Healthcare Cost Information Analysis

Group 4

- Akshata Ravinder
- Han Mo
- Ibrahim Kizil
- Indraneel Somayajula
- Jaya Varshini

TABLE OF CONTENTS

SECTION 1: EXECUTIVE SUMMARY

A health maintenance organization (HMO) is a network or organization that provides health insurance coverage for a monthly or annual fee. An HMO is made up of a group of medical insurance providers that limit coverage to medical care provided through doctors and other providers who are under contract with the HMO.

This project aims to analyze the HMO dataset to identify types of candidates who are more expensive, and why. The main goals are to use data analysis techniques to predict people who are likely to spend more money on health in the next year and provide actionable insights on how their healthcare costs can be reduced. A review and analysis of healthcare cost data can be a helpful tool for understanding the most common and expensive health conditions where claims have been made; examining trends in costs over time; and comparing utilization rates to local, state, or national norms. Analysis of trends in healthcare expenditures will assist with assessing the effects of health promotion programs.

We began the process of determining variables that have the most impact on patient healthcare costs by understanding the dataset variables and building a set of business questions regarding healthcare costs that the given data might be able to answer.

Firstly, we focused on preprocessing the data and identifying missing and null values in the dataset that might cause anomalies in the data analysis output and used methods like data interpolation and imputation to replace these values. After this, we performed exploratory data analysis to build visuals of the data to identify any patterns, spot irregularities within the data, and find possible relationships between the variables to build new assumptions. We then built several predictive data models like Linear regression, Logit, Decision trees, Random Forests, Neural Networks and SVM.

Based on our findings from the initial Exploratory analysis and ML models we developed recommendations for the HMO that would help in reducing healthcare costs. The final recommendations are listed below in Section 7.

SECTION 2: INTRODUCTION

Health Management Organizations (HMO) are providers of comprehensive medical care for voluntary subscribers based on a certain prepaid fee. HMOs usually have their own network of doctors and healthcare providers within their network who agree to offer services at certain amounts, which helps HMOs keep costs in check for their members. Thus, collection, management, and analysis of Healthcare data is extremely crucial for HMOs as it helps in enabling better patient care, strategic pricing, and improved productivity in organizations.

In this project, our dataset consists of healthcare cost information of several patients from a Health Management Organization, and our aim is to analyze the dataset and build a predictive model that helps the HMOs determine which people will be expensive in terms of Healthcare costs.

Such predictive models are useful for HMOs to improve pricing forecasts, predict patients who might be expensive candidates, and find possible methods of reducing costs. We have focused on creating various Machine Learning models and compared and improved their accuracy to determine the best methods to solve our problem.

Core business questions:

- What are the features of our current clients and the distribution of cost?
- how to define the expensiveness of a case.
- What is the determinate variable/ what variables impact on expensive in defining the expensive case?
- Characteristics of expensive clients and recognize 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 and make recommendations to the HMO, in terms of how to lower their total healthcare costs.

SECTION 3: DATASET OVERVIEW AND ASSUMPTIONS

3.1 Dataset Description

The dataset we use consists of various patients' healthcare information in an HMO. It consists of 7582 patient records and 14 columns. The variables in these columns and their detailed explanation can be found in the Appendix. The variable we are trying to predict is 'cost'. This variable contains the total amount the person spent on healthcare during the past year. We used the remaining variables in the dataset to predict this dependent variable.

3.2 Variables used:

X: index of observation

age: Age of the person

location: State of residence

location_type: type of residence (urban or country)

exercise: exercise status

Smoker: if the person smoked during the last year.

bmi: body mass index of the person

yearly_physical: if the person had a yearly check up with their doctor

Hypertension: Describes if the person had Hypertension

Gender: Type: Categorical

Education_level: level of college education the client received

Married: marital status

Num_children: number of children the individual has

Cost: cost of healthcare for the client in the past year

For our initial exploratory Data Analysis, we used all the variables. However, based on each of our ML model requirements, we are going to convert some variable types into numeric or dummy formats to use as inputs according to model requirements.

3.3 Data Assumptions

This section contains information about the assumptions and limitations of the data that will remain constant through our further Exploratory analysis and ML model development.
- One of our main assumptions is that the data is collected from an unbiased sample, but there are still possibilities of natural variations within the data which we will consider to be true outliers.
- Our second assumption is that even though the data sample is collected from patients in the northeastern states of the US, the sample is inclusive of the entire population demographic.

3.4 Data Preprocessing

3.4.1 Data type and missing values

From the original dataset, for variables of smoker, location, location_type, education_level, yearly_physical, exercise, married, gender, they are categorical but the data type is character. For our convenience of further analysis, we need to convert them into factors.

After importing the data file, we have a dataframe with 7582 observations and 14 variables. By applying is.na function to all of the variables, we found 78 missing values in bmi and 80 missing values in hypertension. ImputeTS package is used to interpolate the missing value in hypertension. For hypertension, since it is a binary variables, we assigned 0 for all missing value assuming those clients did not have hypertension

```
‘’
library(imputeTS)
data$bmi<- na_interpolation(data$bmi)
data[is.na(data$hypertension),which(colnames(data) == "hypertension")] <- "0"
‘’
```

3.4.2 Criteria for expensive / non expensive

A new variable named expensive is created in the dataset since our goal is to identify the pattern of expensive clients. The new variable is defined by the cost variable in the original data. We choose the top 20% as the cap, which means the clients with the top 20% of cost will be identified as expensive cases and the rest 80% will be identified as non-expensive cases. Expensiveness will be our target variable for all the following analysis and predictions. From our industry research, 5,000 is widely used to determine the expensive case. In our HMO data, 80% quantile in cost is 5778.9, which is very close to $5,000 industry standard. Therefore, we decide to proceed with the top 20% as an expensive case.

