



IST 687 INTRODUCTION TO DATA SCIENCE

# **Actionable Insights for Health Management Organization (HMO) Report**

**Data Analysis/Predictive modeling-HMO dataset**

**Final project**

**December 2022**

**SUBMITTED BY**

Samarth Mengji | Atharva Vakharkar | Shreya Zope | Somia abdulrehman

## Table of contents

### Contents

<b>1. Description .....</b>	<b>3</b>
<b>2. Project Scope and Objective.....</b>	<b>3</b>
<b>3. Project Deliverables .....</b>	<b>3</b>
<b>4. Business Questions.....</b>	<b>4</b>
<b>5. Data Acquisition .....</b>	<b>5</b>
<b>6. Data cleaning &amp; preprocessing .....</b>	<b>6</b>
<b>7. Descriptive Statistics and Visualizations .....</b>	<b>9</b>
<b>8. Modelling techniques .....</b>	<b>19</b>
<b>9. Interpretations and Actionable Insights .....</b>	<b>27</b>
<b>10. Project Link .....</b>	<b>28</b>

## 1. Description

This project revolves around analyzing healthcare data of huge number of people to provide actionable insight for certain Health Management Organization (HMO), based on the data available, as well as accurately predict which people (customers) will be expensive. The dataset contains healthcare cost information from certain HMO. (Each row in the dataset represents a person. Our goal is to determine why some people are more expensive in terms of health costs based on different health and lifestyle attributes.

## 2. Project Scope and Objective

The scope of this project is to analyze and draw insights from the dataset provided, which contains data regarding person's health based on various factors. The dataset contains healthcare cost information. Each row in the dataset represents a person. There is data of around 7000 people.

We will be concentrating on achieving two high level goals. Predicting people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs), in addition to providing actionable insight and recommendations to the HMO, in terms of how to lower their total health care costs.. Additionally, the objective of this project is to suggest to HMO's patients, the areas where they can improve to decrease their total health care costs.

## 3. Project Deliverables

1. Perform data cleaning to prepare the data for further analysis so that there are no missing or invalid fields in the dataset.
2. Identify the attributes that most affects the health care cost of the customers by applying linear regression/correlation/visualization of the cost distribution across different attributes.
3. Predicting if the patient is expensive or not based on selected attributes by applying support vector machine and formulate actionable insights.
4. Predicting if the patient is expensive or not based on selected attributes by applying decision tree and formulate actionable insights.
5. Building the shiny APP based on the best model achieved.

6. Finally, provide suggestions to the client based on data analysis/predictive modeling and interpretation to enhance and lower their total health care costs. Especially targeting the less satisfied group of people.

## **4. Business Questions**

A vast healthcare and lifestyle data of 7582 people was provided which gives information regarding the factors that drive the overall healthcare cost of that person. Analyzing these factors would help HMO and people to lower their total health care costs.

The business questions that have been recognized and answered through the project are as follows:

1. How does BMI, Age, Exercise, Smoker impact the overall healthcare costs?
2. Why do some parameters like children, location, location type, education, married, gender not affect the overall healthcare cost?
3. Which aspect of person affects the health care cost the most?
4. How can a customer save on healthcare costs?
5. What kind of people will spend the most on health care costs?
6. Why do some people have high healthcare costs?
7. What can HMO do to lower the health care expenses cost?

## 5. Data Acquisition

The data set was made available to us by the course instructors in the format of CSV file

This data set consisted of approximately 7583 rows of the people with varied health and lifestyle attributes with 15 fields such as Age, bmi, gender, children, smoker, location, location\_type, education\_level, yearly\_physical etc.

This data was extensively studied to determine the usable variables. After this initial analysis, the data set was forwarded to the preprocessing phase where all the errors in the data were removed in order to make it usable for further analysis.

### Snapshot of the data

```
{r}  
data <- data.frame(read_csv('HMO_data.csv'))  
head(data)
```

Description: df [6 × 14]

	X <dbl>	age <dbl>	bmi <dbl>	children <dbl>	smoker <chr>	location <chr>	location_type <chr>
1	1	18	27.900	0	yes	CONNECTICUT	Urban
2	2	19	33.770	1	no	RHODE ISLAND	Urban
3	3	27	33.000	3	no	MASSACHUSETTS	Urban
4	4	34	22.705	0	no	PENNSYLVANIA	Country
5	5	32	28.880	0	no	PENNSYLVANIA	Country
6	7	47	33.440	1	no	PENNSYLVANIA	Urban

6 rows | 1-8 of 14 columns

```
{r}  
str(data)
```

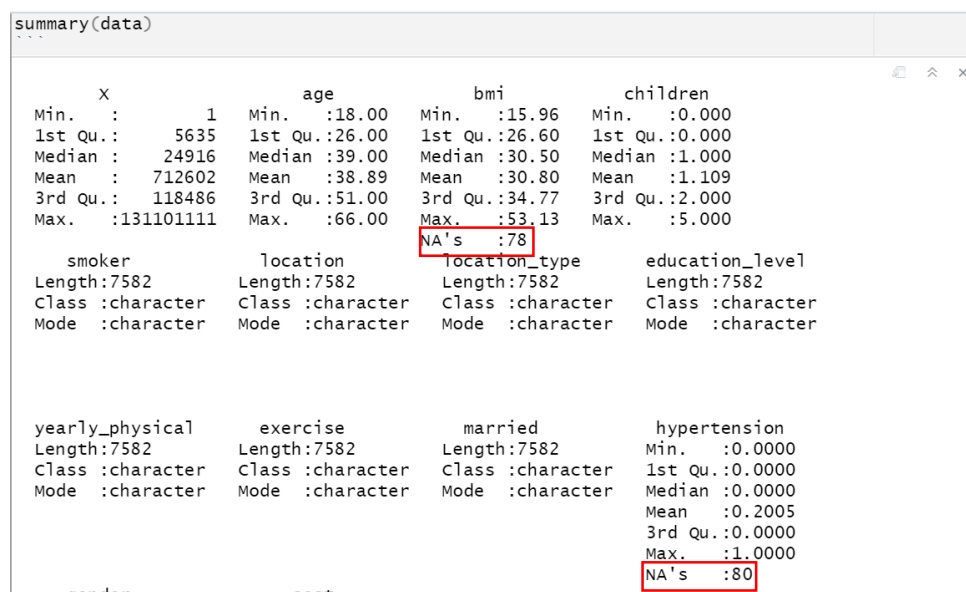
```
'data.frame': 7582 obs. of 14 variables:  
 $ X          : num  1 2 3 4 5 7 9 10 11 12 ...  
 $ age         : num  18 19 27 34 32 47 36 59 24 61 ...  
 $ bmi         : num  27.9 33.8 33 22.7 28.9 ...  
 $ children    : num  0 1 3 0 0 1 2 0 0 0 ...  
 $ smoker      : chr   "yes" "no" "no" "no" ...  
 $ location    : chr   "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...  
 $ location_type : chr   "Urban" "Urban" "Urban" "Country" ...  
 $ education_level : chr   "Bachelor" "Bachelor" "Master" "Master" ...  
 $ yearly_physical : chr   "No" "No" "No" "No" ...  
 $ exercise     : chr   "Active" "Not-Active" "Active" "Not-Active" ...  
 $ married      : chr   "Married" "Married" "Married" "Married" ...  
 $ hypertension : num  0 0 0 1 0 0 0 1 0 0 ...  
 $ gender       : chr   "female" "male" "male" "male" ...  
 $ cost         : num  1746 602 576 5562 836 ...
```

## 6. Data cleaning & preprocessing

### 6.1 Addressing Null values

The dataset consisted of 7583 rows and 15 column variables. It includes different attributes of a person's health and lifestyle like smoking, BMI, age, etc.

