IST 687 – Introduction to Data Science

Project Report on

Analyzing healthcare cost information from HMO (Health Management Organization)

By

Abhijith Pillai (SUID: 422093455)

Alex Bayer (SUID: 229828281)

Bohan Li (SUID: 244617355)

Durgesh Kulkarni (SUID: 432596359)

Table of Contents

1. Introduction

1.1 Description

Health Management Organizations (HMOs) are medical insurance companies that provide health care for a fixed yearly fee. The dataset we were provided comprises 14 columns and information on 7,583 persons. The columns cover a wide range of topics, including the individual's unique identity, age, geographic region, gender, education level, marital status, number of children, and healthcare spending. They also inquire about the individual's physical activity, smoking habits, BMI, yearly physical examination status, and hypertension status.

Our goal is to to provide actionable insight, based on the data available, as well as accurately predict which people (customers) will be expensive. We need to also identify the key drivers for why some people are more expensive.

1.2 Business Questions

We developed and provided the following business questions after evaluating and comprehending the problem statement.

1. What are the major factors that are affecting people's health and making them to spend more money on health?

2. Does region is the factor that is affecting people to be more expensive because of high cost of living?

3. Does size of family matter in health cost expense?

4. People who are more stressed are likely to be more expensive?

5. Overall what can be done to reduce the healthcare expense of a person?

1.3 Technical Details

Packages used in project: -

- tidyverse- The tidyverse is a collection of R packages
- ggplot2- It is used for data visualization
- maps- Provides different map outlines and points

- ggmap- Provides with functions to visualize spatial data and models
- mapproj- Converts latitude/longitude into projected coordinates
- imputeTS- Specializes in time series imputation
- caret- Used for building machine learning models
- kernlab- Used for kernel-based machine learning methods in R
- rpart- Rpart is a powerful machine learning library in R that is used for building classification and regression trees
- arules- The arules package for R provides the infrastructure for representing, manipulating and analyzing transaction data and patterns
- rpart.plot- It is used to create a tree structure of model that is built under rpart

1.4 Objective

We provide consultancy services to HMOs (Health Management Organizations), which are medical insurance organizations that provide healthcare in return for a predetermined yearly fee.

Our goal is to pinpoint the primary reasons why certain people require more medical treatment than others, identify those who will incur significant healthcare expenditures in the following year, and provide the HMO with tailored guidance on how to reduce costs in order to reduce their total healthcare spending.

2. Data Analysis

2.1Analyzing columns of dataset

The dataset contains following attributes

• **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.

2.2 Data acquisition and cleaning

HMO Dataset was provided to perform analysis through a csv file that was stored in AWS S3. After Analysis on each attribute of dataset we found that there were 68 missing values for BMI and 80 missing values for hypertension.

```
#str(dfhco)
sum(is.na(dfhco$X))
sum(is.na(dfhco$age))
sum(is.na(dfhco$bmi))#78 missing values
sum(is.na(dfhco$children))
sum(is.na(dfhco$hypertension))#80 missing values
sum(is.na(dfhco$cost))
```

```
[1] 0
[1] 0
[1] 78
[1] 0
[1] 80
[1] 0
```

We have used interpolation technique to replace missing data. This technique uses either linear, spline or stineman interpolation to replace missing values. Now after interpolation we see that there are 7582 observations of 14 variables. For the new data we had acquired summary as follows.

```
library(imputeTS)
dfhco$bmi<-na_interpolation(dfhco$bmi)
dfhco$hypertension<-na_interpolation(dfhco$hypertension)
summary(dfhco)
```

```
       X                  age              bmi           children       smoker            location
 Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000   Length:7582        Length:7582
 1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000   Class :character   Class :character
 Median :    24916   Median :39.00   Median :30.50   Median :1.000   Mode  :character   Mode  :character
 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_type      education_level    yearly_physical      exercise           married           hypertension
 Length:7582        Length:7582        Length:7582        Length:7582        Length:7582        Min.   :0.0000
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:0.0000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.0000
                                                                                                Mean   :0.2005
                                                                                                3rd Qu.:0.0000
                                                                                                Max.   :1.0000
    gender              cost
 Length:7582        Min.   :    2
 Class :character   1st Qu.:  970
 Mode  :character   Median : 2500
                    Mean   : 4043
                    3rd Qu.: 4775
                    Max.   :55715
```

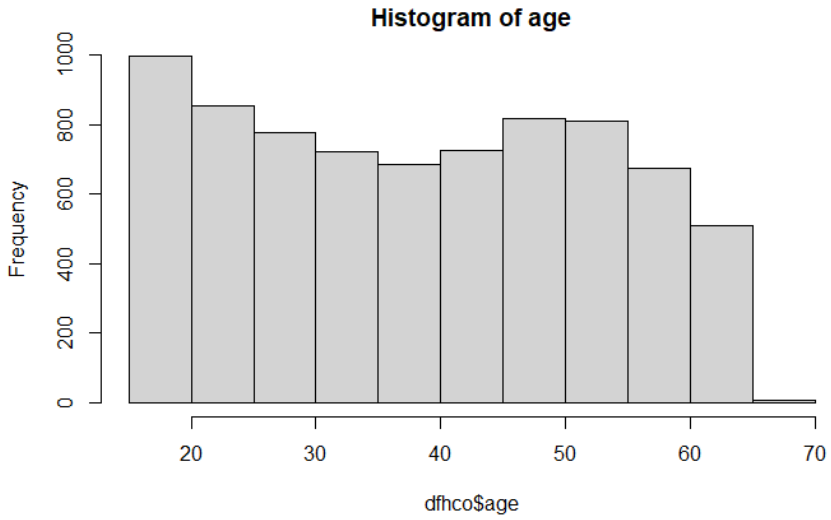
3. Exploratory Analysis

In this step the distribution and/or classification of each property are initially understood by analyzing each one separately. The aim is to check whether there is any association by trying a number of combinations between the characteristics.

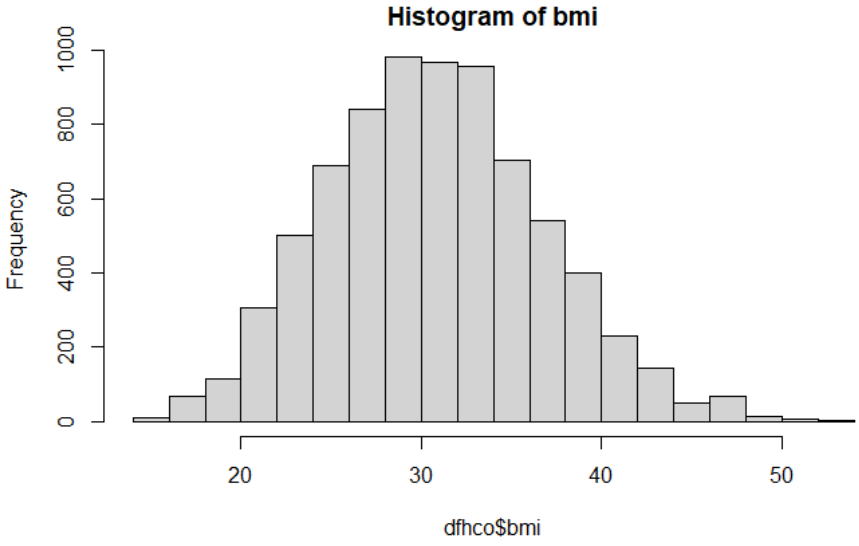
3.1 Histograms

For each numeric attribute which we found significant for that we have plotted the histogram to study its nature.

