Predicting the cost of health-care using data science

Riya Sawant

Sepideh Namvarrad

Aryan Kakade

Di Wang

Introduction:

A person who requires health insurance can choose from a number of insurance companies, each with their own special benefits. A health maintenance organization (HMO), a form of insurance that offers coverage through a network of doctors, is one of the most well-liked insurance providers on the Health Insurance Marketplace. HMO data from the dataset must be evaluated to discover what is pricey and inexpensive. Our study uses analytics to try to provide important information about the Health Maintenance Organization. Some of these traits, including smoking, exercise, BMI, and hypertension, among others, may have an effect on the cost variable. By analyzing historical data that discloses cost based on the choices chosen by consumers when purchasing insurance, it is crucial to estimate if the cost will still be prevalent. This can help the health industry make a more accurate prediction.

Business Aims:

Our aim as consultants is to predict how much a person needs to pay based on the various factors like age, bmi, smoke etc.

Questions:

1. Is there any pattern between the factors for predicting cost?
2. Is smoking related to cost increase?
3. Does checking up from a doctor have an influence on the cost being expensive?
4. Is smoking, yearly physical and exercise influence cost?

Technical Details:

Dataset: The dataset has 7582 rows and 14 columns

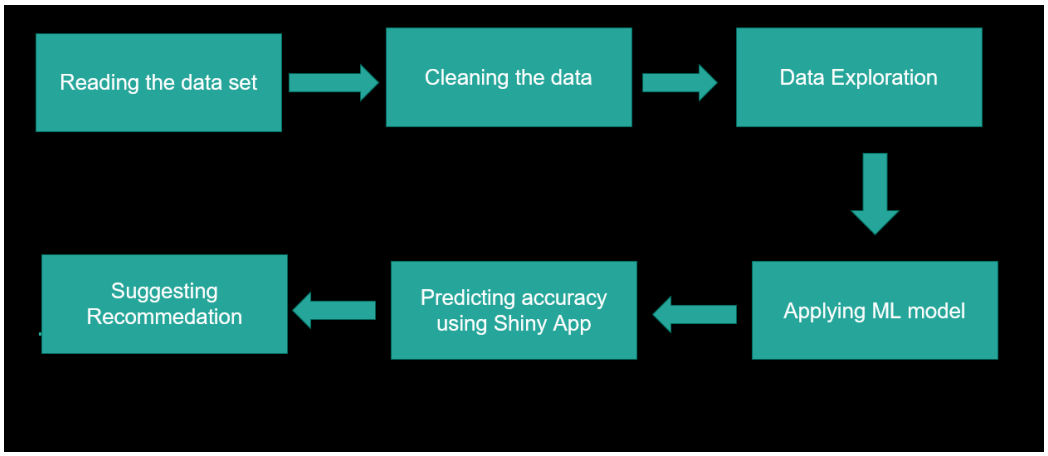
Columns Name	Description
X	Integer, unique numbers
age	Integer, How much old the person is
location	Categorical, name the state in United states where he lives
location _type	Categorical, Is that location urban or rural
Exercise	Categorical, Does he exercise regularly? Non-Active if he/she does exercise regularly other ‘active’
smoker	Integer, ‘1’ if they smoke regularly or ‘0’ if not
bmi	Double, What is his/her body mass index
yearly_physical	Categorical,Do they visit doctor regularly, ‘Yes’ if they do else ‘No’
Hypertension	‘0’ if he does not have hypertension or ‘1’ if they have hypertension

gender	Categorical, Gender of the person “Male” and “Female”
education_level	Categorical, The valid options are “No College Degree", "Bachelor", "Master", “PhD"
married	Categorical, “Yes” if married, “No” if not
num_children	Integer, number of childrens
cost	Integer, total cost of health care

Goal:

1. To predict whether the hospital cost is expensive or not.
2. To build a model with highest accuracy using Shiny app.
3. Recommend solution as per analysis.

Process:



1. **Loading the dataset:** The summary gives the mean, median and mode of all the factors that are being considered for analysis and prediction purpose.

   Code:

   ```
   library(tidyverse)
   #setwd("/Users/sepid/Desktop/Lab-datascience" )
   #data<- read_csv("Sciene.csv")
   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
Length:7582        Length:7582        Length:7582
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


education_level    yearly_physical      exercise
Length:7582        Length:7582        Length:7582
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


  married           hypertension       gender              cost
Length:7582        Min.   :0.0000   Length:7582        Min.   :    2
Class :character   1st Qu.:0.0000   Class :character   1st Qu.:  970
Mode  :character   Median :0.0000   Mode  :character   Median : 2500
                   Mean   :0.2005                      Mean   : 4043
```

2. **Cleaning the data**: Cleaning the data can help you preserve error-free data and can greatly increase its dependability, accuracy, and validity. Before cleaning there were 78 missing values in BMI and 80 missing values in hypertension.