Firstly, all the columns of the data set were summarized to determine null values. It was found that BMI and Hypertension column had 78 and 80 null values as shown below.



```
summary(data)
```

X	age	bmi	children
Min. : 1	Min. :18.00	Min. :15.96	Min. :0.000
1st Qu.: 5635	1st Qu.:26.00	1st Qu.:26.60	1st Qu.:0.000
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.:34.77	3rd Qu.:2.000
Max. :131101111	Max. :66.00	Max. :53.13	Max. :5.000
		NA's :78	

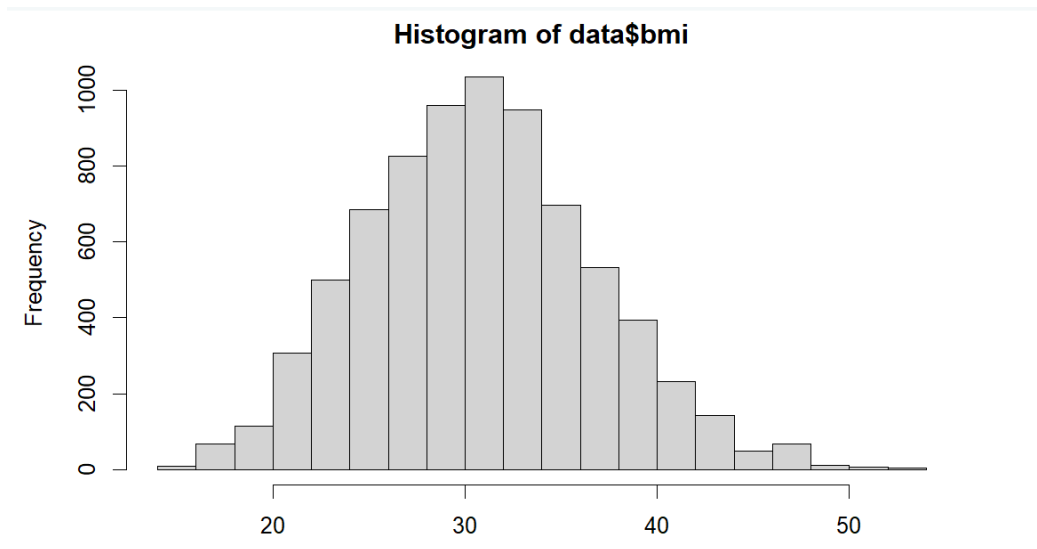
  

smoker	location	location_type	education_level
Length:7582	Length:7582	Length:7582	Length:7582
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

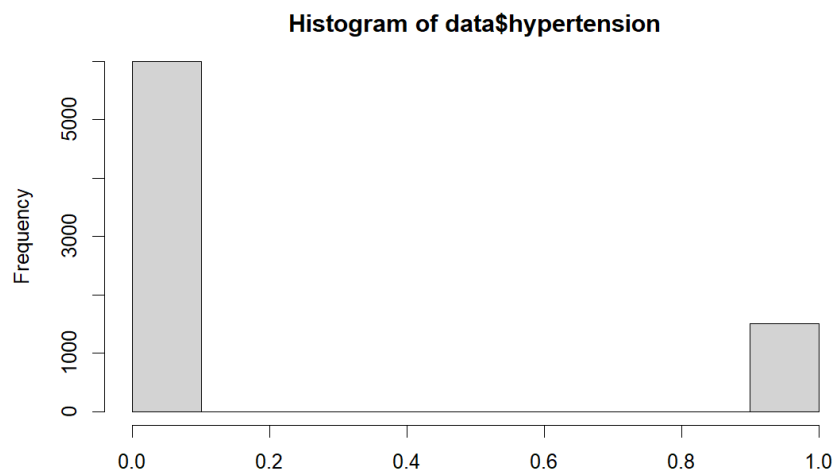
  

yearly_physical	exercise	married	hypertension
Length:7582	Length:7582	Length:7582	Min. :0.0000
Class :character	Class :character	Class :character	1st Qu.:0.0000
Mode :character	Mode :character	Mode :character	Median :0.0000
			Mean :0.2005
			3rd Qu.:0.0000
			Max. :1.0000
			NA's :80

In order to clean the data from the null values, the `replace_na()` function was used to replace the null with the average value for BMI attribute. This value was chosen because this attribute shows normal distribution.



For the hypertension attributes (1 = hypertension, 0 = no hypertension), the null value was replaced by the mode – 0. This due to the nature of this attribute as it can be seen from the plot below.

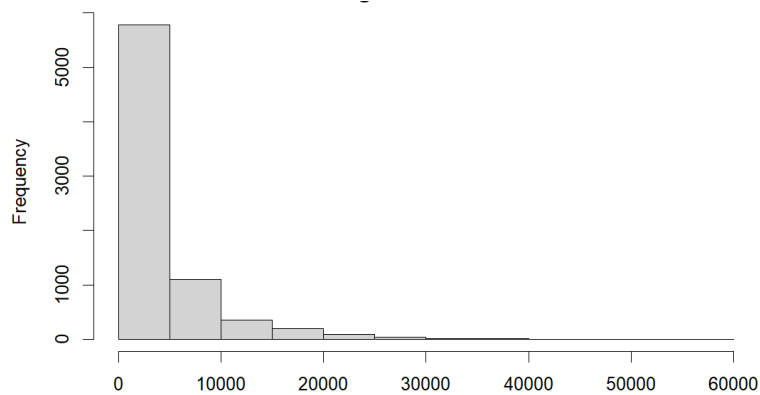


After data munging the number of rows in the data set were 7583 and had 15 columns.

## **6.2 Expensive Attribute:**

The expensive attribute was derived from the cost column

Any cost above 5000 was considered/labeled expensive in the dataset. This threshold was specified by the professor.



Below code was used to create and derive the expensive column:

```
data <- transform(  
  data, expensive= ifelse(cost > 5000, TRUE,FALSE))
```

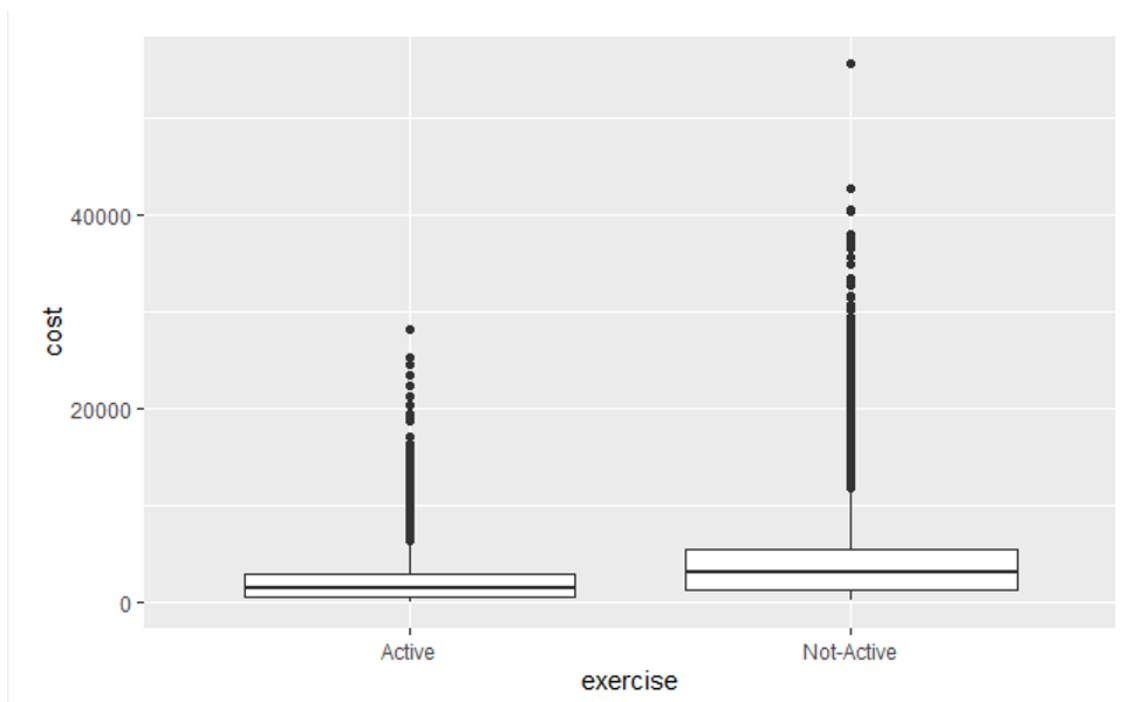
By the end of this step, the data now is ready for exploration and analysis.