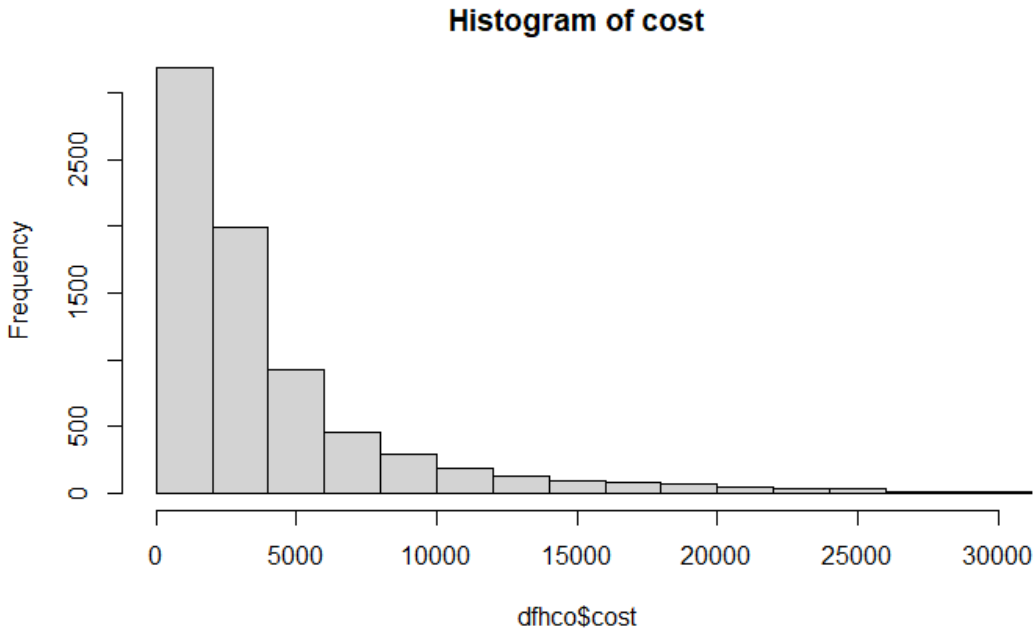
Age: We can observe a bimodal distribution where the first peak is at the age of 18-20 and another similar peaks are at the 45-55.



BMI:- This attribute has a normal distribution, as shown in the image below, with the values from 15.96 to 53.13 and a mean 30.79.



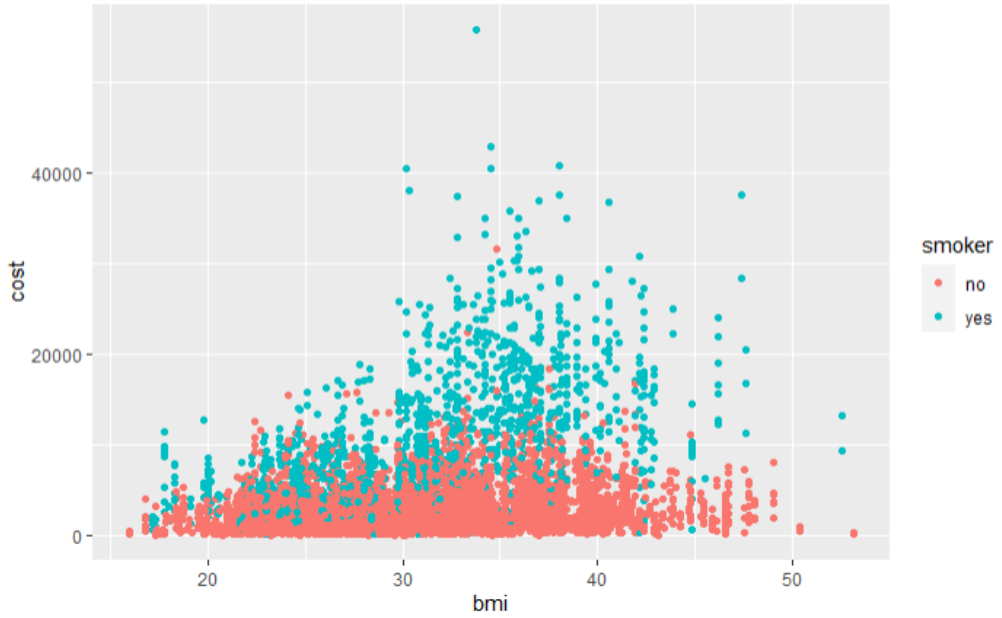
Cost:- The Distribution of cost is right skewed with the bulk of expenses falling below 1000. The prices vary from $2 to $55715, with an average of $4052.



3.2 Scatter Plots

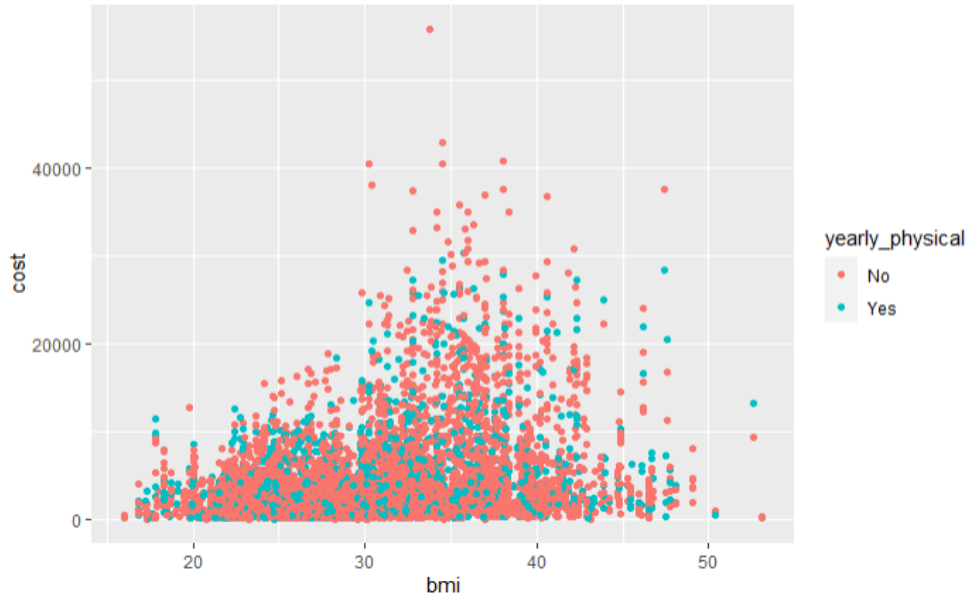
We have used scatterplots to find the correlation of key attributes with respect to cost that helps us to determine the reasons why some people are having higher cost expenses.

We can observe from the scatter plot of bmi, cost, and smoker that nonsmokers have lower costs than smokers. Furthermore, we may notice a densified observation with BMI values greater than 30 (showing obesity) and a higher value of expense and smoker.