   Code:

   ```
   # There are missing values in BMI and hypertension

   #library(tidyverse)

   missing_bm <- nrow(data[is.na(data$bmi),])

   missing_bm


   #library(tidyverse)

   missing_ht <- nrow(data[is.na(data$hypertension),])

   missing_ht
   ```

For filling the missing values in BMI we have used mean of the existing values and similarly for hypertension we have used the technique to fill the values in upper direction.

Code:

```
#Filling the missing value with mean
data$bmi[is.na(data$bmi)]<-mean(data$bmi,na.rm=TRUE)
data
missing_bmi <- nrow(data[is.na(data$bmi),])
missing_bmi


data <- data %>% fill(hypertension, .direction = 'up')
missing_ht <- nrow(data[is.na(data$hypertension),])
missing_ht
```

3. **Finding the significance**: Statistical significance aids in determining whether a result is more likely the result of chance or an important cause. The gain ratio determines the significance of variables.

   Code:

```
library('FSelector')
ratio <- gain.ratio(cost~., data)
ratio
#smoker=0.3300140112
#age=0.1520260647
#exercise=0.0554992165
```

#bmi=0.0401061839

#children=0.0336380767

The five most significant variables after gain ratio analysis are smoker, age, exercise, bmi and children with significance of 0.33, 0.15, 0.05, 0.04, 0.03 respectively. The significance implies that these variables/factors have influence on the cost.

4. **Data Exploration:**

By looking at the visualizations of the data set, we tried to explore any meaningful relationship between different attributes and the cost. We observed many trends such as positive correlations between health care cost and age, smoking habits, yearly checkups with doctors, etc. However, these trends were observable in the graphics that we made, we could not completely rely on them since visualizations can not be the only helpful tool to draw conclusions about the possible relationships between attributes. Therefore we decided to take advantage of more analytical tools to more effectively measure these possibilities. One way to estimate the significance of different attributes as predictive tools is information gain ratio analysis which aims to answer the following questions by using the concept of entropy which can be defined as the level of uncertainty, impurity and disorder in a data set.

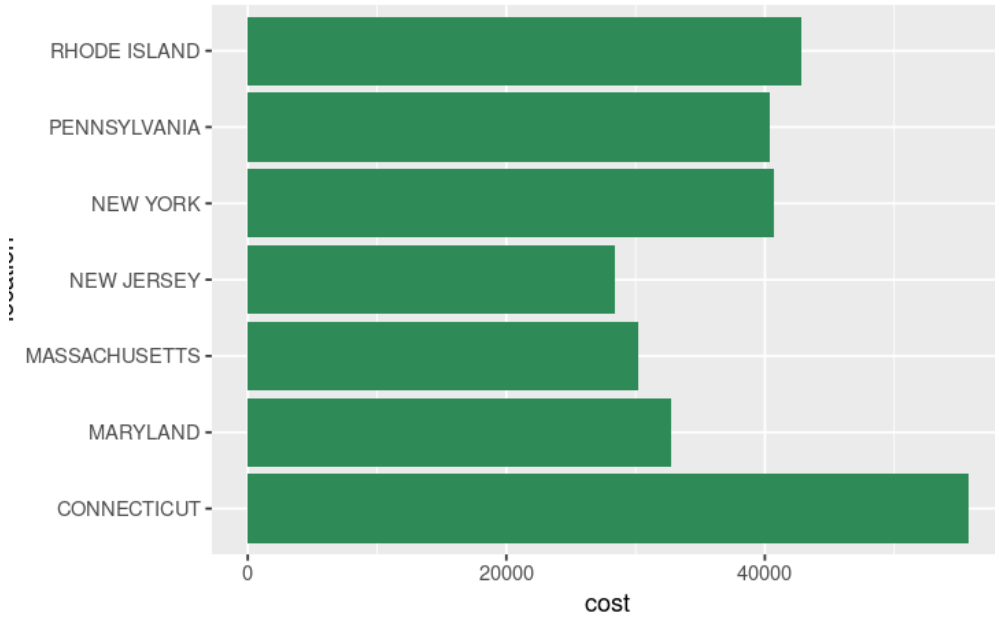


Fig: Cost Vs Location

Insight from the bar plot: Also, we used different graphs to gain insights about the demographics of the data and possible trends and relationships.