SECTION 4: EXPLORATORY DATA ANALYSIS

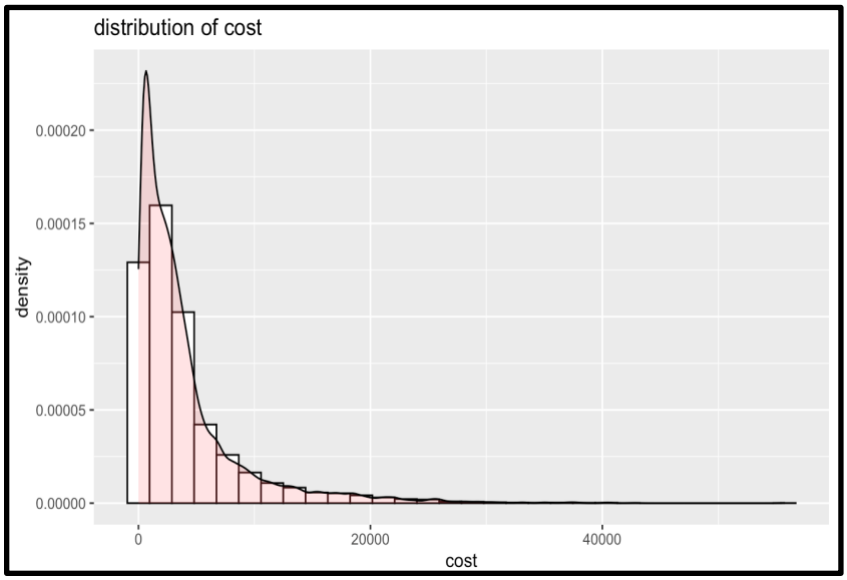
4.1 Descriptive Summary

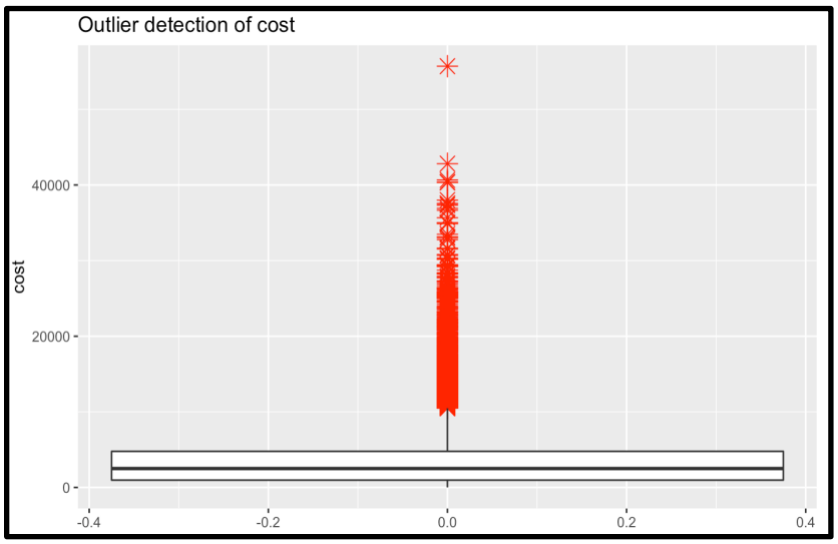
Firstly, we generated a basic descriptive summary for all the variables.

```
       X                  age              bmi            children     smoker
 Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000   no :6103
 1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000   yes:1479
 Median :    24916   Median :39.00   Median :30.50   Median :1.000
 Mean   :   712602   Mean   :38.89   Mean   :30.80   Mean   :1.109
 3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.70   3rd Qu.:2.000
 Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000

          location    location_type           education_level yearly_physical       exercise
 CONNECTICUT  : 611   Country:1903   Bachelor         :4578   No :5699        Active    :1888
 MARYLAND     : 747   Urban  :5679   Master           :1533   Yes:1883        Not-Active:5694
 MASSACHUSETTS: 465                  No College Degree: 759
 NEW JERSEY   : 498                  PhD              : 712
 NEW YORK     : 547
 PENNSYLVANIA :4010
 RHODE ISLAND : 704
        married     hypertension    gender          cost       expensive
 Married    :5060   0:6078       female:3662   Min.   :    2   0:6065
 Not_Married:2522   1:1504       male  :3920   1st Qu.:  970   1:1517
                                               Median : 2500
                                               Mean   : 4043
                                               3rd Qu.: 4775
                                               Max.   :55715
```

Speaking from numbers, the interquartile (from 25% - 75%) of the clients age falls between 26 to 51, whereas the interquartile for bmi is 26.6 to 34 and interquartile for children is 0 to 2. The middle 50 percent of cost goes from 970 dollars to 4775 dollars. Other basic observations are that about 80% of clients are not smokers, 75% of clients live in urban areas, 75% of clients do not have regular yearly physical check, 75% of clients are not active in exercise, 67% of clients are married, and 80% of clients do not have hypertension. This dataset is relatively equally distributed for gender with male taking 52% from the overall observation. About 53% of clients come from Pennsylvania, whereas all the other states, including Connecticut, Maryland, Massachusetts, New Jersey, New York and Rhode Island account for less than 10 % each. Bachelor degree takes 60% of clients' education level whereas master makes 20% and PhD and no college degree contribute for the rest 10%.

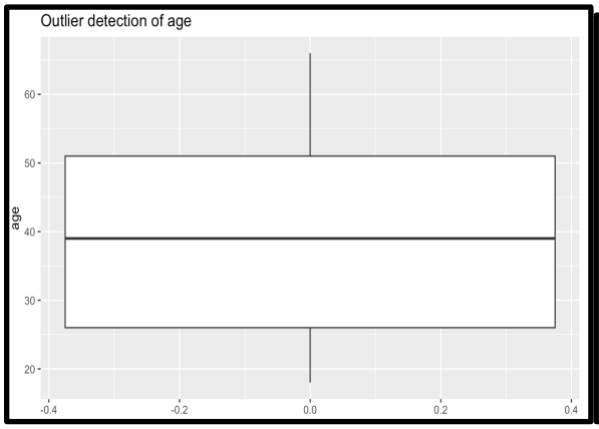
4.4 Box plot for outlier detection

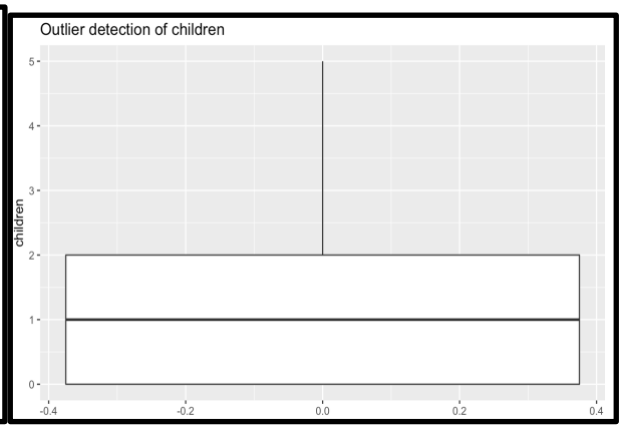




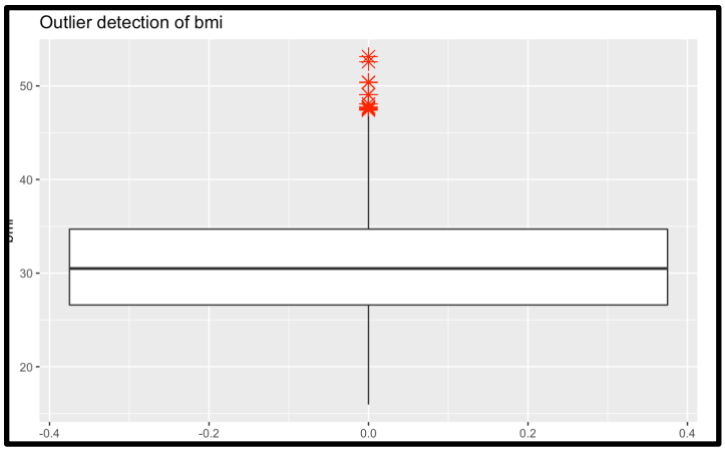
From the histogram of cost, we see a very right skewed distribution. Telling from the plot, most cases have the cost smaller than 20,000 dollars. Engaging the information from the descriptive summary, the middle 50 percent of cost goes from 970 dollars to 4775 dollars, and maximum cost is 55,715 dollars, which indicates that an outlier detection is very necessary. As shown in the box plot, the dataset has a lot of extreme value, which makes more sense for us to convert cost into expensiveness dummy variables for predictive models and step outside from the impact of outliers.

In the next step, we are going to do outlier detection for other numerical variables.