## 7. Descriptive Statistics and Visualizations

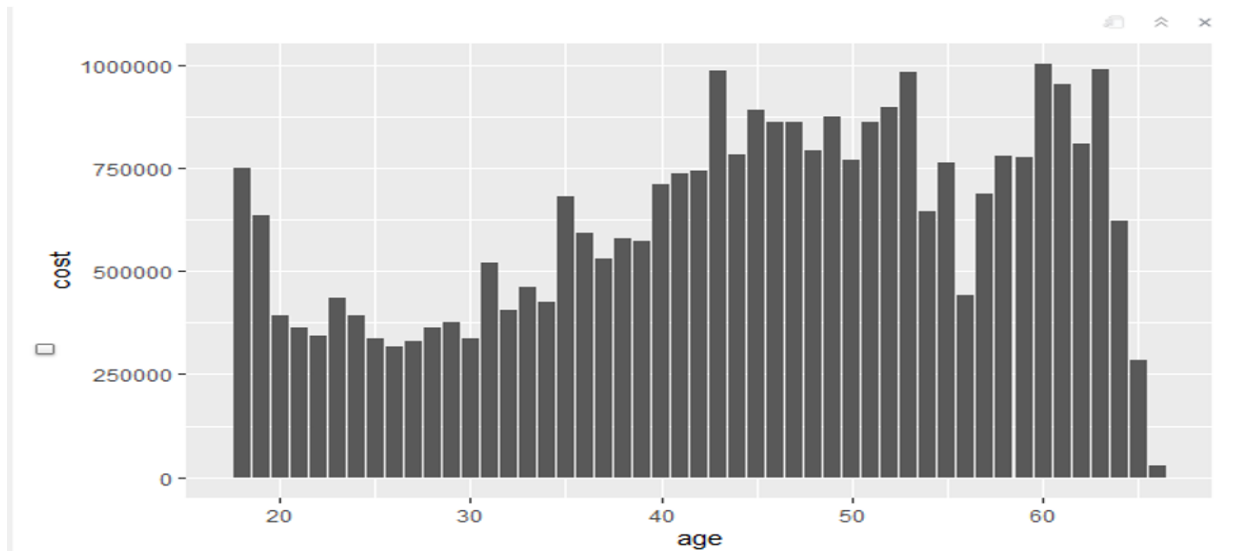
Different types of descriptive analysis were performed to understand the data and gain some insights from it as following:

### 7.1 Exercise VS Cost



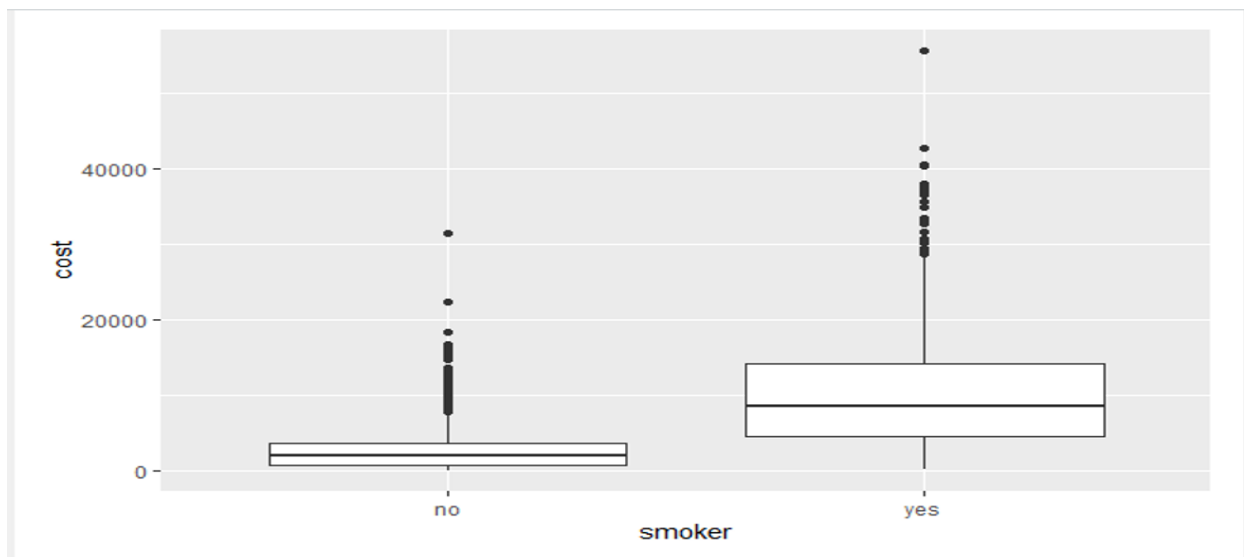
Here we analyzed the overall active and non-active people who exercise and represented it against the cost distribution of healthcare they spend on. We can clearly see that people that are non-active exercisers spend more on their healthcare cost compared to active exercisers. (Higher cost median)

## **7.2 Age VS Cost (Histogram)**



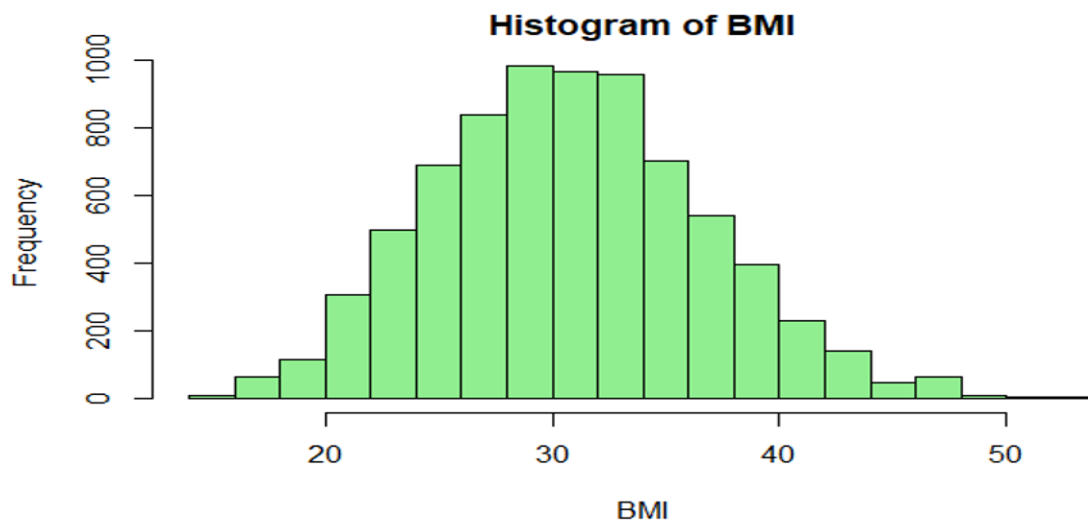
Here we analyzed the age group of people and represented it against the cost of healthcare they spend on. We can clearly see that the cost increases as the age of a person increases. The cost is directly proportional to age