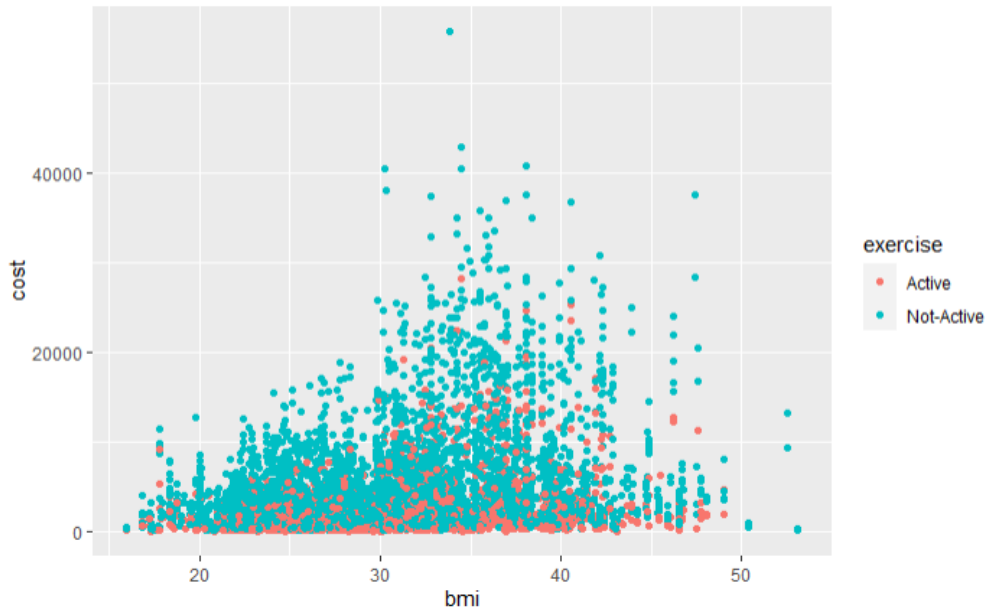


In the scatter plot of bmi,cost and yearly_physical we see a comparison of bmi, cost and whether or not a person receives a yearly physical. Most of the points

are heavily distributed in the same area on the lower end of the cost spectrum indicating that whether or not a person gets a physical annually does not have a significant impact on their health care costs. However, looking at the outliers on the higher end of the cost spectrum we can see that they all did not receive an annual physical indicating that physicals may be an indicator for extreme health care cost.



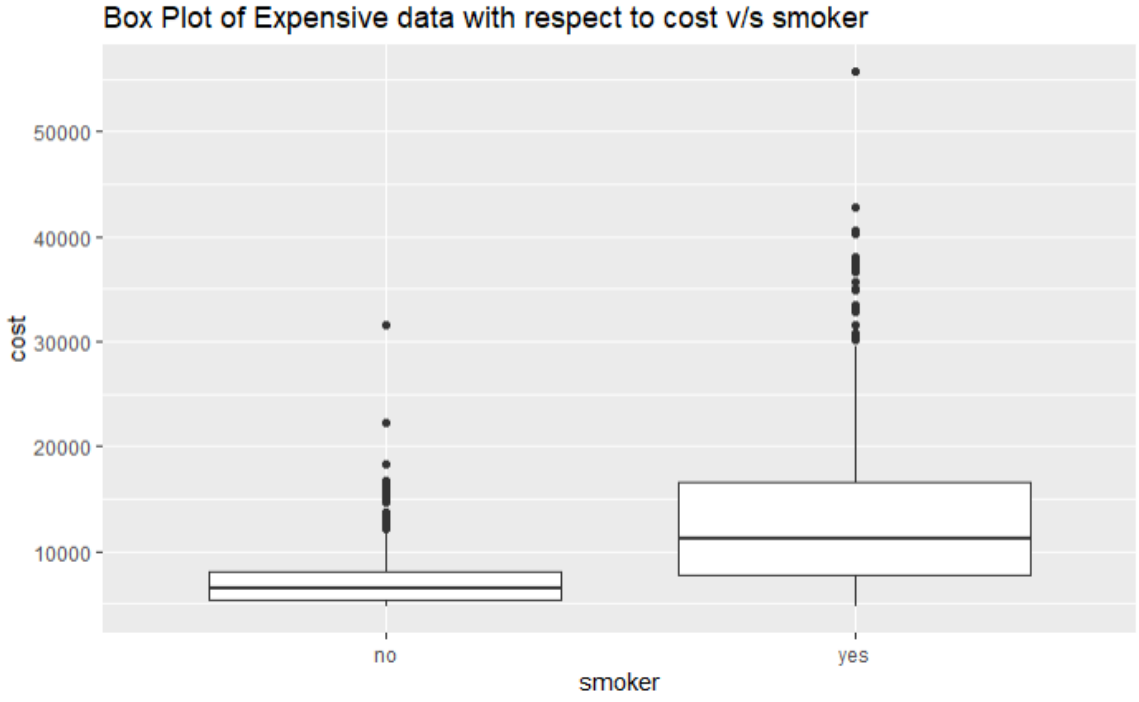
In the scatterplot of bmi, cost and exercise we see the comparison of bmi, cost and activity. Here we see that activity generally does not have a significant effect on bmi with many active and non-active persons across the bmi spectrum. For cost however we see that the non-active population generally has higher costs than the active populations and the cost outliers are all non-active persons.



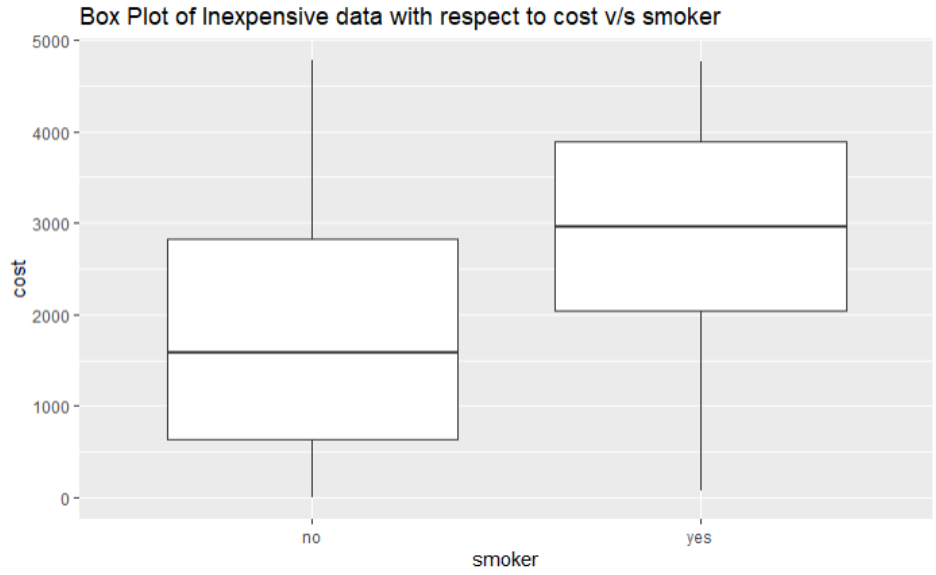
3.3 Box Plots

For boxplots we had filtered dataset into 2 different dataframes such as expensive and inexpensive on basis of cost. If cost is greater than 4775 then that data will go to expensive dataframe else in inexpensive dataframe. We have taken 4775 value on the basis of quantile function. Quantile function gave us the results that 25% people are having cost greater than 4775 in main dataset and we decided to take this value as threshold of expensive status.

