THE
iSCHOOL
Syracuse University

IST 687: Introduction to Data Science

Final Project Report: Analyzing Healthcare Cost Information from a HMO (Health Management Organization)

Abhijeet Lal | Anish Kumar | Madhumitha Saravanan
Pulak Jain | Samragyee Pandey | Saurav Sundararaju Makam

Submitted to:

Prof. Erik Anderson

TABLE OF CONTENTS

INTRODUCTION

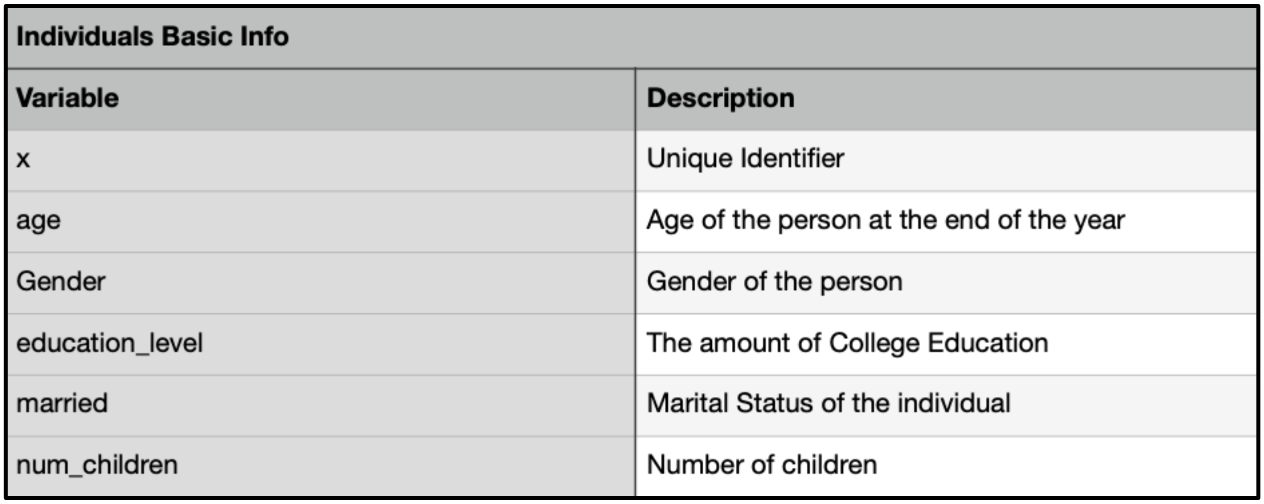
Objective:

Our main objective through this study and analysis is to find out the various factors which trigger healthcare costs for people and provide actionable insight, based on the data available, as well as accurately predict which people (customers) would be expensive. The dataset contains healthcare cost information from an HMO (Health Management Organization). Each row in the dataset represents a person. Our goal is to understand the key drivers for why some people are more expensive (i.e., require more health care), as well as predict which people will be expensive (in terms of health care costs).

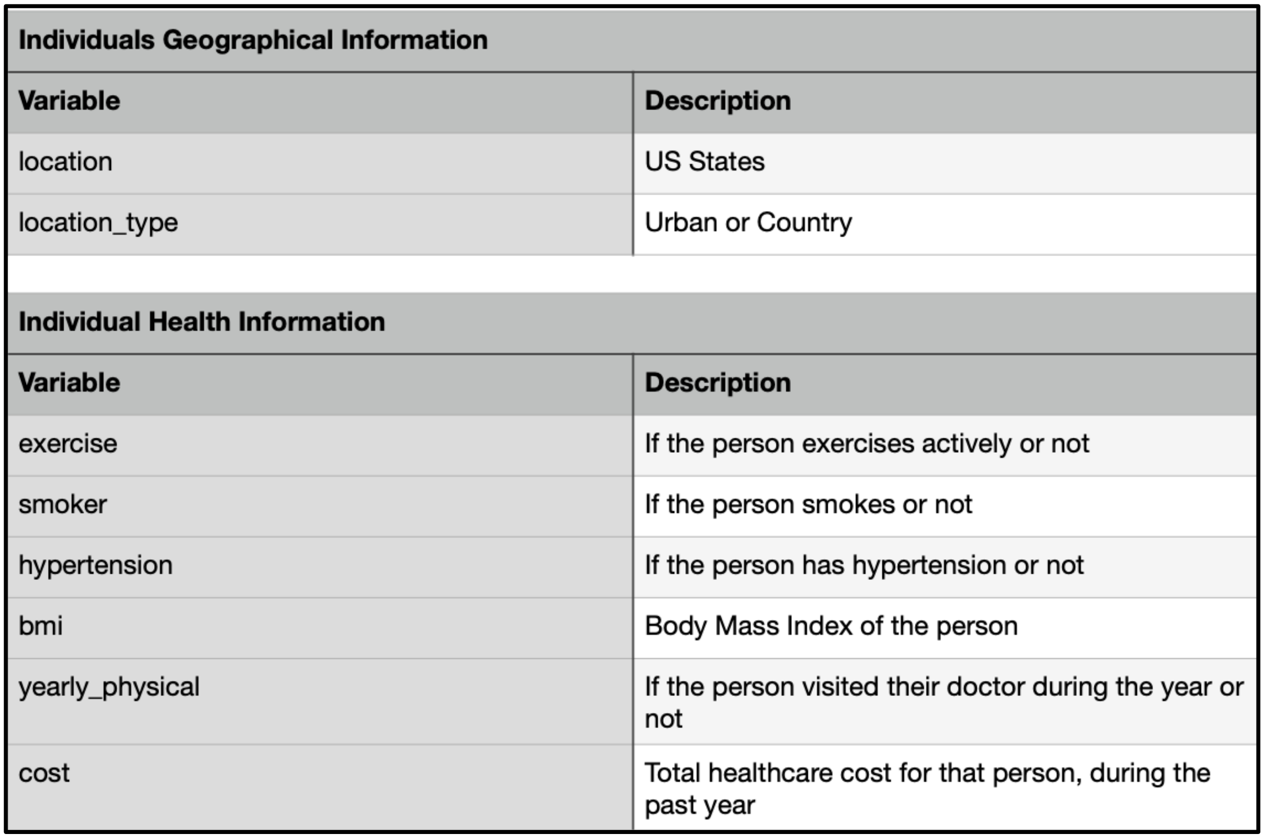
Background:

Health Management Organizations (HMOs) are medical insurance companies that provide health care for a fixed annual fee. The dataset we were given has 14 columns and information on 7,583 people. 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, annual physical examination status, and hypertension status. We give practical information based on the facts at hand through our research, and we also correctly predict which consumers will spend a lot of money on healthcare.