## **7.3 Smoker VS Cost (Box Plot)**



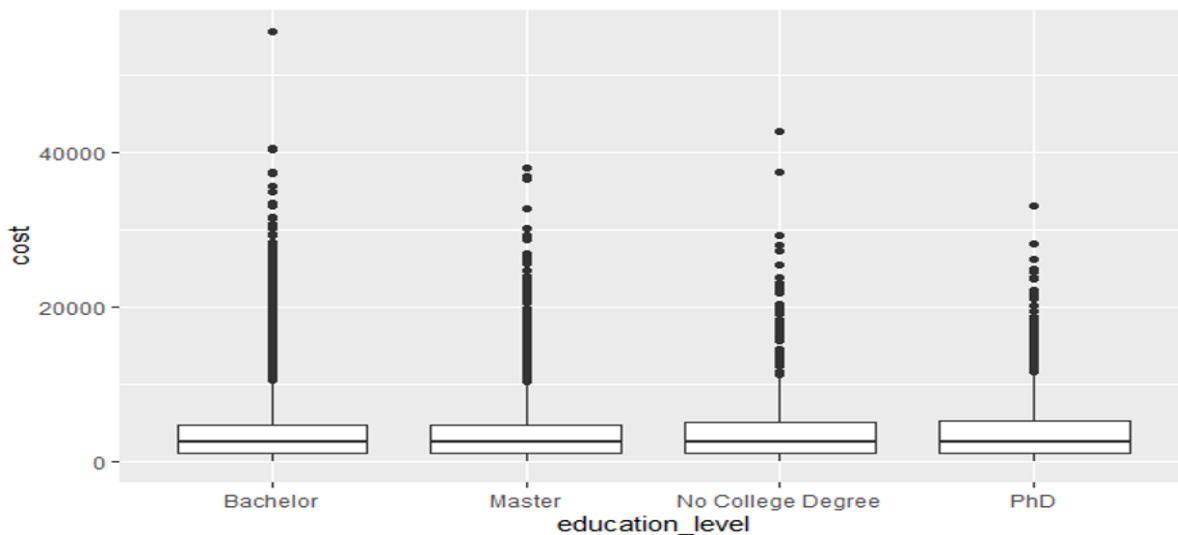
Above, we analyzed the smokers and non-smokers people and represented it against the cost of healthcare they spend on. We can clearly see that people that are smokers spend more on their healthcare cost compared to non-smokers. This suggests that smoking is a key parameter in judging the impact of expenditure on healthcare services.

#### **7.4 BMI (Histogram)**



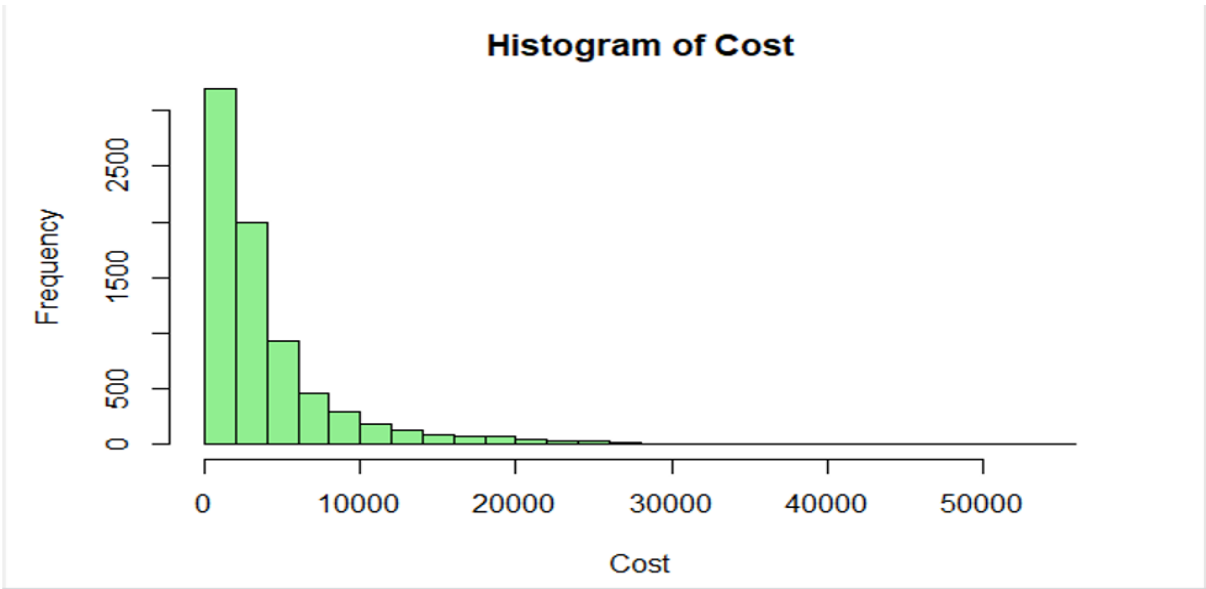
In the above histogram we can see the frequency distribution of BMIs. There are many people having BMI around 25-35 which is above the ideal BMI- below 26. Hence, we get an idea from this plot why people having the above mentioned BMI range spend more on healthcare.

#### **7.5 Education level VS Cost (Box Plot)**



The above graph explains the how much people with different educational backgrounds spend on healthcare services. We cannot interpret much from this factor as they share similar medians.

**7.6 Cost distribution (Histogram)**

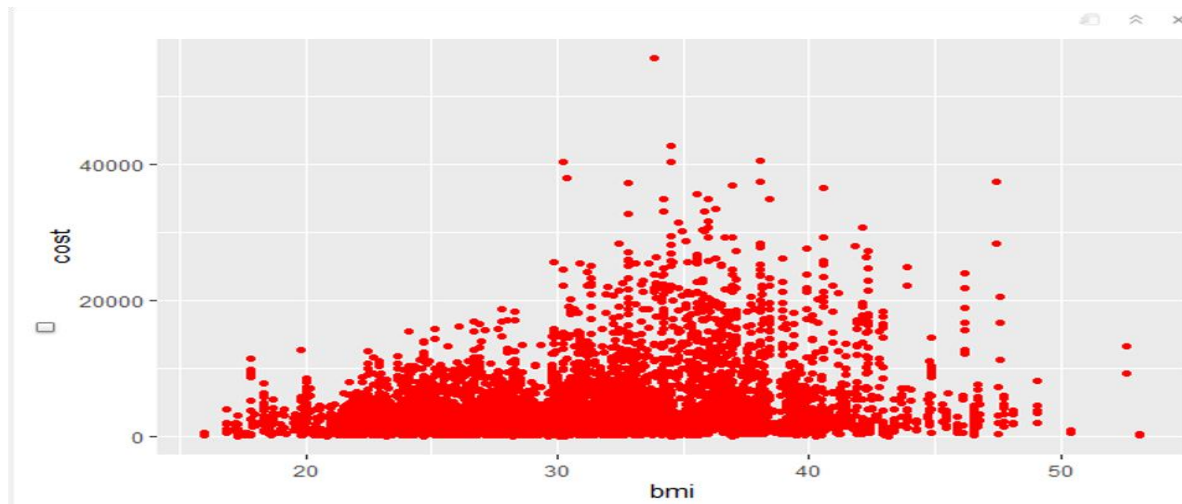


Here we analyzed the cost and represented it against the frequency of people. We can clearly see that 75% of people spend approximately 5000 and lower. (5000 was the threshold to identify expensive patients).

Below is the quantile of the cost distribution.

<pre>{r} quantile(data\$cost)</pre>					
0%	25%	50%	75%	100%	
2	970	2500	4775	55715	

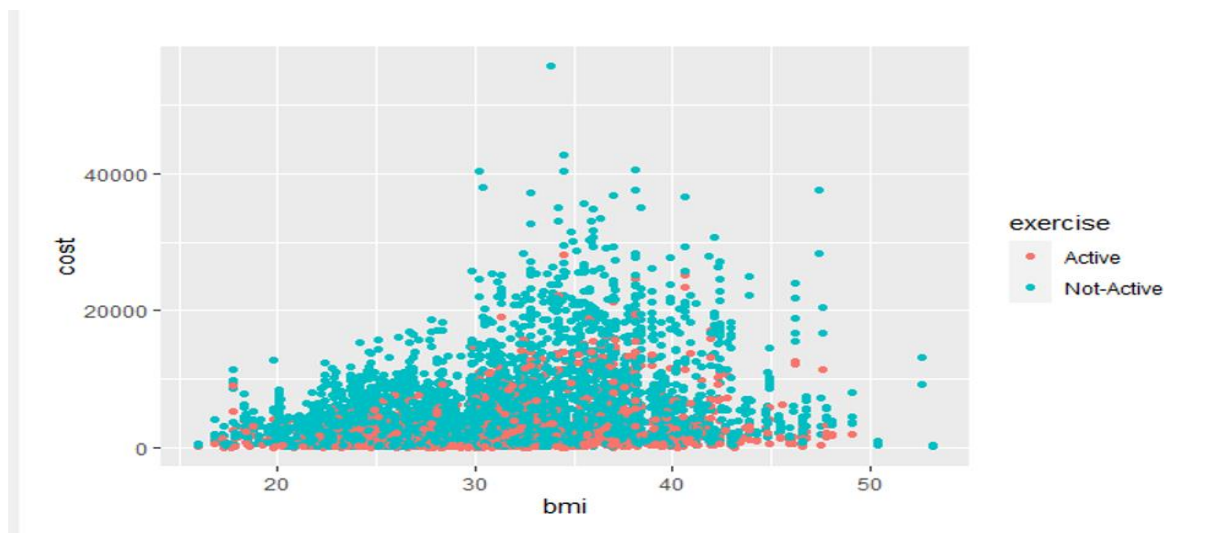
### **7.7 BMI VS COST (Scatter Plot)**



Above, we analyzed the BMI of the people and plotted it against the cost of healthcare they spend on. We can clearly see that people that have BMI above 30.5 spend more on their healthcare cost compared to people who have BMI between 18 to 30.