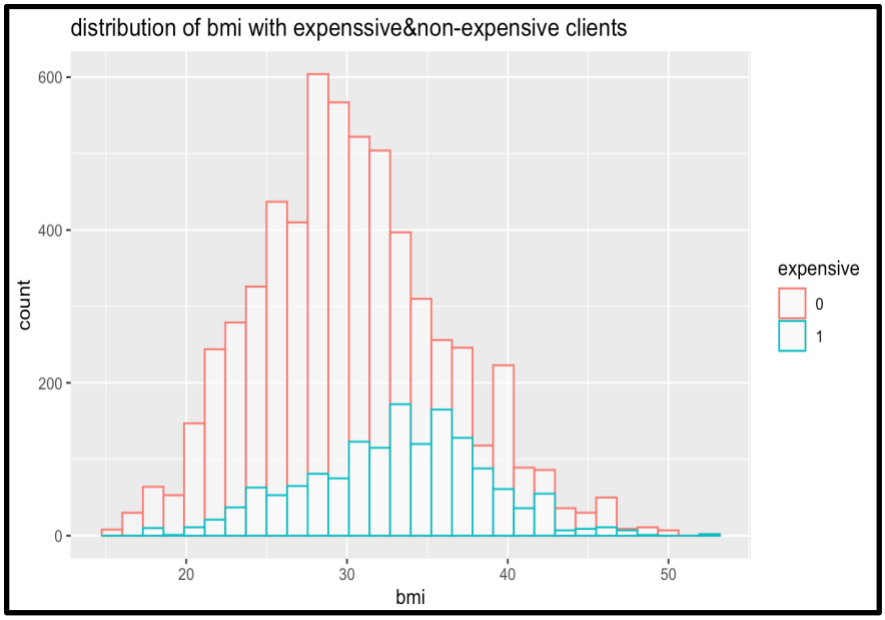


The box plots for age and children show that no extreme value is found in these two variables.

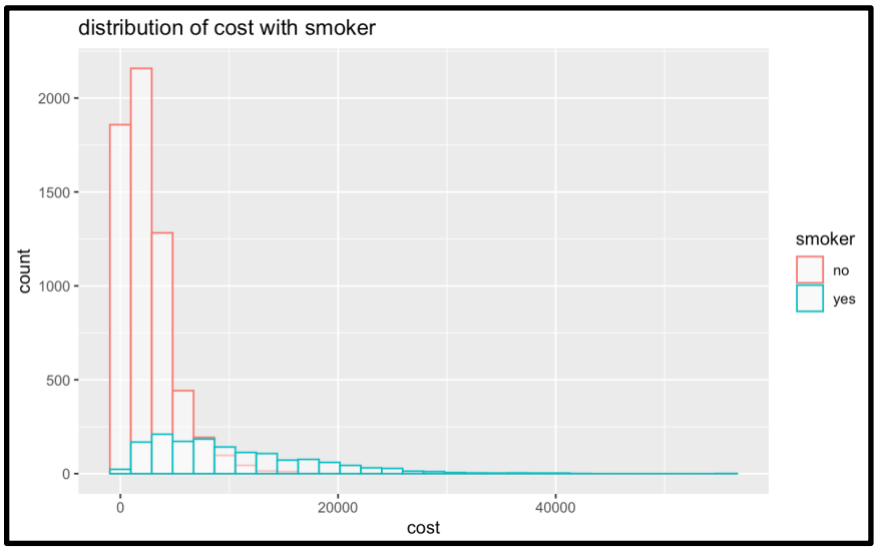


As per our research, some outliers represent natural variations in the population, and they should be left as is in the dataset. These are called true outliers.

4.5 Histogram for distribution analysis



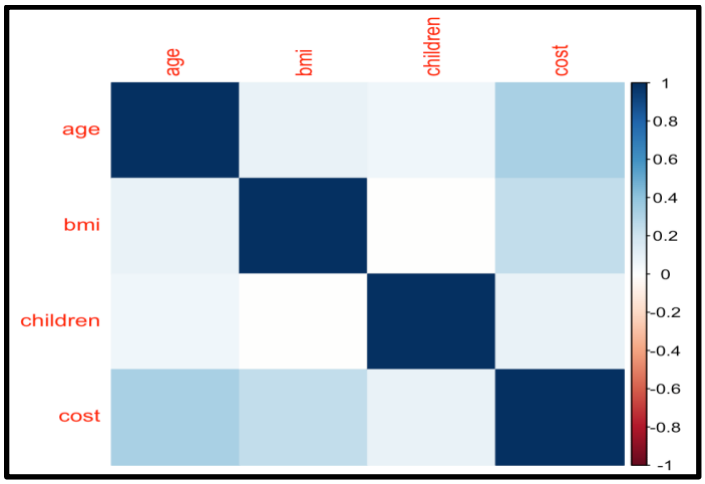
From the above generated histograms, we can see that the red coloured bars represent the non-expensive clients and the blue bars represent the expensive clients. Both the red and blue bars follows uniform distribution and the red bar is more concentrated and have major peaking at 26 - 28 BMI index which means that more number of non-expensive clients are Overweight and when looking at the blue bar , it is majorly concentrated and has peaking at 34-36 BMI index which means that more number of expensive clients are having the problem of Obesity.



4.6 Correlation analysis

Correlation Between numerical Variables:

For the correlation , We have used the Spearman method for age , bmi , children , cost variables because the data of these variables are which are plotted are not symmetrically distributed and there is skewness present in this data of these variables. The correlation index ranges between -1 to +1 and +1 -represents strong positive correlation . -1 -represents strong negative correlation. Before diving into detailed correlation using scatterplot, a correlation matrix is used.