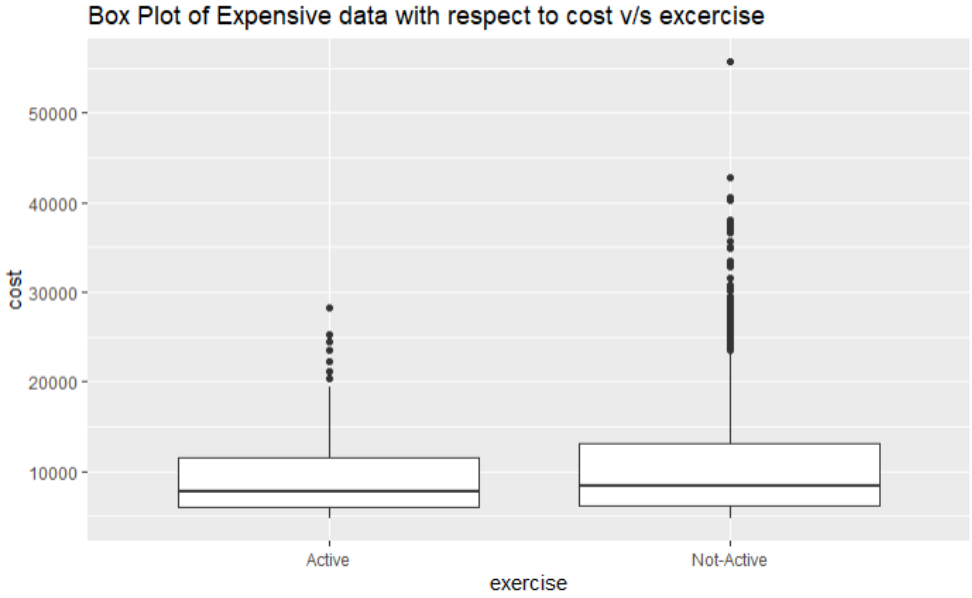
The boxplot of expensive data with respect to cost v/s smoker displays the distribution of smokers in the expensive population compared to their health care costs. We see here the average distribution box of those who are not smokers is significantly lower than the average distribution box of those who are smokers. Looking at the outliers we can see that the majority of non-smoker outlier's cost is in line with the average cost of smokers in this population. Even the most extreme non-smoker outlier has costs below most of the smokers outliers. The outliers for the smokers in this case have costs on the extreme end of the spectrum with many exceeding $30,000



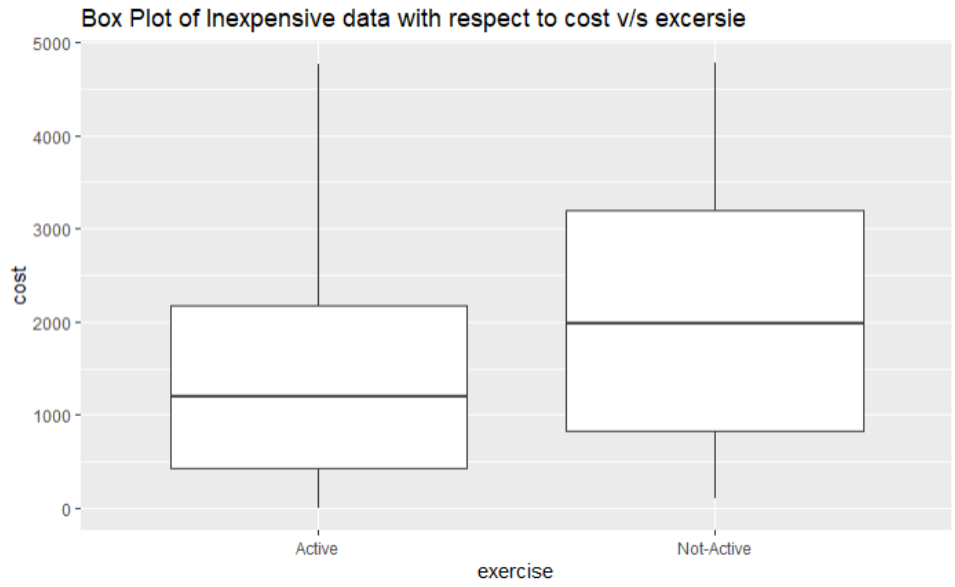
For the inexpensive population we can see that whether or not a person smokes has less of an effect on cost but those who choose not to smoke have a lower average cost. We can also see in this population that even those that do choose to smoke have a much lower average cost than the expensive population.



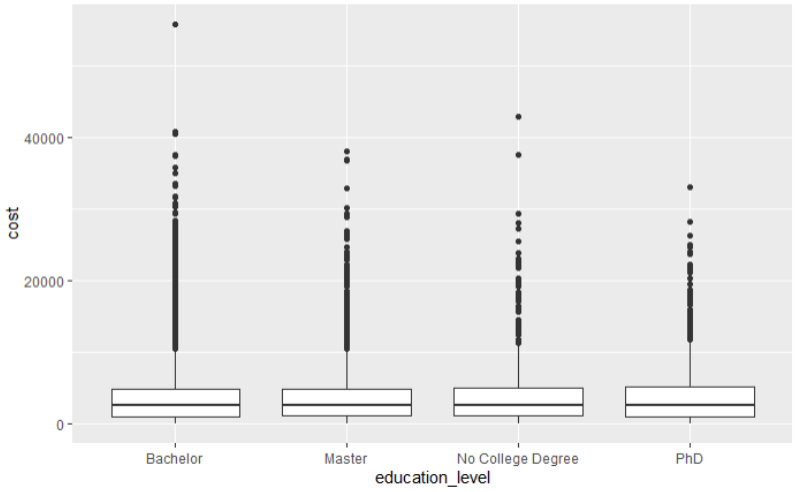
The expensive data with respect to cost v/s excercise boxplot displays the comparison of those who exercise and those who don't in the expensive population against cost. The average distribution shows that exercise has only a minimal effect on cost however the outliers in this example show those with extreme health care costs predominantly don't exercise.

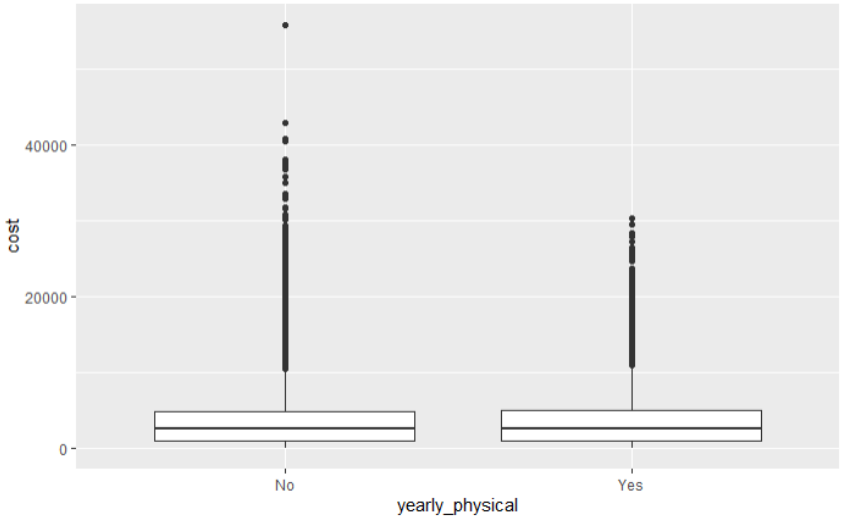


The inexpensive data with respect to cost v/s excercise boxplot displays the comparison of exercise in the inexpensive population. Here we see that those who don't exercise have a higher average cost than those who don't. However, in this population those who chose not to exercise have a lower average cost than those who don't in the expensive population.