Scope:



Individuals Basic Info	
Variable	**Description**
x	Unique Identifier
age	Age of the person at the end of the year
Gender	Gender of the person
education_level	The amount of College Education
married	Marital Status of the individual
num_children	Number of children



Individuals Geographical Information	
Variable	**Description**
location	US States
location_type	Urban or Country

Individual Health Information	
Variable	**Description**
exercise	If the person exercises actively or not
smoker	If the person smokes or not
hypertension	If the person has hypertension or not
bmi	Body Mass Index of the person
yearly_physical	If the person visited their doctor during the year or not
cost	Total healthcare cost for that person, during the past year

BUSINESS QUESTIONS

Initial Business Questions:

1. Predict who will spend a lot of money on health care in the coming year (i.e., who will have substantial healthcare expenses).
2. Provide actionable insight to the HMO on how to reduce total health care expenses by making specific recommendations on how to reduce the cost of healthcare.

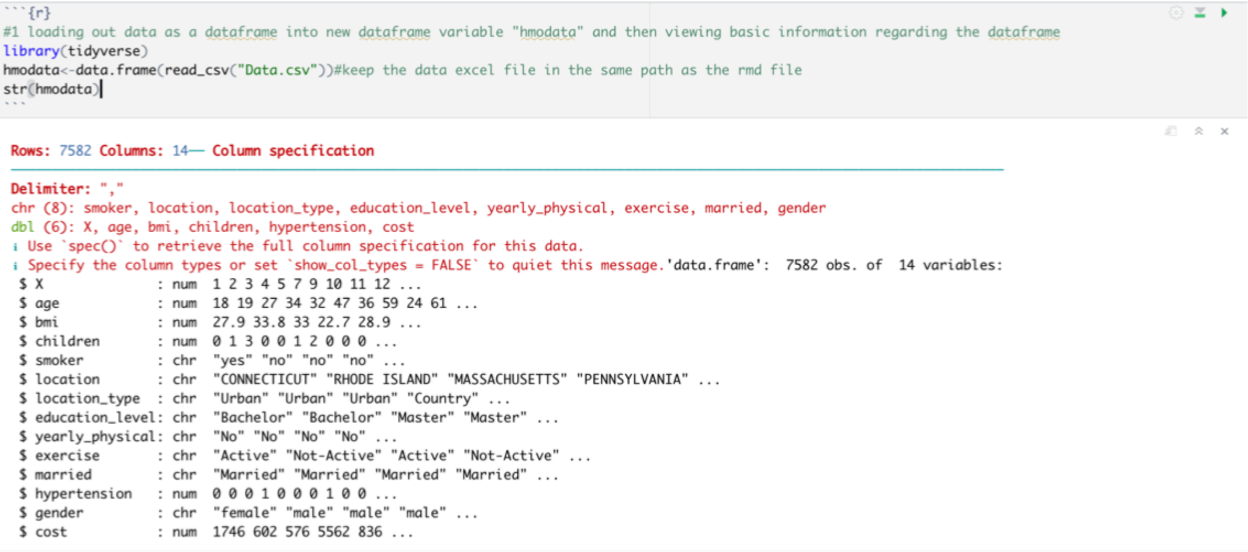
Final Business Questions:

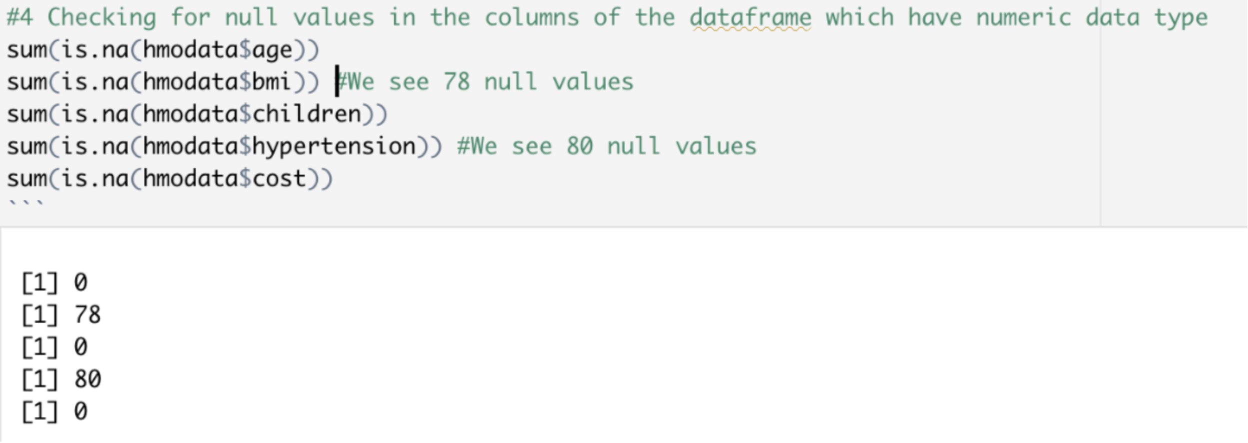
1. What is the general state of health in the United States? (I'm not sure what the perfect measure of health is, but "BMI" can be used.)
2. In which states is the average expenditure larger than, and in which states is it less than, the national average?
3. Which group will be more expensive? People who have hypertension or a BMI that is higher or lower than the average.
4. How much do smokers spend on average?
5. Is there a link between physical activity and personal health? Exercisers are generally less expensive (in terms of healthcare).
6. Are those who get their blood pressure tested once a year less expensive or more expensive than those who don't?