Now , talking about correlation between the variables -

1] Age-

slightly positive correlated with BMI

weak positive correlated with children

good positive correlated with cost

2] BMI -

slightly positive correlated with BMI

weak positive correlated with children

good positive correlated with cost

3] children -

weak positive correlated with Age

weak positive correlated with cost

good positive correlated with BMI

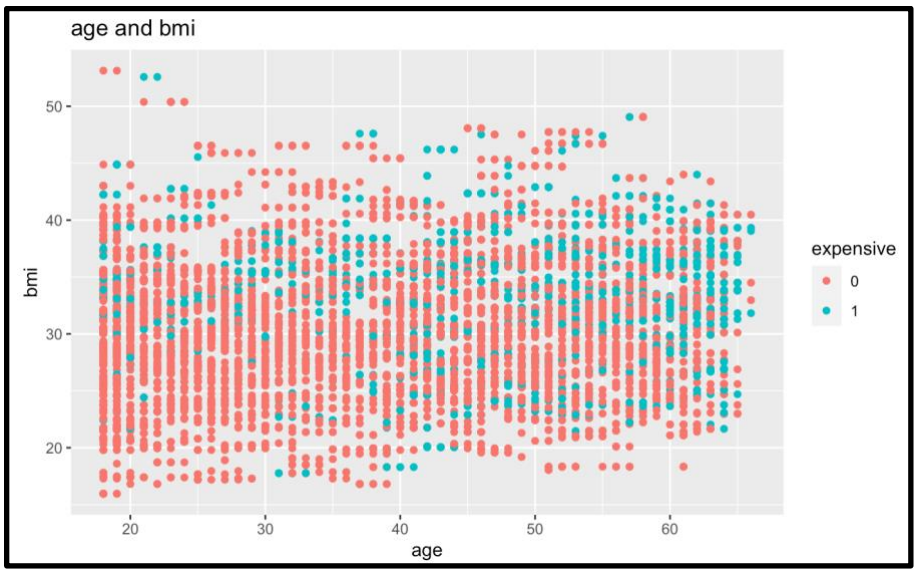
4] cost -

weak positive correlated with children

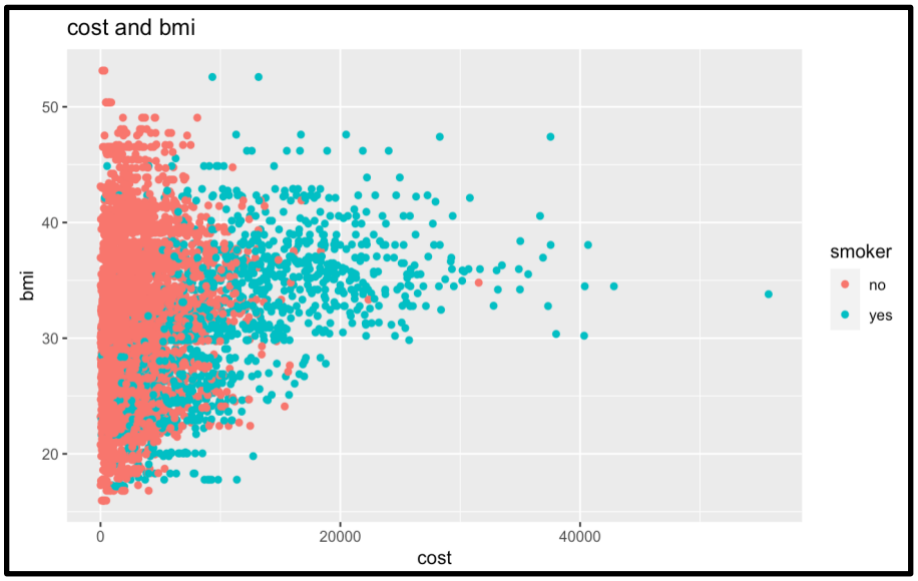
good positive correlated with BMI

good positive correlated with Age

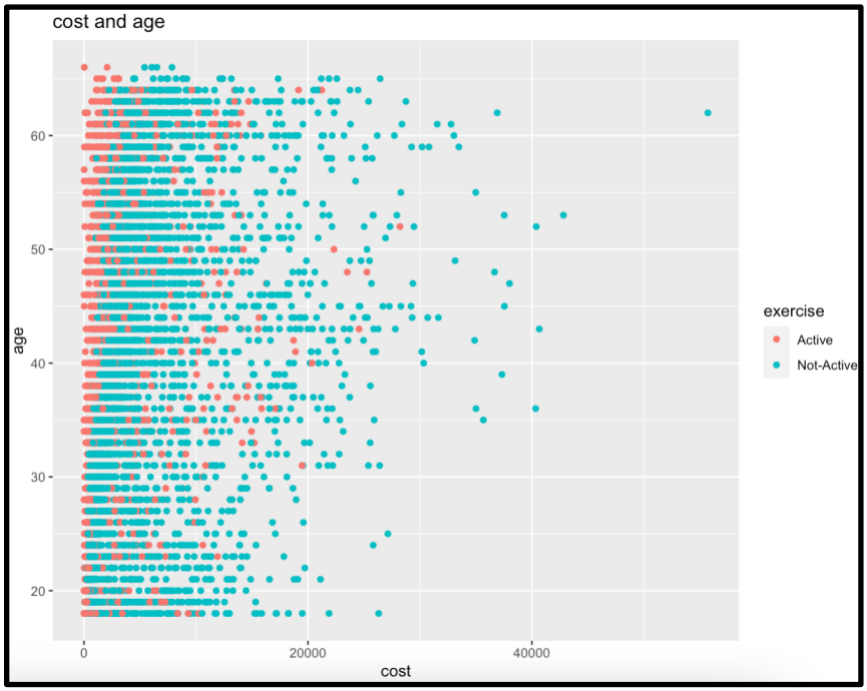
4.6.2 Scatterplots



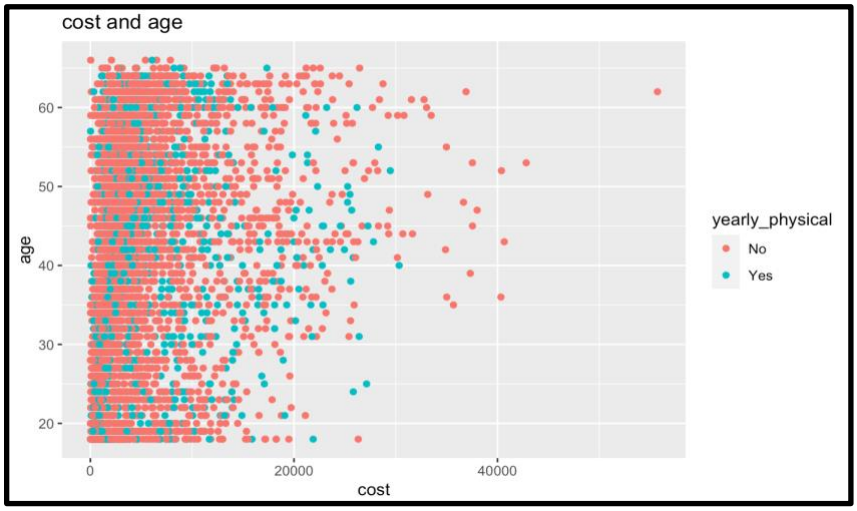
In the above scatterplot, we have tried to determine any possible relationship between the variables **Age, BMI and Cost**. However, the scatterplot does not seem to provide much information as there are a lot of overlapping points and no display of positive or negative correlation, thus it was not a helpful visual.



In the above scatterplot, we have tried to determine a possible relationship between variables **Smoker, BMI and Cost.** According to healthcare experts, A healthy BMI value lies between 18.5 to 24.9. As per the above graph, a person who has a higher BMI, and who is also a smoker, has a higher probability of being an expensive candidate for the HMO.



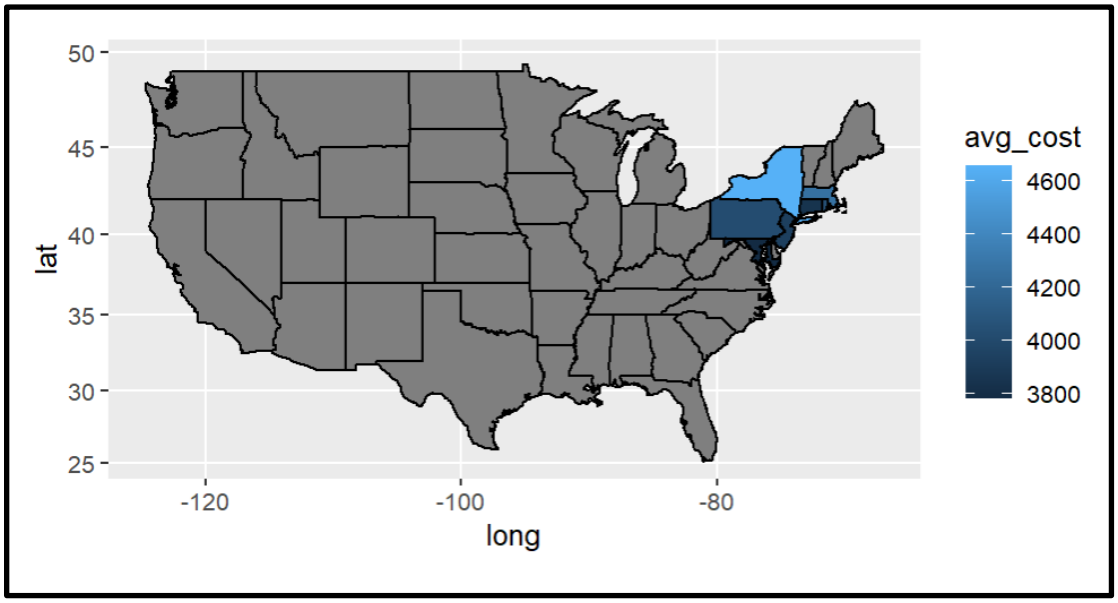
In the above scatterplot, we tried to determine the relationship between the variables **Cost, Age and Exercise**. Based on the scatterplot, candidates who are not active and have a higher age, tend to be more expensive.