The boxplot of education_level vs cost and yearly_physical vs cost shows no relation with respect to cost therefore the boxplots for these attributes are not useful



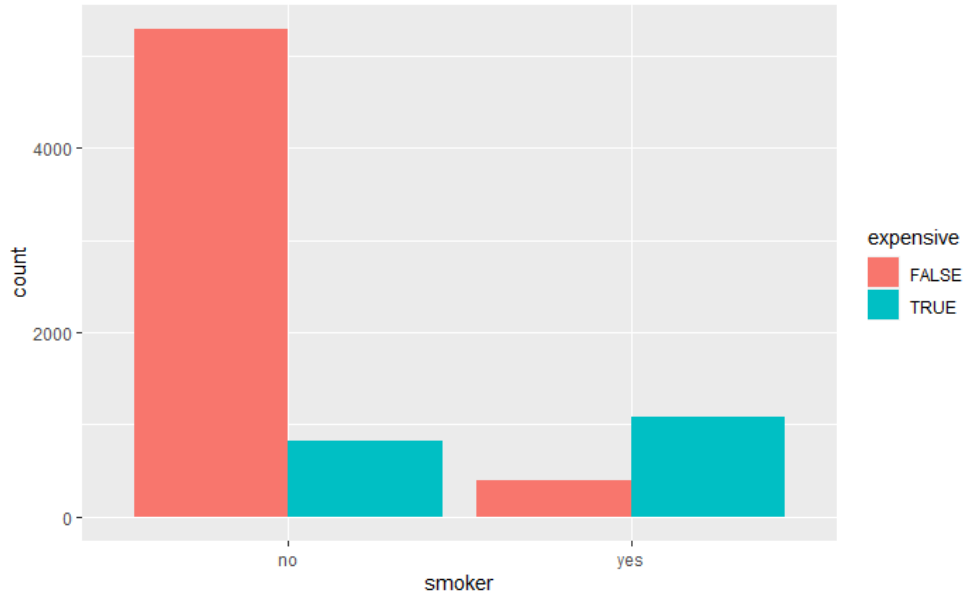


3.3 Bar Plots

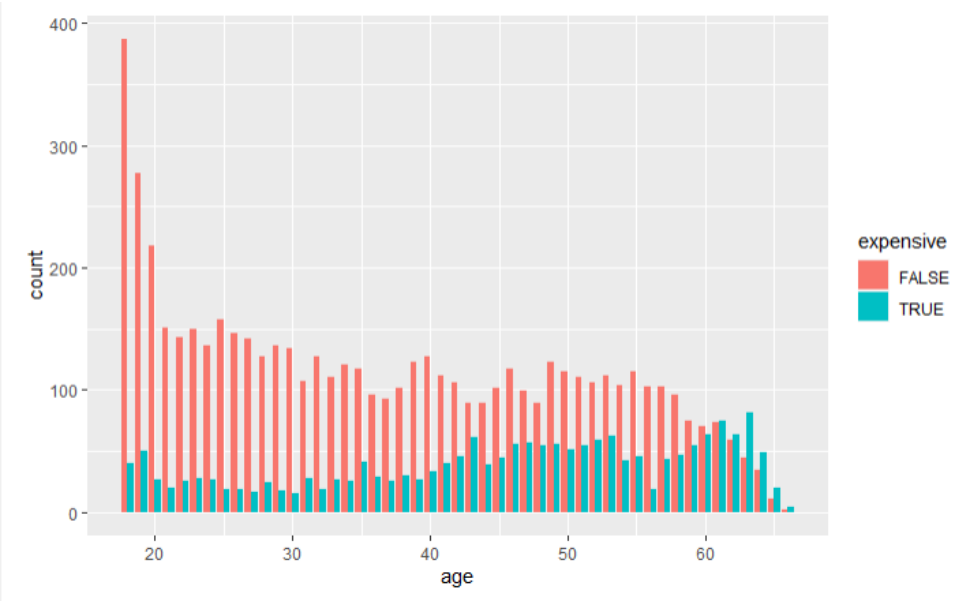
For barplots we have introduced a new column in HMO dataframe i.e expensive which shows the status of person if a person is expensive or not with respect to it's cost. The code for new attribute is as follows.

```
quantile(dfhco$cost)
#Expensive individuals will be those in the top 25% of the population which is cost greater than 4775 according to quantile function
dfhco$expensive <- with(dfhco, ifelse(cost > 4775, 'TRUE', 'FALSE'))
```

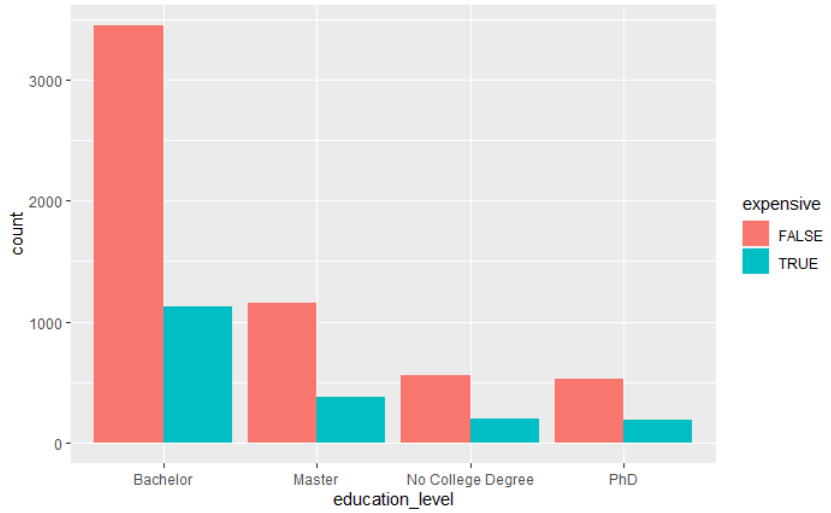
The bar graph of smoker wrt expensive displays whether or not a person is expensive or not based on their smoking status. On the left we see that those who don't smoke are predominantly not expensive. On the right we see the opposite and that a strong majority of those who do smoke are considered expensive.