DATA ANALYSIS

Data Acquisition:

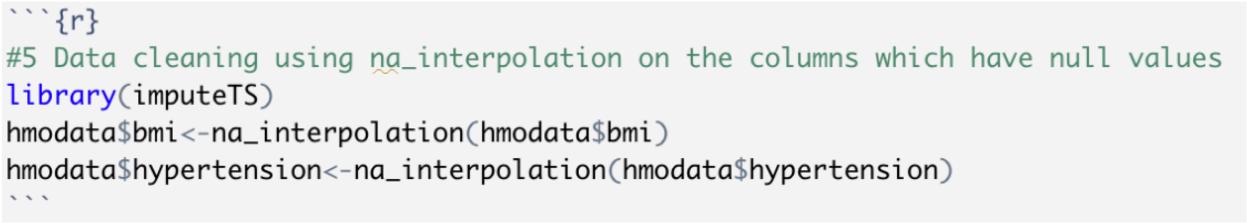
We were provided with a link to the dataset, which we had to copy and save as a.csv file. The collection comprises HMO (Health Management Organization) healthcare cost statistics. There are 14 columns in this data set.csv file. In R Studio, we imported the dataset into a new data frame called "hmodata" by using the read_csv() function.

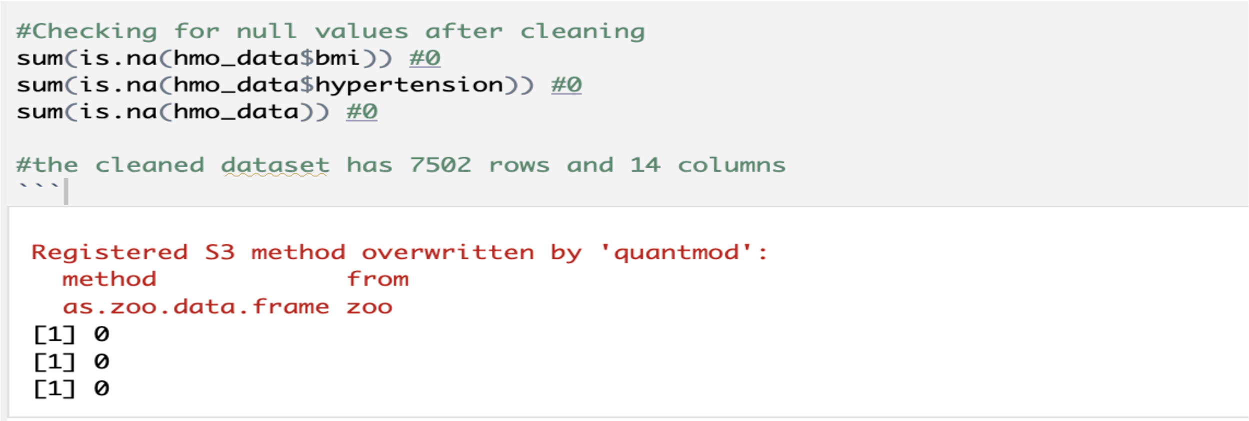


~~~
```{r}
#1 loading out data as a dataframe into new dataframe variable "hmodata" and then viewing basic information regarding the dataframe
library(tidyverse)
hmodata<-data.frame(read_csv("Data.csv"))#keep the data excel file in the same path as the rmd file
str(hmodata)
~~~
```

```
Rows: 7582 Columns: 14— Column specification
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr (8): smoker, location, location_type, education_level, yearly_physical, exercise, married, gender
dbl (6): X, age, bmi, children, hypertension, cost
ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.'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 ...
```

Data Cleansing:

Another issue in the dataset of the healthcare industry is missing values in datasets. Some values in certain features are frequently missing. This is because doctors do not always take all of the essential lab measurements, or the data is lost. e.g. '?', 'n/a', '0', '-10'. As a result, we looked for null values in the data frame's columns with numeric data types and discovered them. Please see the accompanying screenshot.



```
#4 Checking for null values in the columns of the dataframe which have numeric data type
sum(is.na(hmodata$age))
sum(is.na(hmodata$bmi)) #We see 78 null values
sum(is.na(hmodata$children))
sum(is.na(hmodata$hypertension)) #We see 80 null values
sum(is.na(hmodata$cost))
```
```

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

We utilized "na_interpolation" on the "bmi" and "hypertension" columns to remove the null values, as shown in the picture.



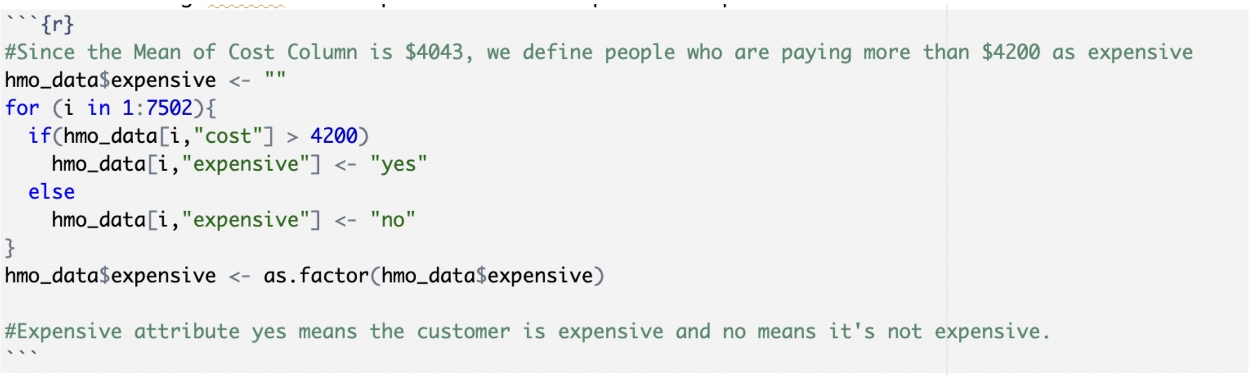
~~~
```{r}
#5 Data cleaning using na_interpolation on the columns which have null values
library(imputeTS)
hmodata$bmi<-na_interpolation(hmodata$bmi)
hmodata$hypertension<-na_interpolation(hmodata$hypertension)
~~~
```

We tested for null values again after cleaning the data with na_interpolation, and they were now 0.