In the above graph we have tried to determine a possible relationship between the factors Age, Yearly Physical and Cost. The scatterplot does not show a strong correlation between these variables, but there are multiple scatterplot points that display that candidates with a higher age, who have not taken a yearly physical exam seem to be costlier candidates.

-

4.7 Map for cost distribution



The dataset covers only the North eastern parts of the United States. From the dataset the average cost in New York State is relatively low compared to the other states like Pennsylvania, Connecticut, Massachusetts.

Data partition for training and test

With all the information from descriptive statistics, we will now use it to work on our predictive models and we will be using *expensiveness* as the dependent variable. The models with be fed by 70 percent of overall observations which work as training dataset, and the model will be tested with the rest 30 percent for predictive accuracy.

SECTION 5: PREDICTIVE MODELS

5.1 Linear regression model

Initially, we performed linear regression to get a general idea of the data and determine the relationship between the dependent and independent variables. Based on the results of Linear regression, the variables Age, BMI, Children, Smoker, Exercise and Hypertension are the most important variables that have a high effect on determining the amount of money an individual spends on healthcare.

```
Call:
lm(formula = cost ~ ., data = trainlinear)

Residuals:
    Min      1Q  Median      3Q     Max
-11869  -1485    -352    1010   41793

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -8943.440    322.429 -27.738  < 2e-16 ***
age                               100.107      3.148  31.800  < 2e-16 ***
bmi                               179.693      7.414  24.236  < 2e-16 ***
children                          222.143     36.278   6.123 9.83e-10 ***
smokeryes                        7588.650    110.701  68.551  < 2e-16 ***
locationMARYLAND                 -189.114    212.037  -0.892   0.3725
locationMASSACHUSETTS             -15.560    235.956  -0.066   0.9474
locationNEW JERSEY                  3.026    232.112   0.013   0.9896
locationNEW YORK                  373.227    226.288   1.649   0.0991 .
locationPENNSYLVANIA              -55.003    167.231  -0.329   0.7422
locationRHODE ISLAND              147.200    213.666   0.689   0.4909
location_typeUrban                  4.685    101.815   0.046   0.9633
education_levelMaster             -47.127    113.026  -0.417   0.6767
education_levelNo College Degree   -4.301    149.167  -0.029   0.9770
education_levelPhD               -318.861    154.361  -2.066   0.0389 *
yearly_physicalYes                139.814    103.336   1.353   0.1761
exerciseNot-Active               2232.826    102.649  21.752  < 2e-16 ***
marriedNot_Married                 91.615     94.280   0.972   0.3312
hypertension1                     445.120    110.962   4.011 6.12e-05 ***
gendermale                         63.266     89.084   0.710   0.4776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3216 on 5287 degrees of freedom
Multiple R-squared:  0.5716,    Adjusted R-squared:  0.57
F-statistic: 371.3 on 19 and 5287 DF,  p-value: < 2.2e-16


        Pearson's product-moment correlation

data:  linearpredict and testlinear$cost
t = 56.012, df = 2273, p-value < 2.2e-16
```

When we observe the median which is -352 which tells that it is not symmetrically distributed and it shows some skewness. The standard errors around the estimates of slope and intercept show the estimated spread of the sampling distribution around these point estimates. The adjusted R -Squared is 57 % which means that the Age, BMI, Children, Smoker, Exercise and Hypertension, gender (predictors) accounts to 57% variability of cost variable (independent variable).

When talking about the major B-Weights contributors - Exercise , Location(Maryland) , Hypertension , Age , Bmi , Children played the major role in creating the effect on the cost variable.

When talking about P-value [<2.2e-16] - It is statistically significant as it is lower than the 0.05 (Significance value) hence , we cannot reject the Null Hypothesis.

5.2 Logit Regression

This algorithm is generally a statistical model that models the probability of the event taking place by having log-odds for the event be a linear combination of one or more independent variables. It is based on the inverse logit function. This Logit function comes from the family of Generalized linear models and it is very helpful in predicting Categorical values such as TRUE or False , 0s or 1s (Binary Outputs).

We got an accuracy of about 83% which is good in predicting the Expensive variable and tells us that the Age, BMI, Children, Smoker, Exercise and Hypertension, gender helps in change of variability of expensive variable.

```
Call:
glm(formula = expensive ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7814  -0.4455  -0.2200  -0.0731   3.2764

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -11.917968   0.471027 -25.302  < 2e-16 ***
age                            0.069124   0.003975  17.388  < 2e-16 ***
bmi                            0.133258   0.008426  15.816  < 2e-16 ***
children                       0.135060   0.038716   3.488 0.000486 ***
smokeryes                      4.267069   0.129429  32.968  < 2e-16 ***
locationMARYLAND              -0.151340   0.238389  -0.635 0.525531
locationMASSACHUSETTS          0.067962   0.249105   0.273 0.784989
locationNEW JERSEY             0.168438   0.254305   0.662 0.507749
locationNEW YORK               0.581538   0.236283   2.461 0.013848 *
locationPENNSYLVANIA           0.095185   0.182632   0.521 0.602240
locationRHODE ISLAND          -0.011453   0.234453  -0.049 0.961040
location_typeUrban            -0.098517   0.109483  -0.900 0.368210
education_levelMaster          0.033877   0.122909   0.276 0.782832
education_levelNo College Degree  0.257033   0.160395   1.603 0.109045
education_levelPhD             0.015110   0.161787   0.093 0.925589
yearly_physicalYes             0.277277   0.109285   2.537 0.011174 *
exerciseNot-Active             2.024991   0.142359  14.224  < 2e-16 ***
marriedNot_Married             0.172345   0.099930   1.725 0.084588 .
hypertension1                  0.357006   0.115608   3.088 0.002015 **
gendermale                    -0.021435   0.096747  -0.222 0.824659
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

When talking about the P- significance test , we can see that the P-value is <2e16 which is conventionally lower than 0.05 [Significance value] , hence it is significant.