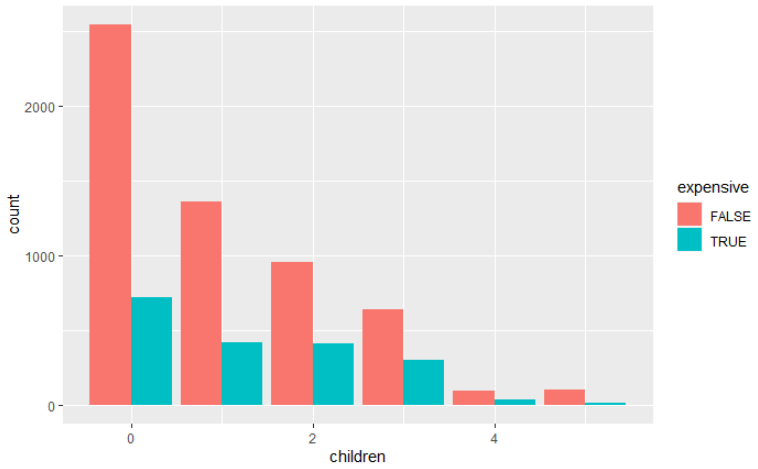


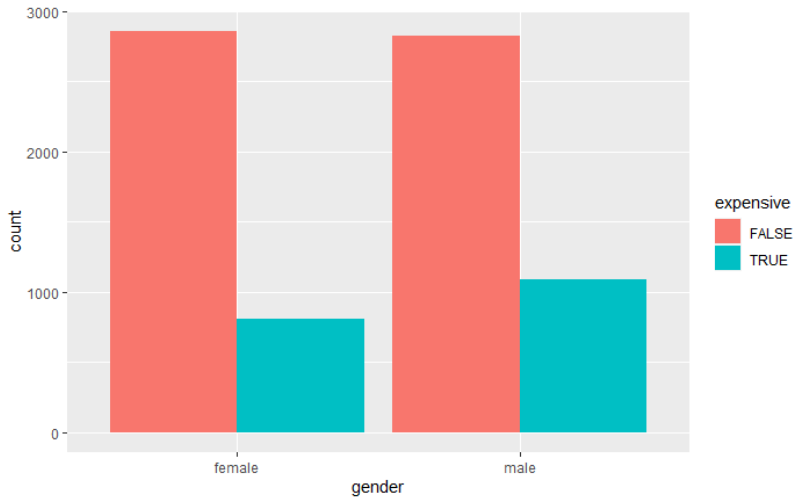
In the bar graph of age wrt expensive we see the distribution on age across the population and how many in each age are expensive or not. At ages below 21 we see predominantly inexpensive people with the distribution generally leveling off afterwards. At ages 45 and above we see the distribution of expensive persons start to rise while inexpensive begin to fall. At ages 60 and above we see the expensive portion of the age ranges begin to become the majority. This indicates that persons above the age of 45 generally begin to experience higher health care costs than other age groups.



The barplots of expensive wrt education_level, expensive wrt children, expensive wrt gender are not useful because they are not highly correlated with cost therefore we cannot derive much information form that barplots.

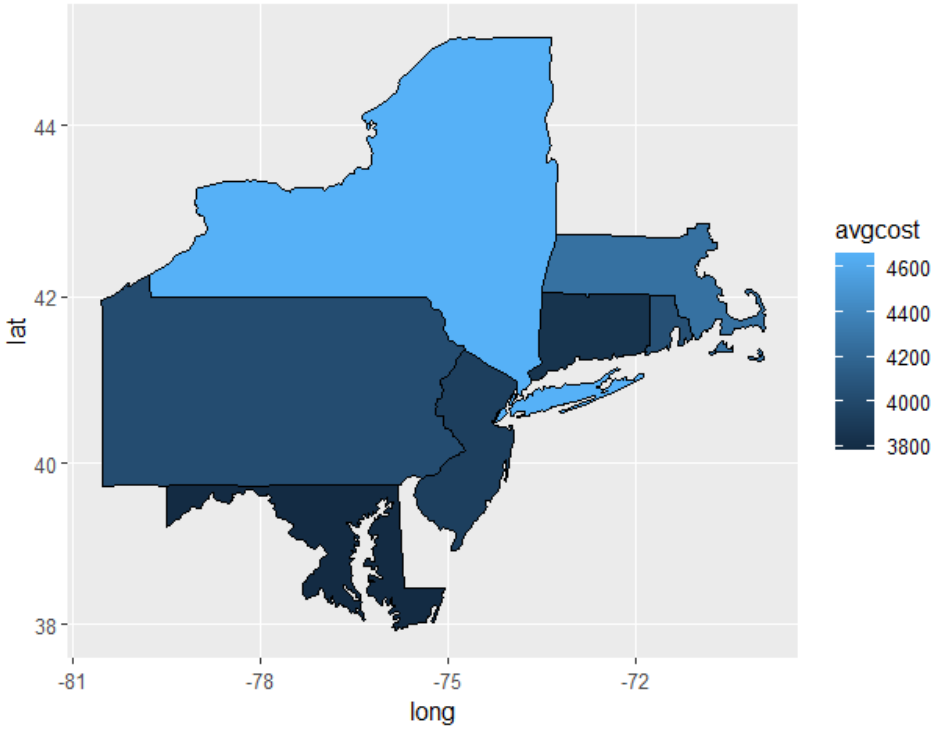






3.3 Map Plots

In the map we see a comparison of the average health care costs by state. Maryland and Connecticut have the lowest average costs while New York has the highest. We can most likely attribute this to several factors such as total population size and the environment of each state. The majority of the population of New York lives in urban environments which is generally less healthy, makes exercise more difficult and has significantly more pollutants compared to rural living.



4. Predictive Models

We may use predictive models to forecast if a person will be pricey next year based on the input data. In the sections that follow, we will look at numerous categorization models and examine their accuracy and other characteristics. All of the models utilized are supervised learning models with classed, discrete outcomes.

We are predicting if a person is expensive or not for that we are going to use the expensive column that we generated for our barplots and based on expensive columns we will train our model while other attributes will act as predictors for our model.

4.1 Support Vector Machines (SVM)

Support vector machines (SVM) are supervised learning models with related learning algorithms that examine data used for classification and regression analysis in machine learning. It is mostly employed in categorization difficulties. Each data item is plotted as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a specific coordinate. The categorization is then carried out by locating the hyper-plane that best distinguishes the two classes. SVMs may do non-linear classification as well as linear classification, implicitly translating their inputs into high-dimensional feature spaces.

We have utilized the svmRadial method, which is a radial basis function, in our R code. We also incorporate some preprocessing by centering and scaling because it is thought to increase model performance. There is no need to separate the training and testing data because they are already separated into distinct datasets. We have created partition on our data where 70% of data will be used for training and 30% will be used for testing The R code is as follows.

```
library(rio)
library(caret)
library(kernlab)
dfhco$expensive<-as.factor(dfhco$expensive)
trainList <- createDataPartition(dfhco$expensive, p=.7, list=FALSE)
trainSet <- dfhco[trainList,]
testSet <- dfhco[-trainList,]
svmmodel<- train(expensive ~ X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical+exercise+married+
hypertension+gender,data=trainSet , method= "svmRadial", trControl=trainControl(method = "none"),
preProcess=c("center","scale"))
```

The confusion matrix that we got while using testing data is as follows

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1640  259
     TRUE     66  309

               Accuracy : 0.8571
                 95% CI : (0.842, 0.8712)
    No Information Rate : 0.7502
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5699

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

            Sensitivity : 0.9613
            Specificity : 0.5440
         Pos Pred Value : 0.8636
         Neg Pred Value : 0.8240
             Prevalence : 0.7502
         Detection Rate : 0.7212
   Detection Prevalence : 0.8351
      Balanced Accuracy : 0.7527