```
#Checking for null values after cleaning
sum(is.na(hmo_data$bmi)) #0
sum(is.na(hmo_data$hypertension)) #0
sum(is.na(hmo_data)) #0

#the cleaned dataset has 7502 rows and 14 columns
```

 Registered S3 method overwritten by 'quantmod':
   method   from
   as.zoo.data.frame zoo
 [1] 0
 [1] 0
 [1] 0
```

Values could be Incorrect: One approach for quickly determining whether there are any inaccurate values in the dataset is to use the "Pandas function df.describe()" to see statistical aspects of a specific feature. It is effective for numerical features.

Data Transformation:

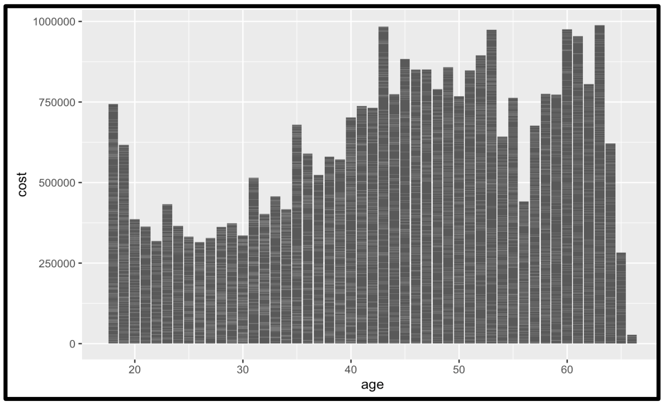
We created a new column, named cost_status, where cost > 4200 is assigned value 1, else 0. .



~~~
```{r}
#Since the Mean of Cost Column is $4043, we define people who are paying more than $4200 as expensive
hmo_data$expensive <- ""
for (i in 1:7502){
  if(hmo_data[i,"cost"] > 4200)
    hmo_data[i,"expensive"] <- "yes"
  else
    hmo_data[i,"expensive"] <- "no"
}
hmo_data$expensive <- as.factor(hmo_data$expensive)

#Expensive attribute yes means the customer is expensive and no means it's not expensive.
```
~~~

DESCRIPTIVE STATISTICS & VISUALIZATIONS

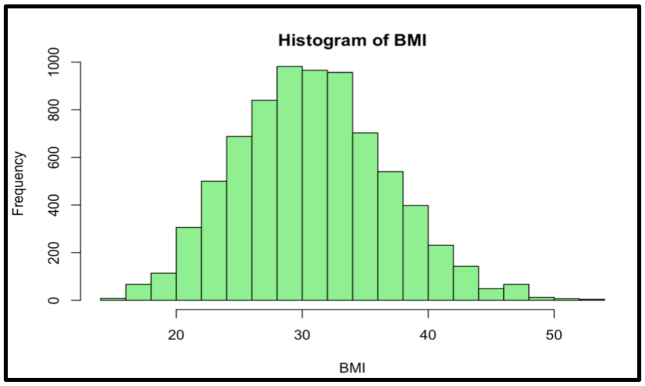
- **Age v/s Cost Barplot:**



And we noticed that prices are high in adolescence, drop dramatically for young people, and then gradually rise with age.

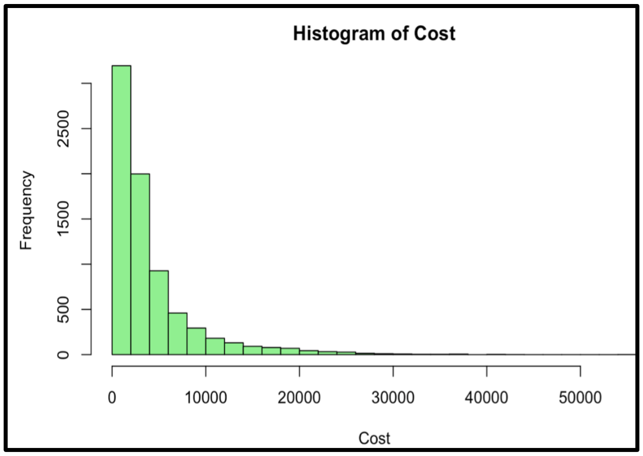
- **Histograms for Distribution of Quantitative Variables:**

A. BMI



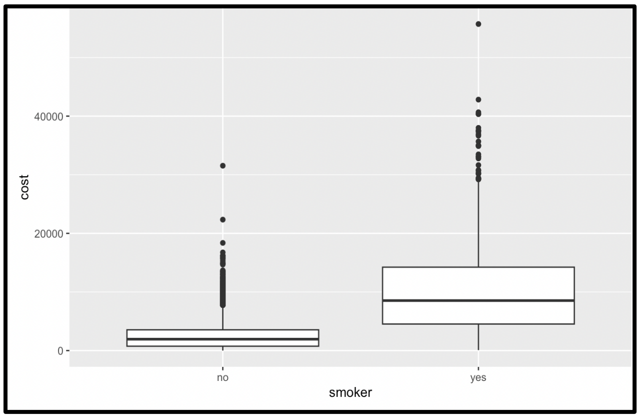
We may notice a normal distribution in the BMI Histogram. The majority of the values cluster in the middle of the range, about 30, and the remainder taper off symmetrically toward either end.

B. COST



When we plot a Cost Histogram, we see a right skewed distribution, which means that the peak of the graph is to the left of the center. That is, those with much greater costs are in short supply.