This suggests that BMI is a key parameter in judging the impact of expenditure on healthcare services.

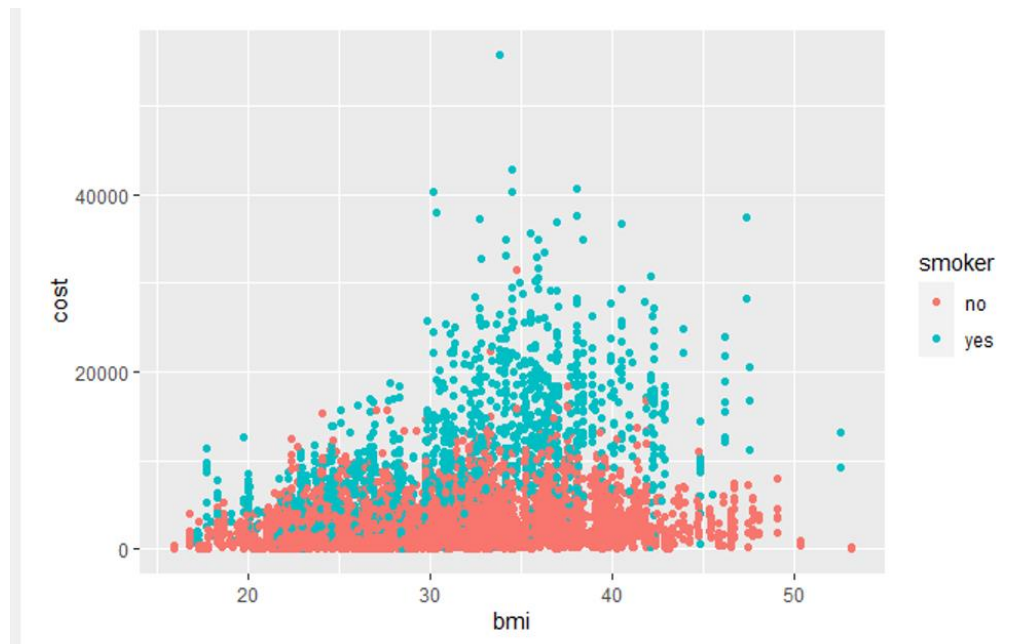
### **7.8 BMI VS COST (Scatter Plot) – Active/Not Active**



In this plot, we analyzed two parameters the BMI and exercise and represented it against the cost of healthcare they spend on. We can clearly see that people that are non-active exercisers and

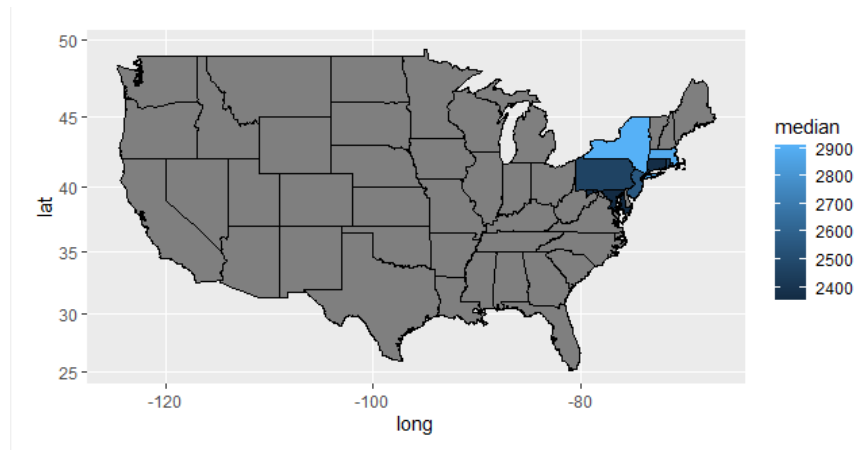
have BMI between 30 to 40 spend more on their healthcare cost.

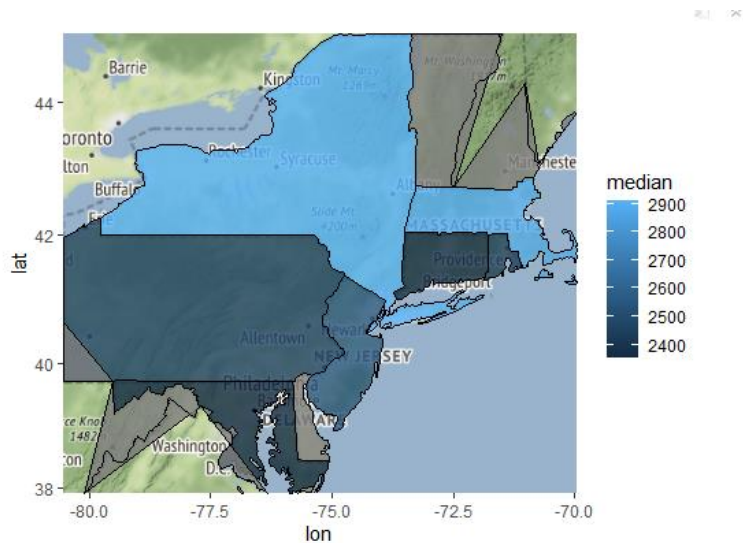
### **7.9 BMI VS COST (Scatter Plot) – Smoker/Not Smoker**



Here we analyzed two parameters the BMI and smoker and represented it against the cost of healthcare they spend on. We can clearly see that people that are smokers and have BMI between 30 to 40 spend more on their healthcare cost.

### **7.10 Location VS COST (Map Plot)**





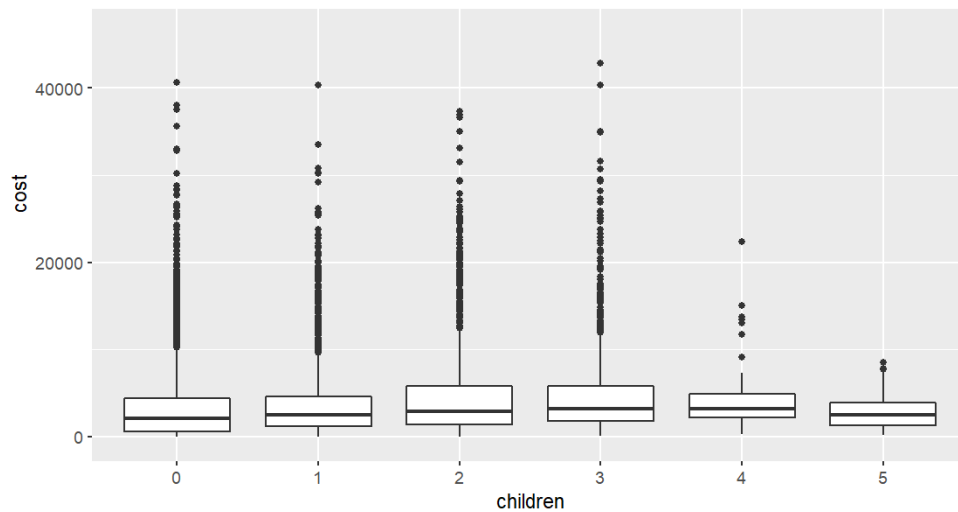
The maps above show that New York has higher median cost compared to other states under study

A tibble: 7 × 3

location <chr>	sum <int>	median <dbl>
CONNECTICUT	2350834	2362.0
MARYLAND	2826778	2352.0
MASSACHUSETTS	1984406	2887.0
NEW JERSEY	1957421	2552.5
NEW YORK	2549844	2910.0
PENNSYLVANIA	16132692	2462.0
RHODE ISLAND	2851757	2448.5