       'Positive' Class : FALSE
```

From SVM we got the accuracy of 85.71%. We got No Information Rate which tells us that largest proportion of the class in our testset we have more false values than true values for expensive status. The p-value is <2.2e-16 which tells us that this model can act significantly for predicting expensive status. Sensitivity is 0.96 or 96% which is calculated as the ratio of number of correct classifications as FALSE vs overall classifications. And the specificity, which is number of correct classifications as TRUE divided by all classifications, is equal to 0.54 or 54%.

4.2 Support Vector Machines (SVM) with K-Fold Cross Validation

The k-fold cross-validation is commonly used to evaluate the effectiveness of SVMs with the selected hyper-parameters. It is known that the SVM k-fold cross-validation is expensive, since it requires training k SVMs. The training dataset is randomly partitioned into k-separate pieces in this scenario, with each chunk having nearly the same balance of outcomes. The approach then trains a model with one of the k-separate pieces removed from the training data and used as the test dataset for the model built with the remaining k-1 chunks. After that, integrate the findings of the k- distinct models into one overall model, and then present the overall model's error rate and accuracy. In our situation, we choose k=10, and the code is as follows:

```
svmmodel.kfold<- train(expensive ~
X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical+exercise+married+hypertension+gender,data=trai
nSet , method= "svmRadial", trControl=trainControl(method = "repeatedcv",number=10), preProcess=c("center","scale"))
```

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1662  275
     TRUE     44  293

               Accuracy : 0.8597
                 95% CI : (0.8448, 0.8737)
    No Information Rate : 0.7502
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.567

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

            Sensitivity : 0.9742
            Specificity : 0.5158
         Pos Pred Value : 0.8580
         Neg Pred Value : 0.8694
             Prevalence : 0.7502
         Detection Rate : 0.7309
   Detection Prevalence : 0.8518
      Balanced Accuracy : 0.7450

       'Positive' Class : FALSE
```

From the confusion matrix we see there is slight increase in accuracy with respect to SVM therefore we can conclude that using k-fold cross validation there is no major change in accuracy and thus this method does not improve our model.

4.3 Regression Trees

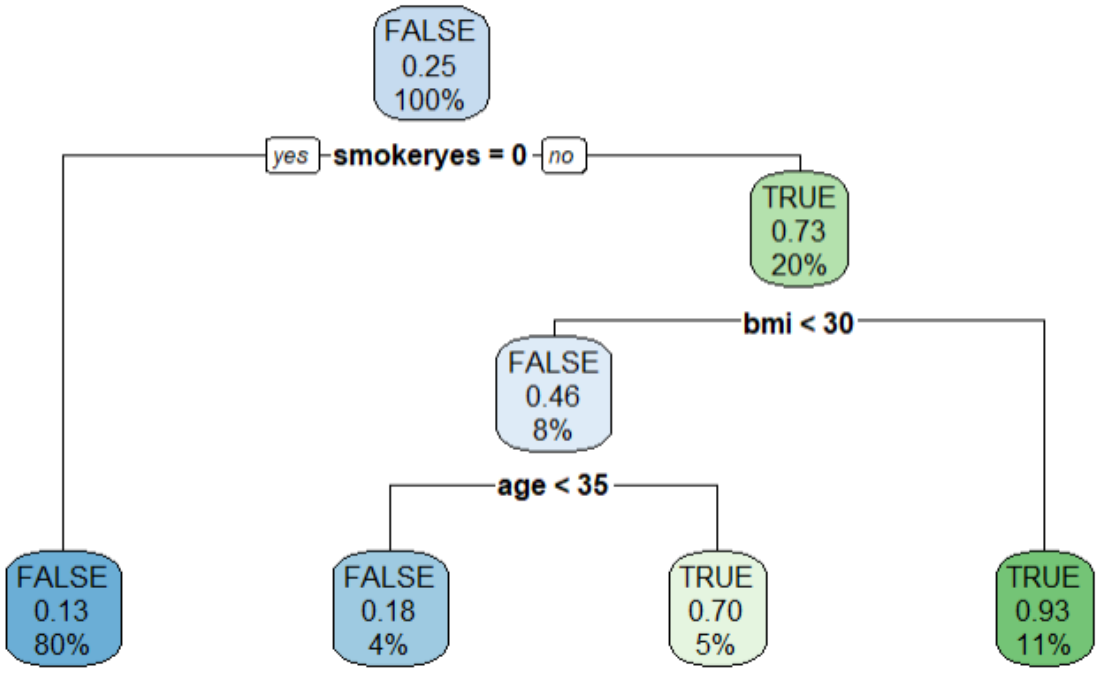
Regression Trees are one of the fundamental machine learning techniques that more complicated methods, like Gradient Boost, are based on. They are useful for times when there isn't an obviously linear relationship between what you want to predict, and the things you are using to make the predictions. In this we are using Recursive Partitioning and Regression Tree algorithm (rpart) model to predict expensive status. The code for rpart is as follows.

```
modelrpart<-train(expensive ~ X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical+exercise+married
+hypertension+gender, method = "rpart",data = trainSet)
```

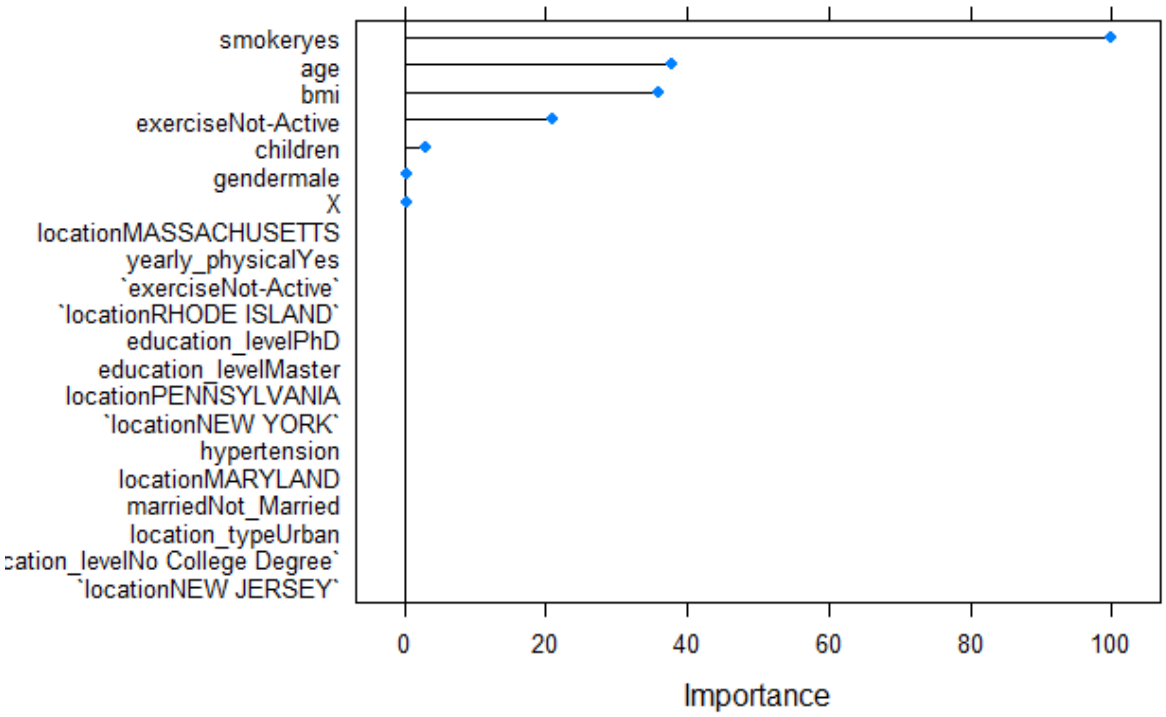
Confusion Matrix for Rpart is as follows

