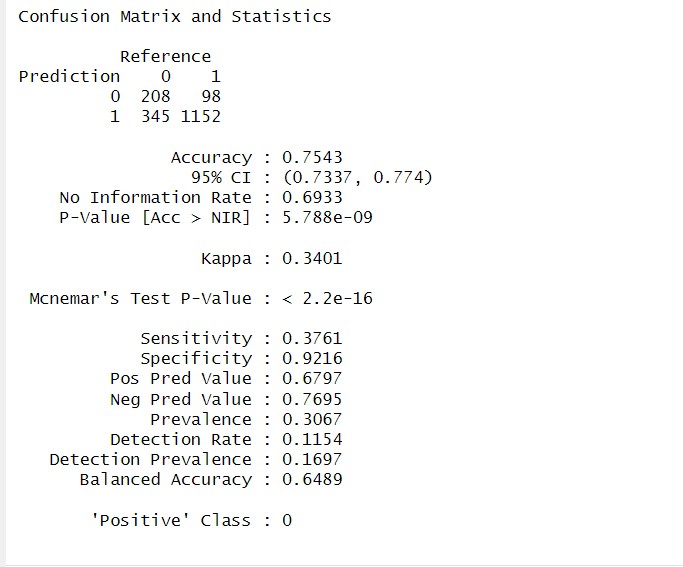
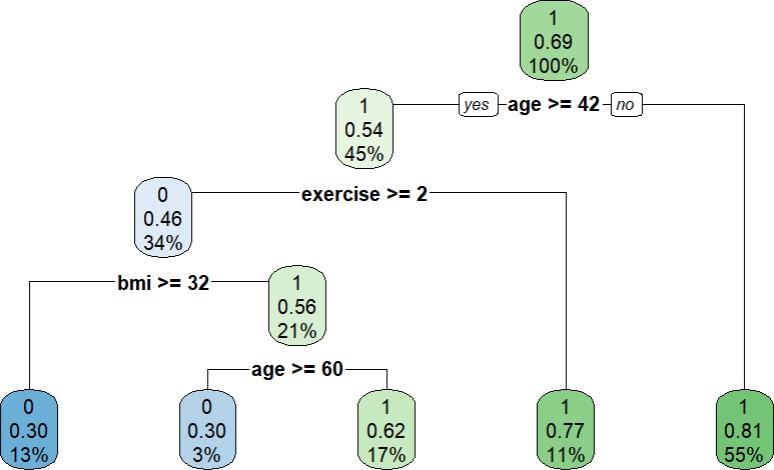




**Decision Tree**

A decision tree is a graph that displays options and their outcomes as a tree. The edges of the graph indicate the conditions or rules for making decisions, whereas the nodes in the graph represent an event or a choice. It primarily appears in R-based machine learning and data mining applications**.**





## Recommendations

Some of the recommendations we would like to give to HMO from the dataset of the customers is as follows:

* As a healthcare organization, HMO should raise awareness campaigns for exercise and anti-smoking in customers who belong to the age group of 41-50 and early. These campaigns can help people understand the impact of smoking and not smoking on their health which will in return reduce the healthcare cost.
* These campaigns should start with the highest number of customers impacted, that is Pennsylvania working down towards the states as per the impact.
* Fitness push on a Monthly Basis: The insurance business should organize marathons, game days, matches, etc. as part of its push to get people physically active.