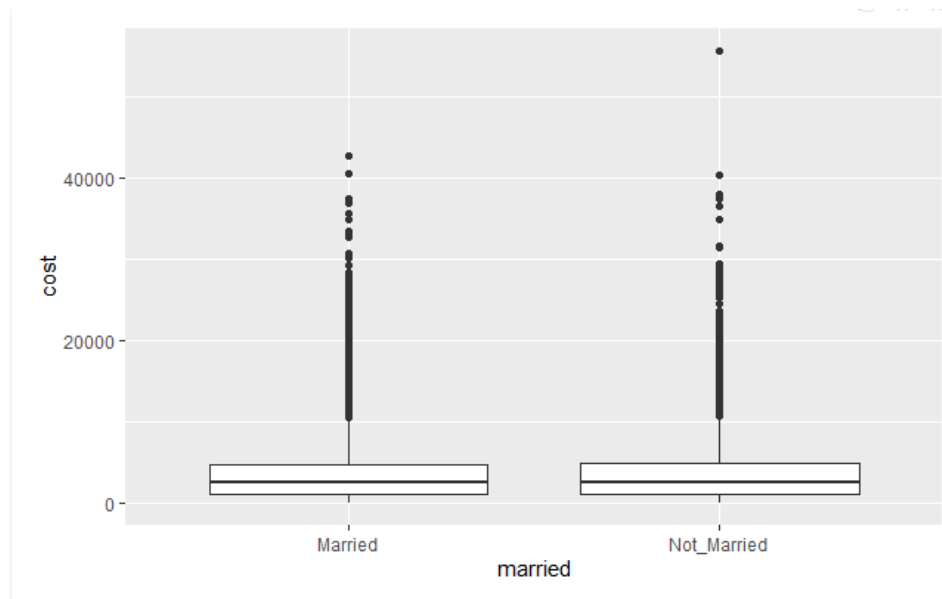
New York and Massachusetts spend quite more than other mentioned states.

### **7.11 Number of children VS COST (Box Plot)**



There is slight correlation between the number of children the person has and the cost. Those with 2 or 3 and 4 children seem to have higher median.

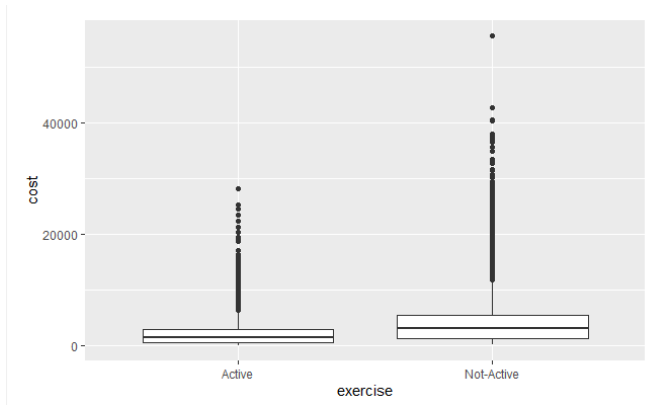
### **7.12 Social Status VS COST (Box Plot)**



There is no change in distribution based on the social status.

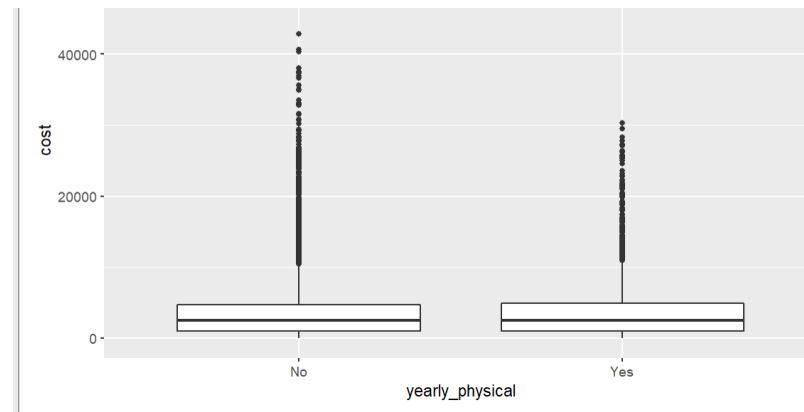


### **7.13 Exercise VS COST (Box Plot)**



The exercise status of the person seems to have some impact on the cost level as it can be seen from the Boxplot above. Not-Active people have higher cost median in comparison with Active people.

### **7.14 Yearly Physical VS COST (Box Plot)**



There is no change in distribution based on the yearly physical visits.

### **Correlation between Cost, BMI and Age.**

After running the correlation function on the above variables, it was found the correlation coefficient is 0.32 between the cost and age. And 0.25 between Cost and BMI which indicates that there is a relationship between cost& age and cost & BMI. Although it is relatively weak.

Table below shows the output result.

```
163 {r}
164 numerical_data <- data[,c('age', 'bmi', 'cost')]
165 res <- cor(numerical_data)
166 round(res, 2)
167
```

	age	bmi	cost
age	1.00	0.09	0.32
bmi	0.09	1.00	0.25
cost	0.32	0.25	1.00

## Conclusion:

From the above executed analysis, we decided to exclude some of the features with small impact on the cost and only include the strong attributes in building the Machine learning models.

### Features included:

Smoker, Exercise, BMI, and Age

## 8. Modelling techniques

Various models have been experienced for predicating expensive attribute based on the information obtained from the dataset. These models give an understandable representation of real-life information from the datasets.

The first step taken was to select the most related features, based on the previous analysis. The features selected were: Smoker, Exercise, BMI, and Age

Secondly, the dataset was divided into two parts, training dataset and testing data set. This step is necessarily to prevent the model from over-fitting.

```
{r}
set.seed(111)
trainList <- createDataPartition(y=HMO_data$expensive, p=.70,list=FALSE)
trainSet <- HMO_data[trainList,]
testSet <- HMO_data[-trainList,]
```

Then the following models have been implemented:

### 8.1 Linear Regression Model

The linear regression model was implemented using the continuous variable of the cost. The goal behind this was to examine the relationship between the cost and other features that we have selected.

The result of the model can be seen as below:

```

>>>
lm(formula = cost ~ ., data = HMO_data)

Residuals:
    Min       1Q   Median       3Q      Max
-12321  -1514   -376    989   41978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8778.260    224.018   -39.19  <2e-16 ***
bmi             181.523     6.261    28.99  <2e-16 ***
age             103.574     2.634    39.33  <2e-16 ***
smokeryes      7690.012    93.830    81.96  <2e-16 ***
exercisNot-Active 2268.430    85.981    26.38  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3237 on 7577 degrees of freedom
Multiple R-squared:  0.569,    Adjusted R-squared:  0.5687
F-statistic: 2500 on 4 and 7577 DF,  p-value: < 2.2e-16

```

From the result above of the P value, it can be seen that there is a significant relationship between cost and the parameters selected. However, the Adjusted R-squared is only 0.5687

## 8.2 Support Vector Machine (SVM)

Secondly, we used SVM modeling algorithm to predict the health care expensive attribute of the patient by using the significant features mentioned above.

Similar to what we did in linear regression model, we divided the dataset into training and testing so that we can check and validate our results.

```

57
58 ▾ ``{r}
59 set.seed(111)
60 trainList <- createDataPartition(y=HMO_data$expensive, p=.70,list=FALSE)
61 trainSet <- HMO_data[trainList,]
62 testSet <- HMO_data[-trainList,]
63 ▴ ``
64
65 ▾ ``{r}
66 model <- ksvm(data= trainSet, expensive~., C=5, CV =3, prob.model =TRUE)
67 ▴ ``
68

```

\*Set.seed will help to get same results whenever you run the SVM.