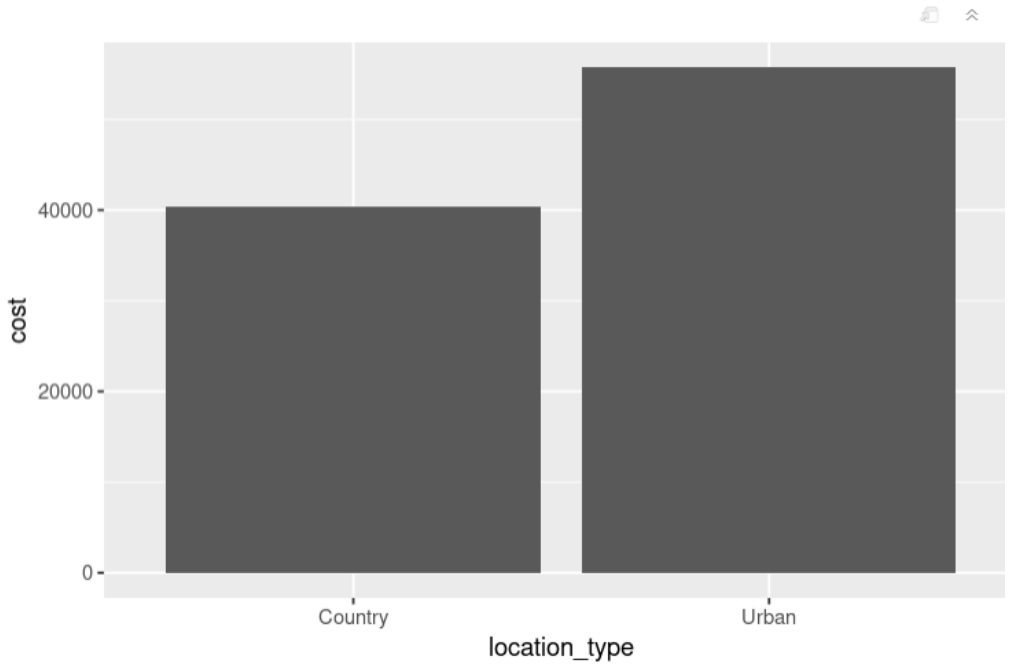


Fig: Location_type Vs Cost

Insight from the bar plot: The first graph shows that health care cost in urban areas is higher than rural areas.

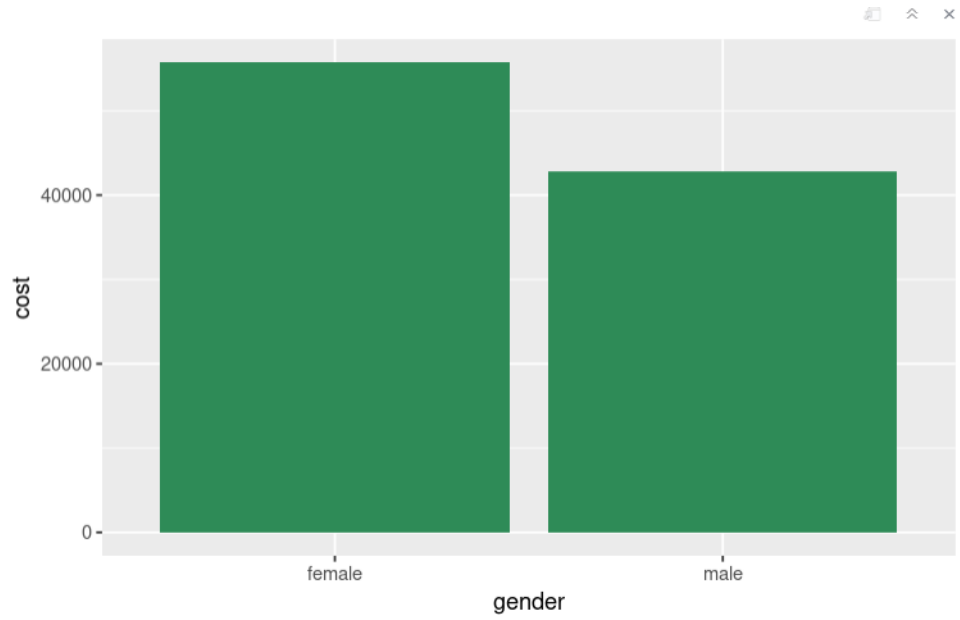


Fig: Gender Vs Cost

Insight from the bar plot: This graph shows that this number is also higher among women compared to men