```
[1] 2275
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1212  209
         1  170  684

               Accuracy : 0.8334
                 95% CI : (0.8174, 0.8485)
    No Information Rate : 0.6075
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.648

 Mcnemar's Test P-Value : 0.05095

            Sensitivity : 0.8770
            Specificity : 0.7660
         Pos Pred Value : 0.8529
         Neg Pred Value : 0.8009
             Prevalence : 0.6075
         Detection Rate : 0.5327
   Detection Prevalence : 0.6246
      Balanced Accuracy : 0.8215

       'Positive' Class : 0
```

5.3 Neural networks

Neural networks, also known as artificial neural networks, are a part of machine learning and have gradually become very important in deep learning algorithms. A neural network model is composed of an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an according weight and threshold. With its special characteristic of hidden layers, the neural network model provides more flexibility and profoundness in exploring the data pattern. In this HMO data consulting case, we are going to put training data into a neural network model and the algorithm learns and improves its accuracy over time

Neural network data preprocessing

Neural networks require the input variables to be in numeric form. In the HMO dataset, most of our variables, namely 'location', 'location_type', 'Exercise', Smoker', 'yearly_physical', 'Hypertension', 'Gender', 'Education_level', 'Married', are categorical. Since Neural networks only consider pre-processed data, unlike linear and logistic regression models whose algorithm automatically converts categorical variables into dummy variables, we manually converted the categorical variables into a numeric format before using it as input for the model. Finally, we used the following variables as an input for the model – 'Age',' bmi', 'children', 'smoker', 'exercise' and 'hypertension'. We created dummy variables for categorical variables and then left out one class of each dummy variable to prevent redundant information.

Neural network with two nodes in one layer

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1797  166
         1   47  265

               Accuracy : 0.9064
                 95% CI : (0.8937, 0.918)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16
```

With two nodes one layer neural network models, we made 2062 correct predictions, which consisted of 1792 correct non-expensive prediction and 265 correct expensive prediction, out of total 2275 testing observations. The predictive accuracy is 0.9064 and with 0.8937 to 0.918 accuracy with 95% confidence interval.

Neural network with three nodes in one layer.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1806  168
         1   38  263

               Accuracy : 0.9095
                 95% CI : (0.8969, 0.9209)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16
```

With three nodes one layer neural network models, we made 2069 correct predictions, which consisted of 1806 correct non-expensive prediction and 263 correct expensive prediction, out of total 2275 testing observations. The predictive accuracy is 0.9095 and with 0.8969 to 0.9209 accuracy with 95% confidence interval. From the testing result, we have the basic conclusion that the three nodes model performed slightly better than the two node model.

Neural network with four nodes in one layer.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1811  168
         1   33  263

               Accuracy : 0.9116
                 95% CI : (0.8992, 0.923)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16
```

With a four-node and one-layer neural network model, we made 2074 correct predictions, which consisted of 1811 correct non-expensive predictions and 263 correct expensive predictions, out of a total of 2275 testing observations. The predictive accuracy is 0.9116 and with 0.8992 to 0.923 accuracy with 95% confidence interval. From the testing result, four nodes model performed the best so far.

5.4 Association Rules

To generate the association rules for a particular dataset, important prerequisites are that the variables have to be of data type factor and the dataset needs to be converted to transaction. Meanwhile, unlike other predictive models, association rule analysis focuses more about data patterns appearing together. So we are not going to split the data into training and testing. Instead we will put all the data into the association rule model and find the rules with most instances. After all necessary data preprocessing for association rule analysis, we generated rules with a Support value of 0.05 and Confidence value of 0.70, we obtained 19 association rules that help us understand what the important variables are when predicting expensive clients.

```
[8]  {smoker=yes,
      yearly_physical=No,
      exercise=Not-Active}       => {expensive=1} 0.08150884  0.7463768 0.10920601 3.730408   618
[9]  {smoker=yes,
      exercise=Not-Active,
      hypertension=0}            => {expensive=1} 0.08599314  0.7502877 0.11461356 3.749955   652
[10] {smoker=yes,
      location_type=Urban,
      exercise=Not-Active,
      gender=male}               => {expensive=1} 0.05170140  0.7951318 0.06502242 3.974087   392
[11] {smoker=yes,
      yearly_physical=No,
      exercise=Not-Active,
      gender=male}               => {expensive=1} 0.05130572  0.7826962 0.06554999 3.911933   389
[12] {smoker=yes,
      exercise=Not-Active,
      hypertension=0,
      gender=male}               => {expensive=1} 0.05367977  0.7857143 0.06831970 3.927018   407
[13] {smoker=yes,
      education_level=Bachelor,
      exercise=Not-Active,
      hypertension=0}            => {expensive=1} 0.05038248  0.7519685 0.06700079 3.758355   382
```

```
     lhs                            rhs              support confidence   coverage     lift count
[1]  {smoker=yes,
      exercise=Not-Active}       => {expensive=1} 0.10894223  0.7605893 0.14323398 3.801442   826
[2]  {smoker=yes,
      married=Married,
      gender=male}               => {expensive=1} 0.05565814  0.7068677 0.07873912 3.532940   422
[3]  {smoker=yes,
      exercise=Not-Active,
      gender=male}               => {expensive=1} 0.06831970  0.7932619 0.08612503 3.964741   518
[4]  {smoker=yes,
      location=PENNSYLVANIA,
      exercise=Not-Active}       => {expensive=1} 0.05803218  0.7586207 0.07649697 3.791603   440
[5]  {smoker=yes,
      education_level=Bachelor,
      exercise=Not-Active}       => {expensive=1} 0.06475864  0.7600619 0.08520179 3.798807   491
[6]  {smoker=yes,
      exercise=Not-Active,
      married=Married}           => {expensive=1} 0.07465049  0.7753425 0.09628066 3.875179   566
[7]  {smoker=yes,
      location_type=Urban,
      exercise=Not-Active}       => {expensive=1} 0.08177262  0.7598039 0.10762332 3.797517   620
```

```
[14] {smoker=yes,
      location_type=Urban,
      exercise=Not-Active,
      married=Married}          => {expensive=1} 0.05724083  0.7695035 0.07438671 3.845996   434
[15] {smoker=yes,
      yearly_physical=No,
      exercise=Not-Active,
      married=Married}          => {expensive=1} 0.05552625  0.7626812 0.07280401 3.811898   421
[16] {smoker=yes,
      exercise=Not-Active,
      married=Married,
      hypertension=0}           => {expensive=1} 0.05816407  0.7696335 0.07557373 3.846646   441
[17] {smoker=yes,
      location_type=Urban,
      yearly_physical=No,
      exercise=Not-Active}      => {expensive=1} 0.05961488  0.7446458 0.08005803 3.721756   452
[18] {smoker=yes,
      location_type=Urban,
      exercise=Not-Active,
      hypertension=0}           => {expensive=1} 0.06436296  0.7484663 0.08599314 3.740851   488
[19] {smoker=yes,
      yearly_physical=No,
      exercise=Not-Active,
      hypertension=0}           => {expensive=1} 0.06502242  0.7391304 0.08797151 3.694190   493
```

For the first rule, we have our determining factors and their subsequent values as given:

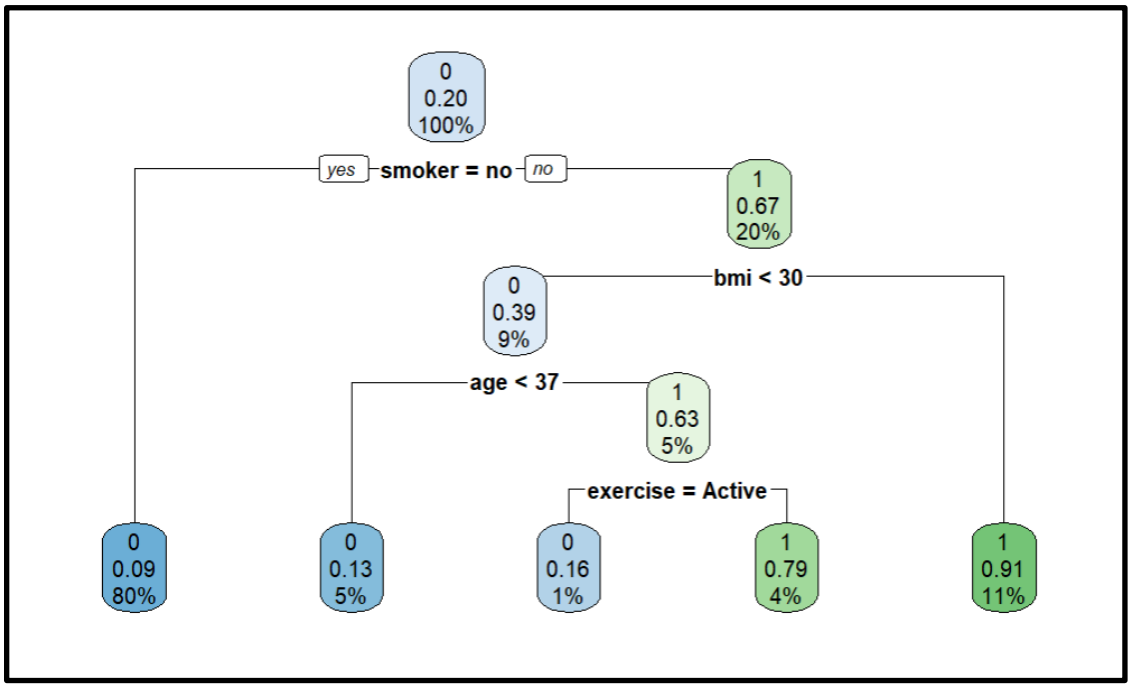
expensive = 1, smoker = yes + exercise=Not-Active. This rule is describing the expensive clients who are smokers and are not active with exercise. This type of clients had a support of approximately 11% (0.109), which means that they were in the dataset for about 11% of the time, and had a confidence of 0.761, This rule has the highest Lift value as 3.801, which also supports an association between our antecedent and consequent. Smokers with not-active exercise status are the most possible clients that will make expensive cost.

For the second rule, we had identical factors as the first rule with the addition of a new factor married = married. This had a Support of approximately 6% (0.056) as well and confidence of 0.707 and a slightly lower Lift value of 3.533, which still indicating an association between the smoker, not-active exerciser and expensive client, but this time, we also know those clients are possibly married.

In this way, we could observe the rules and found a lot of common factors in most of them, such as the smoker, exercise as Not-Active, hypertension as 0, and yearly_physical as No, which means a lot of expensive clients share the same life patterns, including smoking, not active exercise status, not having hypertension and not going to yearly physical check. Lowering the confidence level value and generating more rules could give us a deeper idea of more factors that led to expensive cases.

5.5 Decision Tree

Decision Trees are versatile Machine Learning algorithms which can perform both classification and regression tasks. We can visualize each decision made which makes decision tree a great tool for decision analysis. To train a decision tree we are splitting the data into train and test set. Since its easy to understand decision visualization for classification models we are classifying the cost into expensive as 1 and non-expensive as 0. We get the following decision tree plot for our model:



The above model gives the decision based on the most correlated variables age,BMI,exercise,smoker.It is clear that out of the total population the probability of a person being non-expensive is 80 percent while the remaining 20% are expensive if the person is a smoker and their probability is 67 percent. And as the node branches to smokers, it checks the BMI of the person if the BMI is less than 30 and smoker then the probability of that person's health cost being non-expensive is 39 percent and they cover 9 percent of the population under smokers. While if the person has a BMI above 30 the possibility that the person's medical cost will be expensive is 91 percent and 11 percent out of the total expensive population . And further observations from under the node BMI less than 30 suggest that a person being less than 37 years old will be inexpensive and the probability of that occurrence is 13 percent and they cover 5 % of the total expensive pool. While a person above 37 years with BMI less than 30 might be expensive and the probability of that occurrence is 63% and they include 5% of the population. Further analysis from the tree suggests that people with smoking issues, BMI less than 30, and who are in their late thirties or above and don't have an exercise regimen might be expensive and the possibility of that is 79 percent and they cover 4% out of the non-exercise population.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1814  172
         1   30  259

               Accuracy : 0.9112
                 95% CI : (0.8988, 0.9226)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6691

 Mcnemar's Test P-Value : < 2.2e-16
```

With the decision tree we made an accuracy of 91.12% and No Information Rate of 0.8105. For 95% confidence intervals the predictive accuracy ranges between 89.88 % to 92.26% .

5.6 Random Forest

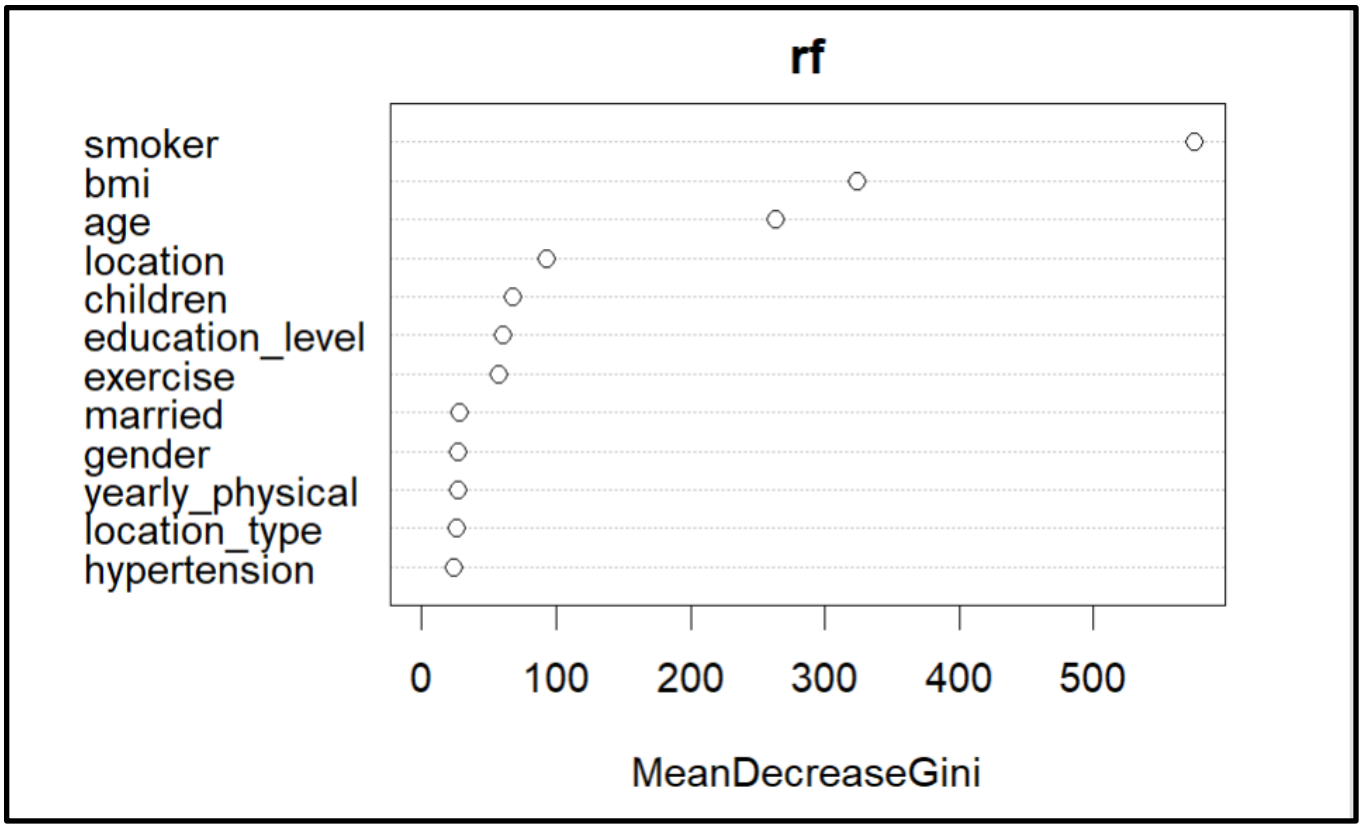
Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The Random forest for the dataset generated 500 trees. WIth 3 variables tried at each split.