## Output:

```
model
\\
Support Vector Machine object of class "ksvm"

SV type: C-svc (classification)
parameter : cost C = 5

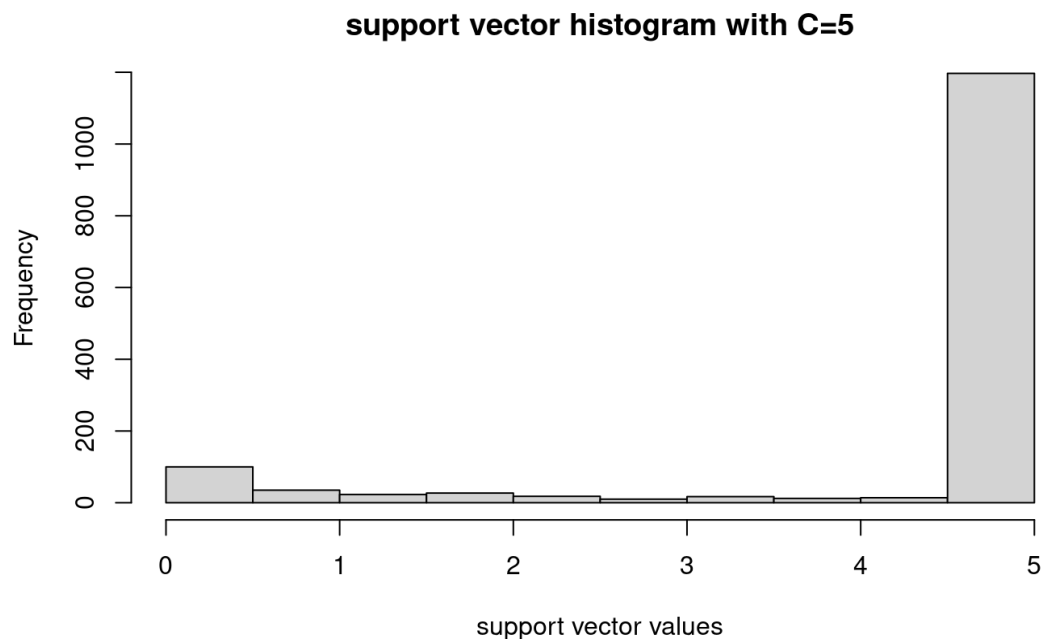
Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 0.515084267225532

Number of Support Vectors : 1453

Objective Function Value : -6239.629
Training error : 0.112472
Probability model included.
```

## Supporting Histogram:

```
{r}
hist(alpha(model)[[1]],main='support vector histogram with C=5',xlab='support vector values')
\\
```



## Prediction and Accuracy:

After training the model with the training dataset we run the predict function to evaluate our model against the testing data as below:

```
{r}
svmPred <- predict(model,newdata = testSet)
svmPred
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
[14] FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE
[27] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[40] FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE TRUE
[53] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  TRUE
[66] TRUE  FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[79] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
[92] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
[105] TRUE  FALSE FALSE FALSE FALSE FALSE FALSE TRUE  TRUE  FALSE FALSE FALSE FALSE
[118] FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE
[131] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE TRUE
[144] FALSE FALSE FALSE TRUE  FALSE TRUE  FALSE FALSE TRUE  FALSE FALSE TRUE  FALSE
[157] FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE TRUE  TRUE  FALSE FALSE
[170] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

The confusion matrix of the SVM model can be found as below:

```
{r}
confMatrix <- table(svmPred, testSet$expensive)
confMatrix
```

```
svmPred      0      1
0 1696  219
1   38  321
```

The SVM model produces 0.88 accuracy, 0.97 Sensitivity and 0.59 Specify. With P-Value smaller than 0.005 which makes the result statistically significant. It can also be observed that the model accuracy is higher than no information rate 0.7625.

```
{r}
confusionMatrix(svmPred,testSet$expensive )
```

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
 FALSE 1696  219
  TRUE    38  321

      Accuracy : 0.887
      95% CI   : (0.8732, 0.8997)
 No Information Rate : 0.7625
 P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.6472

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

      Sensitivity : 0.9781
      Specificity : 0.5944
   Pos Pred Value : 0.8856
   Neg Pred Value : 0.8942
    Prevalence    : 0.7625
  Detection Rate  : 0.7458
 Detection Prevalence : 0.8421
```

### 8.3 Decision Tree Model

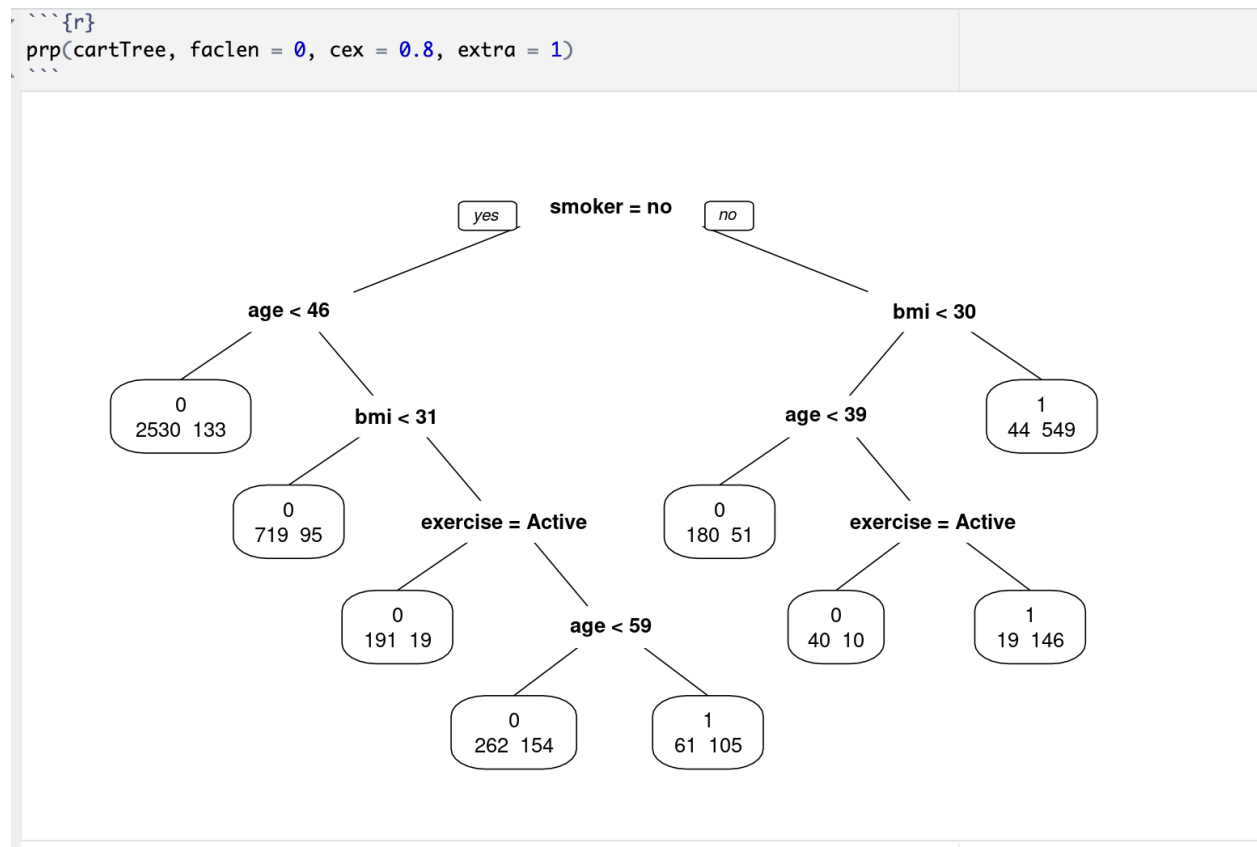
Thirdly, we used decision tree algorithm to predict the health care expensive attribute of the patient by using the significant features mentioned before.

Firstly, data was divided into training and testing sets.