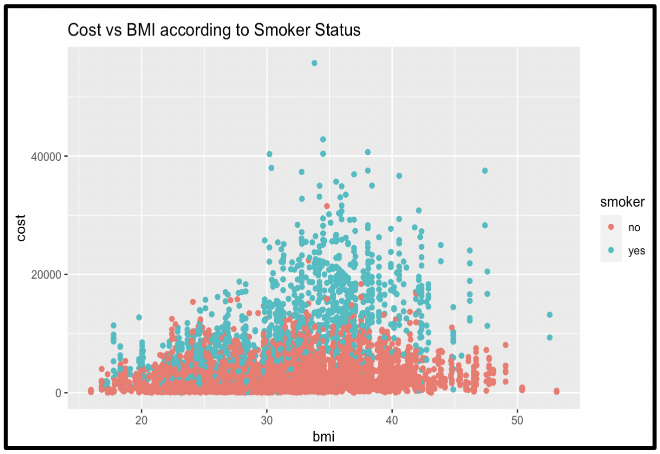
- **Box plot to determine the outliers:**



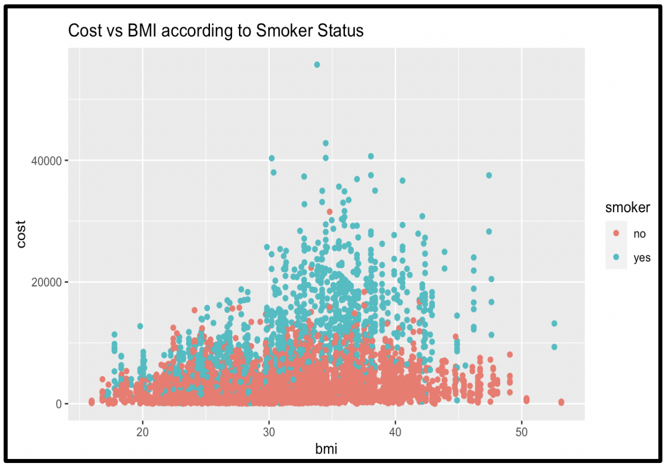
Remarks: We can see that the costs for smokers are higher than for non-smokers.

- **Scatterplots**

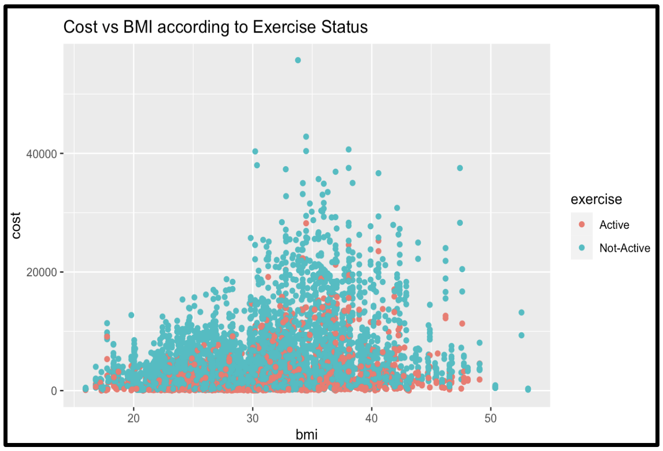
A. Smokers:



B. Getting Physical Tests Regularly:

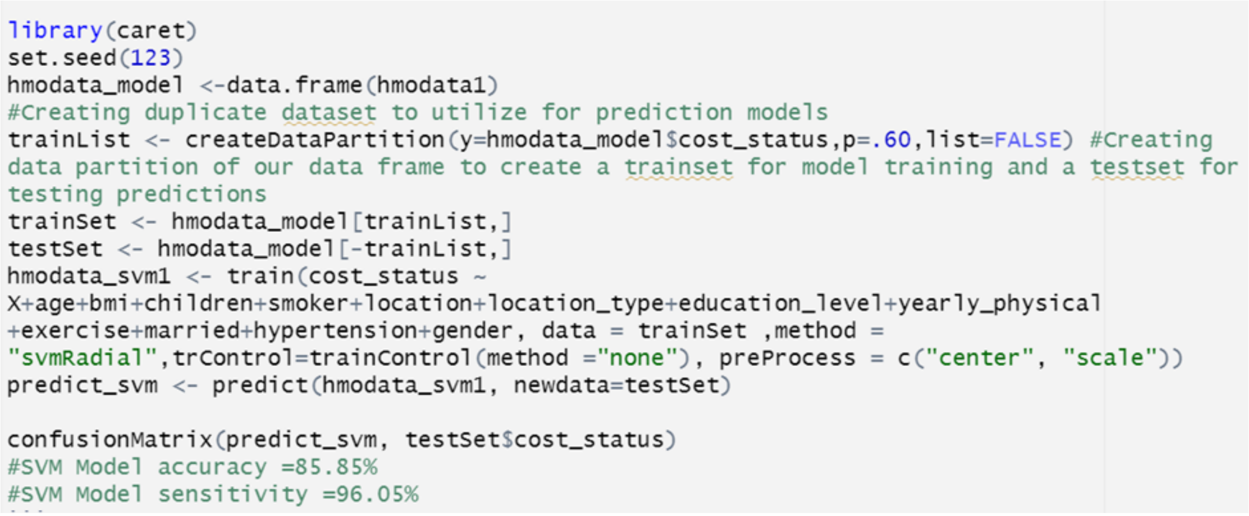


C. Exercise Impact:

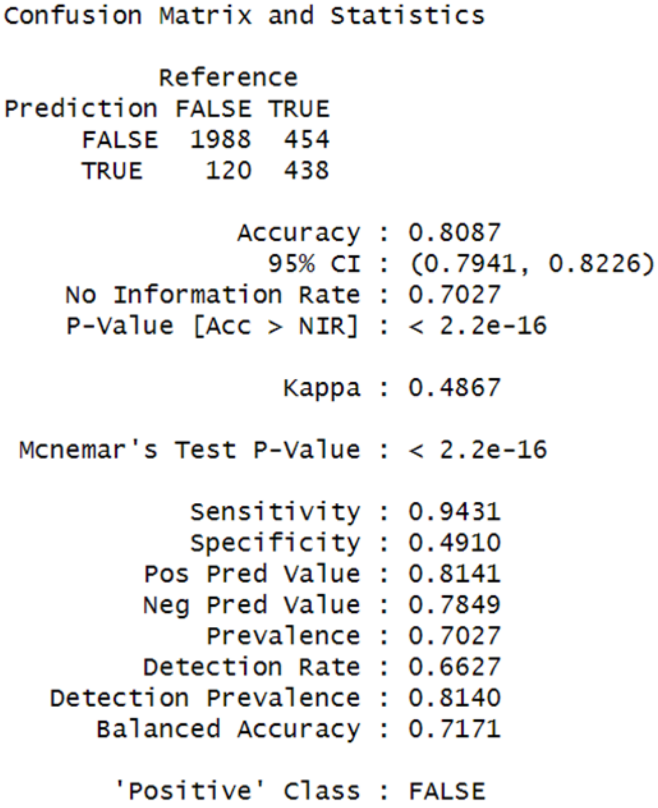


USE OF MODELING TECHNIQUES:

SVM PREDICTION MODEL:

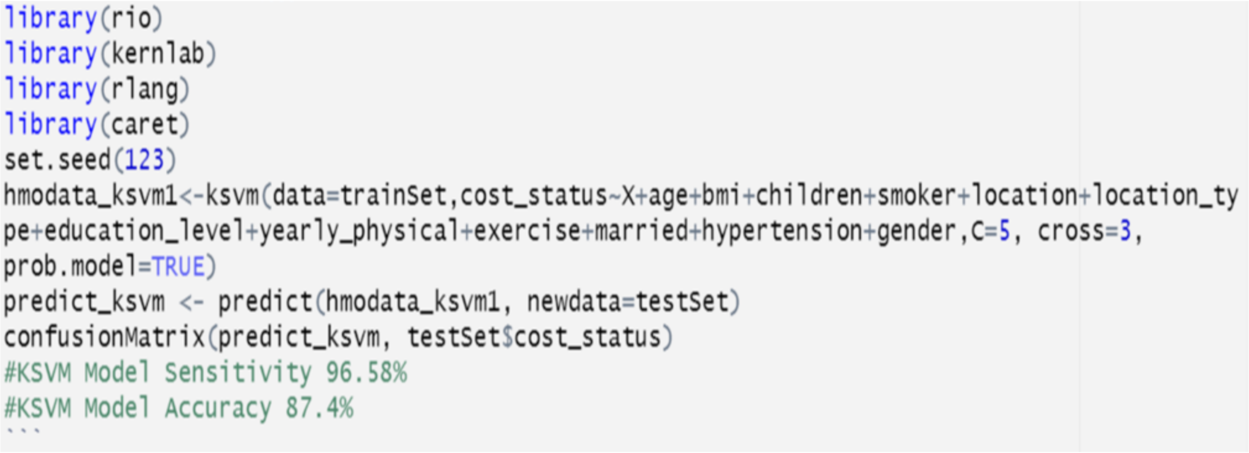


```
library(caret)
set.seed(123)
hmodata_model <-data.frame(hmodata1)
#Creating duplicate dataset to utilize for prediction models
trainList <- createDataPartition(y=hmodata_model$cost_status,p=.60,list=FALSE) #Creating
data partition of our data frame to create a trainset for model training and a testset for
testing predictions
trainSet <- hmodata_model[trainList,]
testSet <- hmodata_model[-trainList,]
hmodata_svm1 <- train(cost_status ~
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"))
predict_svm <- predict(hmodata_svm1, newdata=testSet)

confusionMatrix(predict_svm, testSet$cost_status)
#SVM Model accuracy =85.85%
#SVM Model sensitivity =96.05%
```

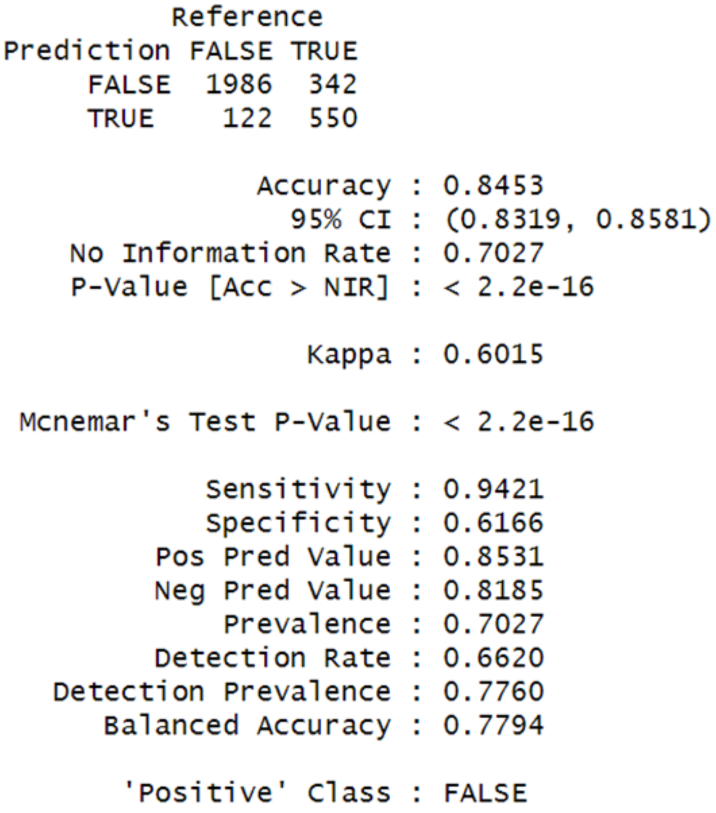


```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1988  454
     TRUE    120  438

               Accuracy : 0.8087
                 95% CI : (0.7941, 0.8226)
    No Information Rate : 0.7027
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4867

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

            Sensitivity : 0.9431
            Specificity : 0.4910
         Pos Pred Value : 0.8141
         Neg Pred Value : 0.7849
             Prevalence : 0.7027
         Detection Rate : 0.6627
   Detection Prevalence : 0.8140
      Balanced Accuracy : 0.7171

       'Positive' Class : FALSE
```