```
Call:
 randomForest(formula = expensive ~ ., data = train, proximity =
TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 9.36%
Confusion matrix:
     0   1 class.error
0 4121 100  0.02369107
1  397 689  0.36556169
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1818  155
         1   26  276

               Accuracy : 0.9204
                 95% CI : (0.9085, 0.9312)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7074

 Mcnemar's Test P-Value : < 2.2e-16
```

The random forest gave an accurate prediction of 92.04%, with No Information Rate of 0.8105.And it is clear that the random forest gives slightly greater accuracy than the decision tree model. The 95% confidence interval for Random Forest ranges from 90.85% to 93.12%.

We can also get a clear perspective of the variables with more important variables in the model with the 'varImpPlot' function.



From the above plot the variables such as age, BMI, smoker are more relevant variables to make predictions for cost.

5.7 SVM(Support Vector Machine)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression. For building the support vector machine we divided the dataset into train and test taking 70 percent in train set and remaining 30 percent in test set. The model is built to predict whether a person's medical cost is expensive or inexpensive, by the expensive variable as our dependent variable. We obtained the following finding:

Iteration 1:

For iteration 1 we are considering all variables and it gives as an accuracy of 90.4 percent with NIR of 0.8105 and 95 percent confidence interval in the range of 0.8913 to 0.916

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1800  174
         1   44  257

               Accuracy : 0.9042
                 95% CI : (0.8913, 0.916)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6472

 Mcnemar's Test P-Value : < 2.2e-16
```

Iteration 2:

For this iteration we are considering variables which have correlation such as age,BMI,children,Smoker, Yearly Physical, Exercise,Hypertension and Marriage Status. The SVM for these variables gave an accuracy of 90.73 percent with NIR of 0.8105 and 95 percent confidence interval in the range of 0.8946 to 0.9189.

```
[1] "SVM For age, BMI , Children, Smoker, Yearly Physical,
Excercise,Hypertension,Married"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1808  175
         1   36  256

               Accuracy : 0.9073
                 95% CI : (0.8946, 0.9189)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6554

 Mcnemar's Test P-Value : < 2.2e-16
```

Iteration 3:

For iteration 3 we removed marriage status variable and got an accuracy of 90.68 percent with NIR of 0.8105 and 95 percent confidence interval in the range of 0.8941 to 0.9184

```
[1] "SVM without married variable"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1809  177
         1   35  254

               Accuracy : 0.9068
                 95% CI : (0.8941, 0.9184)
    No Information Rate : 0.8105
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6527

 Mcnemar's Test P-Value : < 2.2e-16
```

SECTION 6: MODEL COMPARISON

MODEL	VARIABLES USED	ACCURACY	P-VALUE?/NIR
Linear Regression	X, age, location, location_type, exercise, Smoker,bmi,yearly_physical, Hypertension, Gender, Education_level Married, Num_children, Cost	**57%**	**<2.2 e-16**
Logistic Regression	X, age, location, location_type, exercise, Smoker,bmi,yearly_physical, Hypertension, Gender, Education_level Married, Num_children, Cost	**83%**	**<2 e-16**
Neural Networks	age, bmi, smoking, Hypertension, Exercise, Children	**91.8% (two nodes) 92.0% (3 nodes)**	**<2.2 e-16**
Decision Trees	X, age, location, location_type, exercise, Smoker,bmi,yearly_physical, Hypertension, Gender, Education_level Married, Num_children, Cost	**91.12%**	**<2.2 e-16**
Random Forests	X, age, location, location_type, exercise, Smoker,bmi,yearly_physical, Hypertension, Gender, Education_level Married, Num_children, Cost	**92.42%**	**<2.2 e-16**
SVM	age,BMI,children,Smoker,Yearly_Physical, Exercise,Hypertension and Marriage Status	**90.73%**	**<2.2 e-16**

SECTION 7: CONCLUSION AND RECOMMENDATIONS

Conclusion

With goals of understanding expensive clients and reducing the cost, we found that the clients who tend to be top 20% in cost have the characteristics of the following

1. Age has a positive impact on the cost, which means the older the clients are, the more possible for them to be an expensive client
2. Higher bmi also leads to higher costs. Body mass index (BMI) is a person's weight in kilograms divided by the square of height in meters. BMI is an inexpensive and easy screening method for weight categories—underweight, healthy weight, overweight, and obesity. The bmi value bigger than 25.0 usually indicates overweight.
3. For other health indicators, clients who smoke, have hypertension, not being active with exercise, not going to regular physical checks usually have higher medical expenditures.

Recommendation

Having different price points for customers based on certain factors like Smoking, Married, etc. would be beneficial for the HMOs. Since our analysis proves that customers with certain habits are more prone to health issues, identifying such customers through health questionnaire forms, examining previous health records, and suggesting a plan with a low, medium, or high premium based on this would be helpful.

Since age is a significant variable to predict cost based on our analysis, providing services like yearly health checkup reminders would be beneficial in detecting any health issues early, thus reducing future healthcare costs.

Partnering with major fitness platforms and offering subsidized online fitness and mindfulness services to clients would also help in improving the overall health of the client, and reducing the risk of healthcare expenditure, thus saving money for the HMO.

Offer family (adult + children) plans, at a slightly lower rate, but still beneficial for the HMO because it is highly unlikely that all members of the family will have health issues at the same time, but it will ensure that everyone in the family joins the same HMO.

SECTION 8: APPENDIX - METADATA

VARIABLE	DESCRIPTION
X	Type: Integer Unique identifier for each person
age	Type: Integer Age of the person (at the end of the year)
location	Type: Categorical State in which the person lives
location_type	Type: Categorical Describes the environment in which the person lives Urban Country
exercise	Type: Categorical Describes if the person exercised regularly in the last year. Active Not Active
Smoker	Type: Categorical Describes if the person smoked during the last year. Yes No
bmi	Type: Integer Describes the body mass index of the person. The BMI is a measure that uses height and weight to determine if the individual is healthy.
yearly_physical	Type: Categorical Describes if the person had a yearly check up with their doctor. Yes No
Hypertension	Type: Categorical Describes if the person had Hypertension Yes No
Gender	Type: Categorical Describes gender of the person Male Female

Education_level	Type: Categorical Describes the level of college education the person received. No College Degree Bachelor Master PhD
Married	Type: Categorical Describes marital status of the person. Married Not Married
Num_children	Type: Integer Contains information about the number of children the individual has.
Cost	Type: Integer Contains the total cost of healthcare for that person in the past year.

SECTION 9: REFRENCES

- https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20046974
- -https://stats.stackexchange.com/questions/140401/quantifying-data-completeness-in-healthcare
- https://www.investopedia.com/terms/h/hmo.asp
- https://www.gao.gov/products/hrd-82-31
-