Then, in order to grow our decision tree, we first loaded the rpart package. After that the rpart() function was used, specifying the model formula, data, and method parameters. In this case, we want to predict the expensive patient based on the features of BMI, smoker, age and exercise. so our call to rpart() was as following:

```
cartTree <- rpart(expensive~., data = trainSet, control = c(maxdepth = 5, cp=0.002))
```

Below is the output of the model:



## Prediction and Accuracy:

To check the prediction accuracy of the model we ran the below predict function on the test set that we created earlier.

```
predictValues <- predict(cartTree, newdata=testSet, type = "class")
```

Then we run the confusion Matrix function and the result was as below:

```
confusionMatrix(predictValues, testSet$expensive)
...

Confusion Matrix and Statistics

      Reference
Prediction 0      1
0      1686    209
1       48    331

      Accuracy : 0.887
      95% CI : (0.8732, 0.8997)
    No Information Rate : 0.7625
    P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.6522

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

      Sensitivity : 0.9723
      Specificity : 0.6130
    Pos Pred Value : 0.8897
    Neg Pred Value : 0.8734
      Prevalence : 0.7625
    Detection Rate : 0.7414
    Detection Prevalence : 0.8333
    Balanced Accuracy : 0.7926

    'Positive' Class : 0
```

The decision tree model produces 0.88 accuracy, 0.97 Sensitivity and 0.61 Specify. With P-Value smaller than 0.005 which makes the result statistically significant. It can also be observed that the model accuracy is higher than no information rate 0.7625.

Additionally, we tried the MLR package (Machine Learning in R) to tune our decision tree hyperparameters. The package functions work to build the model using different parameters combination and then recommend the best combination. The code and result were as follows:



```
[ ] param_grid_multi <- makeParamSet(
  makeDiscreteParam("maxdepth", values=1:30),
  makeNumericParam("cp", lower = 0.001, upper = 0.01),
  makeDiscreteParam("minsplit", values=1:30)
)
```

```
▶ dt_tuneparam_multi <- tuneParams(learner="classif.rpart",
  task=d.tree.params,
  resampling = resample,
  measures = measure,
  par.set=param_grid_multi,
  control=control_grid,
  show.info = TRUE)
```

```
☐ Streaming output truncated to the last 5000 lines.
[Tune-y] 7751: acc.test.mean=0.8852675; time: 0.0 min

[Tune-x] 7752: maxdepth=12; cp=0.009; minsplit=26

[Tune-y] 7752: acc.test.mean=0.8852675; time: 0.0 min

[Tune-x] 7753: maxdepth=13; cp=0.009; minsplit=26

[Tune-y] 7753: acc.test.mean=0.8852675; time: 0.0 min

[Tune-x] 7754: maxdepth=14; cp=0.009; minsplit=26
```

```
▶ best_parameters_multi = setHyperPars(
  makeLearner("classif.rpart", predict.type = "prob"),
  par.vals = dt_tuneparam_multi$x
)

best_parameters_multi
```

```
☐ Learner classif.rpart from package rpart
Type: classif
Name: Decision Tree; Short name: rpart
Class: classif.rpart
Properties: twoclass,multiclass,missings,numerics,factors,ordered,prob,weights,featimp
Predict-Type: prob
Hyperparameters: xval=0,maxdepth=14,cp=0.005,minsplit=5
```

```
[ ] cartTree <- rpart(expensive~, data = trainSet, control = c(xval=0,maxdepth=14,cp=0.005,minsplit=5))
```

```
[ ] prp(cartTree, facLen = 0, cex = 0.8, extra = 1)
```



The best hyperparameters combination for higher accuracy were

Xval = 0 , maxdepth = 14, cp = 0.005 and minsplit = 5

The output of the confusion matrix was as follow:

```
▶ predictValues <- predict(cartTree, newdata=testSet, type = "class")
confusionMatrix(predictValues, as.factor(testSet$expensive))
```

```
☐ Confusion Matrix and Statistics
```

```

      Reference
Prediction FALSE TRUE
FALSE  1686  209
TRUE    48   331

      Accuracy : 0.887
      95% CI : (0.8732, 0.8997)
No Information Rate : 0.7625
P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.6522

McNemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9723
Specificity : 0.6130
Pos Pred Value : 0.8897
Neg Pred Value : 0.8734
Prevalence : 0.7625
Detection Rate : 0.7414
Detection Prevalence : 0.8333
Balanced Accuracy : 0.7926

```

The new hyperparameters produced same accuracy/sensitivity and specificity. Thus we kept our initial parameters as they are.

From the results of SVM and the decision tree model above. Both models show similar results in term of Accuracy and Sensitivity. However, we decided to implement the decision tree model as it has higher Specificity result.

#### **8.4 Validation**

From the above results we ensured that our models and results are valid by doing the following:

- 1- Dividing the dataset into training and testing to avoid over-fitting
- 2- Ensure the P-value is below 0.005. Which indicates that the results are statistically significant
- 3- Calculate the error rate and accuracy and build the confusion matrix.

## 9. Interpretations and Actionable Insights

We have drawn below insights based on the descriptive analysis and predictive modeling exercise that we have done previously:

HMO can save cost by taking preventive approach as following:

1. Smoking is considered the main factor to drive the cost, Thus, HMO can initiate health campaigns to encourage its customers to cease smoking. This result can be observed clearly from the output of the decision tree as the root node was the feature of being smoker or not.
2. HMO also can develop wellness programs to encourage its customer to adopt healthy lifestyles. (Reduce BMI, and become more physically active). The wellness program can include: premium discounts, cash rewards, gym memberships. From the output of the decision tree we can see that BMI threshold is 30. Members who have higher BMI (above 30) does have high chance of expensive health care.
3. Higher Insurance plan price for customers aged 60 and above. 60 was drawn from the decision tree output. However, we have some concerns regarding this point as it might be considered as discrimination against old people.
4. In more restrictive scenarios, higher insurance plan can also be imposed on patients who smoke. Also the insurance plan may not cover any respiratory illness that might be linked to smoking.

## 10. Project Link

Shiny App was developed to help HMO easily used our predictive model to predict patients who might be expensive.

The link to the prediction tool can be found below:

[https://sabdelra.shinyapps.io/HMO\\_Tool/](https://sabdelra.shinyapps.io/HMO_Tool/)

The screenshot displays the HMO Prediction Tool interface. On the left, there are input sections for 'HMO inout file' and 'HMO solution file', both with 'Browse...' buttons and 'Upload complete' status. Below these is a 'Number of Rows' input field set to '5'. A note states: 'This tool is used to predict if HMO customer is going to be expensive or not'. The main panel shows 'Descriptive Analytics on the data you upload' with expandable sections for 'Age Distribution', 'BMI Distribution', 'Smoker Percentage', 'Active Percentage', and 'Snapshot of HMO file data'. Below this is the 'Predictive Model results' section, which includes a detailed explanation of the confusion matrix and three large colored boxes showing 'Accuracy:65', 'Sensitivity:75', and 'Specificity:50'. The bottom section, 'Confusion Matrix and Statistics', displays a table and various statistical metrics.

**HMO Prediction Tool**

**HMO inout file**  
Browse... HMO\_TEST\_data.csv  
Upload complete

**HMO solution file**  
Browse... HMO\_TEST\_data.csv  
Upload complete

**Number of Rows**  
5

This tool is used to predict if HMO customer is going to be expensive or not

**Descriptive Analytics on the data you upload**

- Age Distribution +
- BMI Distribution +
- Smoker Percentage +
- Active Percentage +
- Snapshot of HMO file data +

**Predictive Model results**

The box below shows the confusion matrix of our predictive model, we are basically interested to look at three main indicators; Accuracy, Sensitivity and specificity. The Accuracy is basically how well the model is in general in predicting the true values. Sensitivity on the other hand is how well the model is in predicting the true positive class (non expensive patients) while specificity is how well the model is in predicting the negative class (expensive patients). Based on our model we expect 88% accuracy, 97% sensitivity (97% of the patients who are non expensive will be predicted as non expensive) and 61% specificity (61% of the patients who are expensive will be predicted as expensive)

**Accuracy:65** **Sensitivity:75** **Specificity:50**

**Confusion Matrix and Statistics**

```
[1mindexing] [0m] [34m0.csv] [0m] [=====] [32m11.05MB/s] [0m, e]

Confusion Matrix and Statistics

      Reference
Prediction FALSE TRUE
FALSE      9      4
TRUE       3      4

Accuracy : 0.65
95% CI : (0.4078, 0.8461)
No Information Rate : 0.6
P-Value [Acc > NIR] : 0.4159

Kappa : 0.2553

McNemar's Test P-Value : 1.0000
```