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1660  263
     TRUE     46  305

               Accuracy : 0.8641
                 95% CI : (0.8493, 0.8779)
    No Information Rate : 0.7502
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5845

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

            Sensitivity : 0.9730
            Specificity : 0.5370
         Pos Pred Value : 0.8632
         Neg Pred Value : 0.8689
             Prevalence : 0.7502
         Detection Rate : 0.7300
   Detection Prevalence : 0.8456
      Balanced Accuracy : 0.7550

       'Positive' Class : FALSE
```

Thus from confusion matrix we can observe accuracy is increased to 86.41%, No Information rate has no change, sensitivity increased to 97.30% and specificity increased to 53.70% as compared to all the previous models that we tested. Now let's see the visualization of tree.



From the tree we can see the classification of data. For e.g when a person is smoker then it is mostly classified into expensive category and if not then other attributes like bmi and age is checked and according to that our model classifies the person is expensive or not. From tree we can observe smoker is major factor for classification we will get clear idea after plotting variable importance graph.



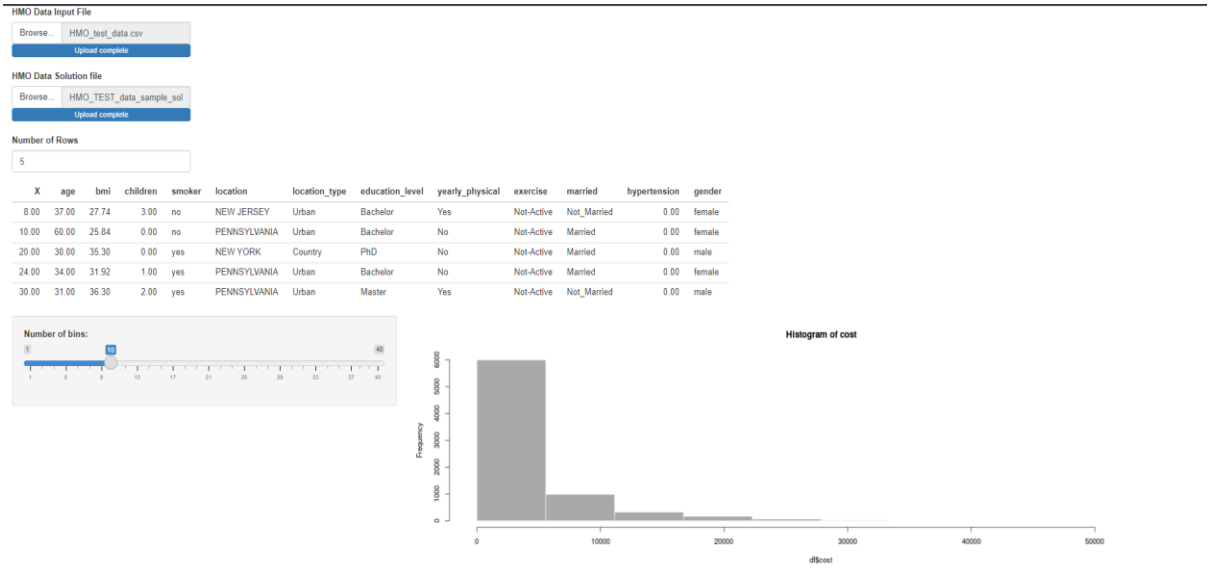
From Importance plot we can clearly observe that smoker is major factor that our model uses for prediction.

5. Shiny App

Shiny is a web application framework for R that enables to build interactive web applications. Shiny apps are useful to communicate information as interactive data explorations instead of static documents. A Shiny app is composed of a user interface ui which controls the layout and appearance of the app, and a server() function which contains the instructions to build the objects displayed in the user interface. Shiny apps permit user interaction by means of a functionality called reactivity.

We have deployed our app on shinyapps.io :- https://607uic-durgesh-kulkarni.shinyapps.io/IDSProject/

Screenshots of our shinyapp




```
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

            Sensitivity : 0.7500
            Specificity : 0.5000
         Pos Pred Value : 0.6923
         Neg Pred Value : 0.5714
             Prevalence : 0.6000
         Detection Rate : 0.4500
   Detection Prevalence : 0.6500
      Balanced Accuracy : 0.6250

       'Positive' Class : FALSE
```

```
# A tibble: 20 × 2
       X expensive
   <dbl> <fct>
 1     8 FALSE
 2    10 FALSE
 3    20 TRUE
 4    24 TRUE
 5    30 TRUE
 6    31 TRUE
 7    35 TRUE
 8    39 TRUE
 9    41 FALSE
10    42 FALSE
11    58 TRUE
12    68 FALSE
13    69 FALSE
14    70 FALSE
15    71 FALSE
16    72 FALSE
17    76 FALSE
18    79 FALSE
19    82 FALSE
20    89 FALSE
```

Waiting for 607uic-durgesh-kulkarni.shinyapps.io...

6. Recommendations

From on our analysis we found that if a person who is smoker can be expensive next year while this is not only just one factor other factors also do matter significantly. Other factors that can be significant factors are age, exercise, yearly physical check-up. For e.g a person who is non-smoker, does exercise actively and also goes for chek-up every year then there is high chance that the person will be inexpensive next year.

Our models indicate a reduction of 10% in BMI, cessation of smoking, 10% of hypertension rate and increase in exercise rate will yield:

$7,000 in average cost savings/person for the expensive population

$13,265,000: potential total cost savings

$3,748,310: Total project cost of wellness program.

$1,758,560 in premium discounts (Based on average ACA premiums, (Forbes, 2022))

$1,989,750: in additional costs $9,516,690 net savings from program 31% total of total HMO cost

According to this analysis we had generated some recommendations for HMO that are as follows.

Create a mandatory wellness program for people that meet certain conditions

- Turn 45 years of age
- Have average annual cost exceeding $5,000

Wellness Program has following features

- Gym membership reimbursement ($500)
- Smart Fitness Watch ($50)
- 150min/week exercise goal requirement.
- Smoking cessation assistance ($500)
- 5% premium discount

- **Reduce BMI:** According to the CDC a reduction of 5-10% of total body weight is likely to improve blood pressure, cholesterol, blood sugars and reduce the risk of chronic disease. (CDC, 2022)
- **Reduce number of smokers:** According to the WHO quitting smoking dramatically decreases risk of heart disease, multiple cancers and increases life span. (WHO, 2020)

- **Health benefits of exercise:** 150min/week of exercise has been shown to risk of injury, heart disease and some cancers and increase quality of life in older adults (Langhammer, 2018)
- **Reduce Hypertension:** Smoking cessation, reduced body weight and increased exercise have been shown to reduce hypertension (Mayo Clinic, 2022)