KVSM PREDICTION MODEL:

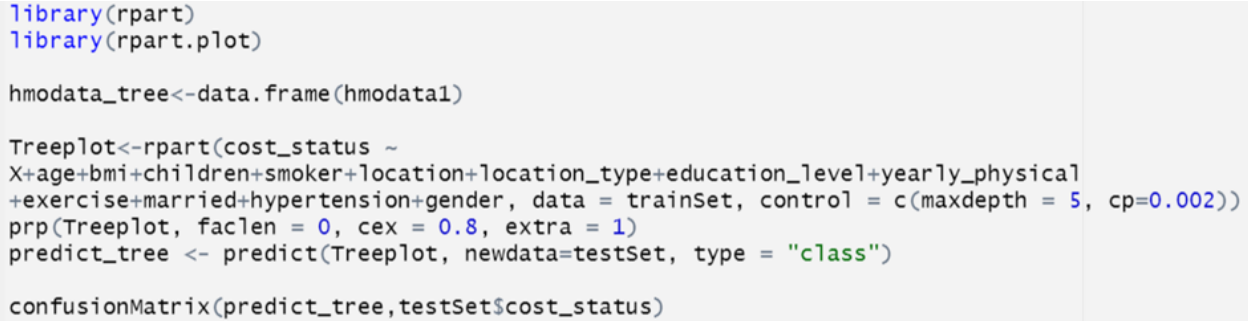


~~~
library(rio)
library(kernlab)
library(rlang)
library(caret)
set.seed(123)
hmodata_ksvm1<-ksvm(data=trainSet,cost_status~X+age+bmi+children+smoker+location+location_ty
pe+education_level+yearly_physical+exercise+married+hypertension+gender,C=5, cross=3,
prob.model=TRUE)
predict_ksvm <- predict(hmodata_ksvm1, newdata=testSet)
confusionMatrix(predict_ksvm, testSet$cost_status)
#KSVM Model Sensitivity 96.58%
#KSVM Model Accuracy 87.4%
```
~~~

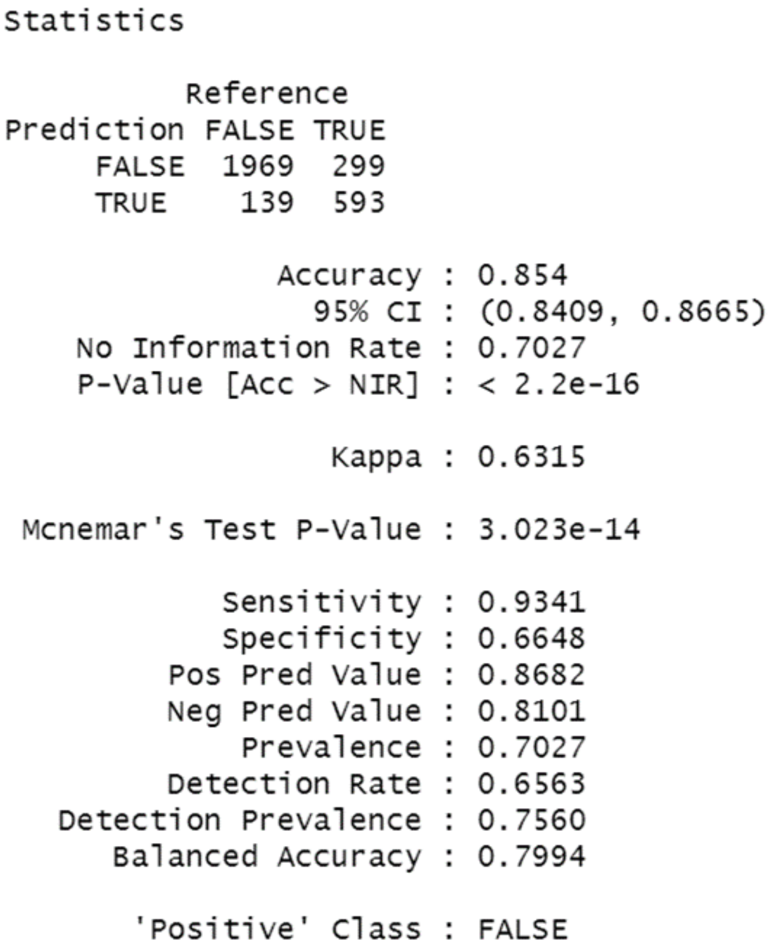


```
          Reference
Prediction FALSE TRUE
     FALSE  1986  342
     TRUE    122  550

               Accuracy : 0.8453
                 95% CI : (0.8319, 0.8581)
    No Information Rate : 0.7027
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6015

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

            Sensitivity : 0.9421
            Specificity : 0.6166
         Pos Pred Value : 0.8531
         Neg Pred Value : 0.8185
             Prevalence : 0.7027
         Detection Rate : 0.6620
   Detection Prevalence : 0.7760
      Balanced Accuracy : 0.7794

       'Positive' Class : FALSE
```

RPART MODEL:



```
library(rpart)
library(rpart.plot)

hmodata_tree<-data.frame(hmodata1)

Treeplot<-rpart(cost_status ~
X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical
+exercise+married+hypertension+gender, data = trainSet, control = c(maxdepth = 5, cp=0.002))
prp(Treeplot, faclen = 0, cex = 0.8, extra = 1)
predict_tree <- predict(Treeplot, newdata=testSet, type = "class")

confusionMatrix(predict_tree,testSet$cost_status)
```



```
Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1969  299
     TRUE    139  593

               Accuracy : 0.854
                 95% CI : (0.8409, 0.8665)
    No Information Rate : 0.7027
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6315

 Mcnemar's Test P-Value : 3.023e-14

            Sensitivity : 0.9341
            Specificity : 0.6648
         Pos Pred Value : 0.8682
         Neg Pred Value : 0.8101
             Prevalence : 0.7027
         Detection Rate : 0.6563
   Detection Prevalence : 0.7560
      Balanced Accuracy : 0.7994

       'Positive' Class : FALSE
```

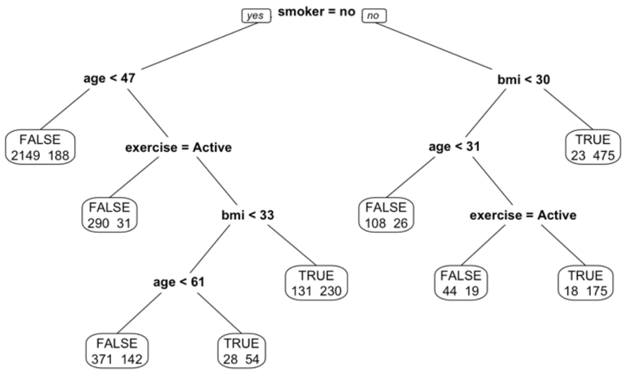
ACTIONABLE INSIGHTS / OVERALL INTERPRETATION OF RESULTS

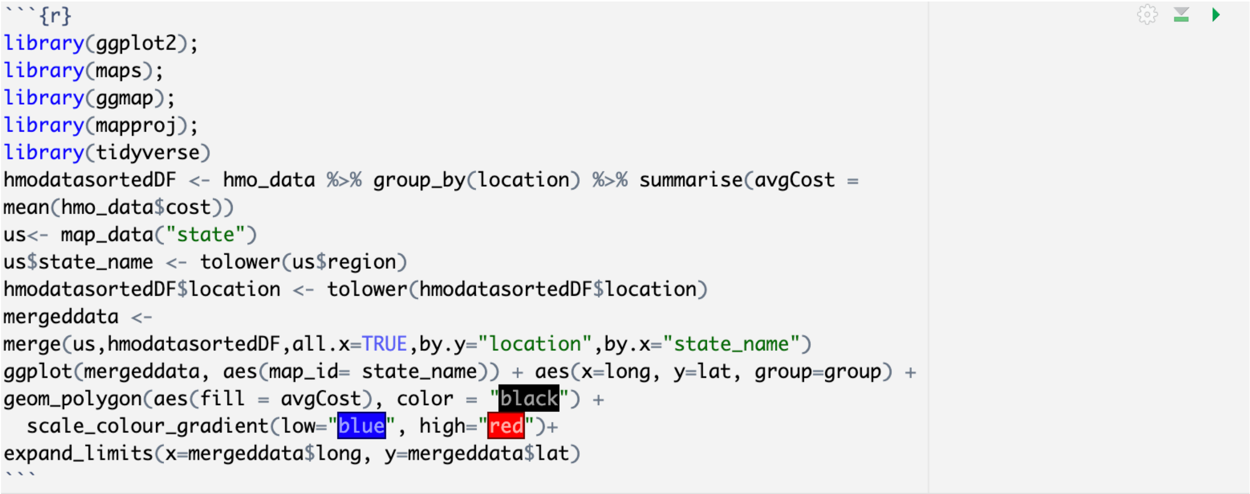
1. **Encourage Healthy Habits:** The data suggests that individuals who are physically active tend to have lower medical costs. The HMO could consider promoting and incentivizing healthy habits such as regular exercise to help prevent chronic health conditions that can lead to costly medical interventions.

2. **Offer Smoking Cessation Programs:** The data suggests that smokers tend to have higher medical costs. The HMO could consider offering smoking cessation programs to help individuals quit smoking and reduce their risk of developing costly healthy conditions.

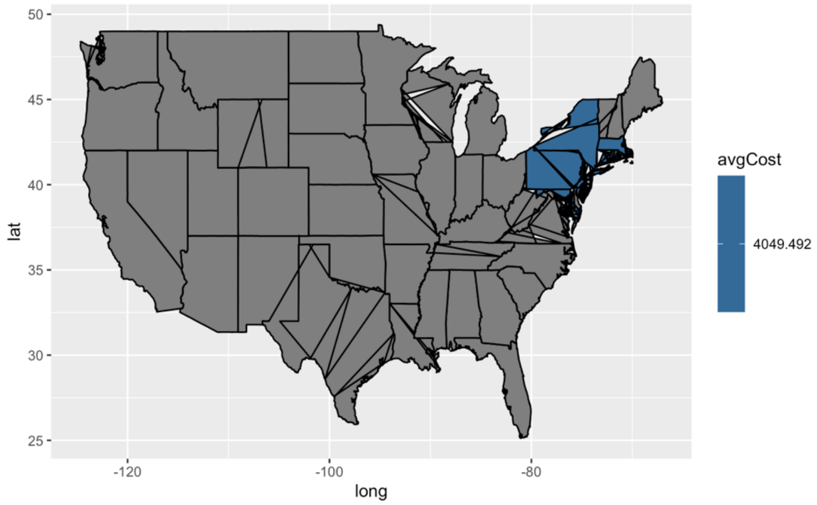
3. **Monitor High-Risk Individuals:** The data also indicates that individuals with hypertension tend to have higher medical costs. The HMO could consider implementing a monitoring system to identify and manage high-risk individuals with chronic health conditions to prevent costly complications.

APPENDIX-CODE: (Things which, in our opinion, did not materialize as expected)





~~~
```{r}
library(ggplot2);
library(maps);
library(ggmap);
library(mapproj);
library(tidyverse)
hmodatasortedDF <- hmo_data %>% group_by(location) %>% summarise(avgCost =
mean(hmo_data$cost))
us<- map_data("state")
us$state_name <- tolower(us$region)
hmodatasortedDF$location <- tolower(hmodatasortedDF$location)
mergeddata <-
merge(us,hmodatasortedDF,all.x=TRUE,by.y="location",by.x="state_name")
ggplot(mergeddata, aes(map_id= state_name)) + aes(x=long, y=lat, group=group) +
geom_polygon(aes(fill = avgCost), color = "black") +
  scale_colour_gradient(low="blue", high="red")+
expand_limits(x=mergeddata$long, y=mergeddata$lat)
```
~~~


50
45
40
35
30
25
lat
-120
-100
-80
long
avgCost
4049.492