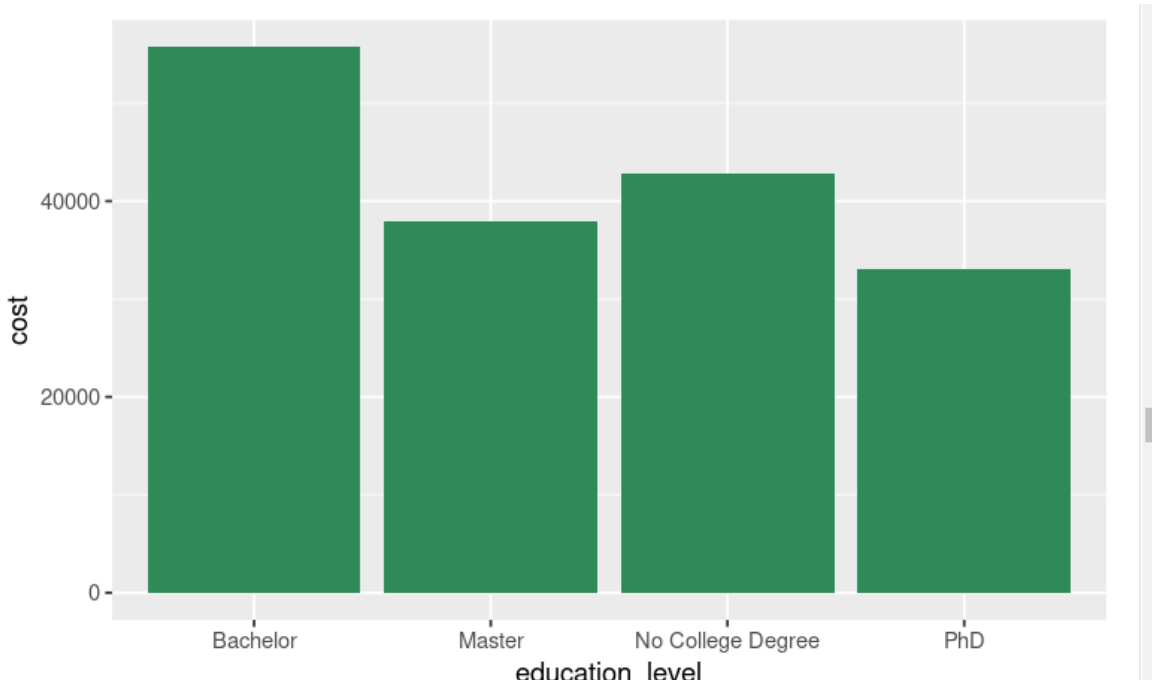


Fig: Education_level Vs Cost

Insight from the bar plot: This picture decides the cost of healthcare between people with various levels of education and as it shows, this number is highest among people with bachelor degrees. However we can not draw a new conclusion by looking at other groups as they are very similar.

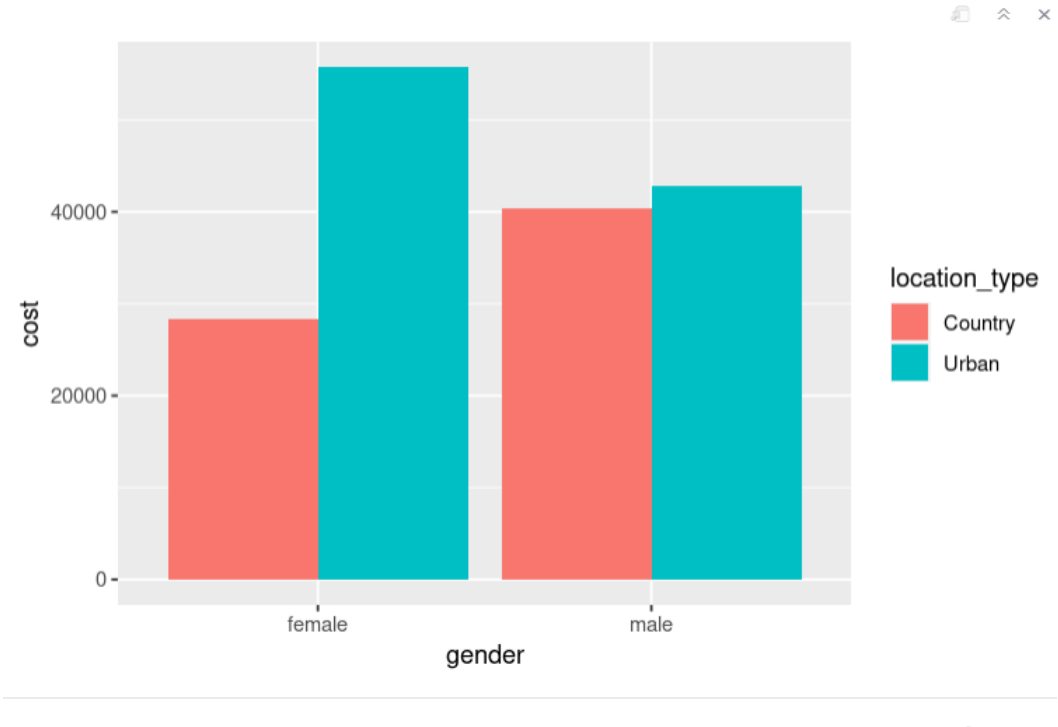


Fig: Gender Vs Cost Vs Location_type

Insight from the bar plot:

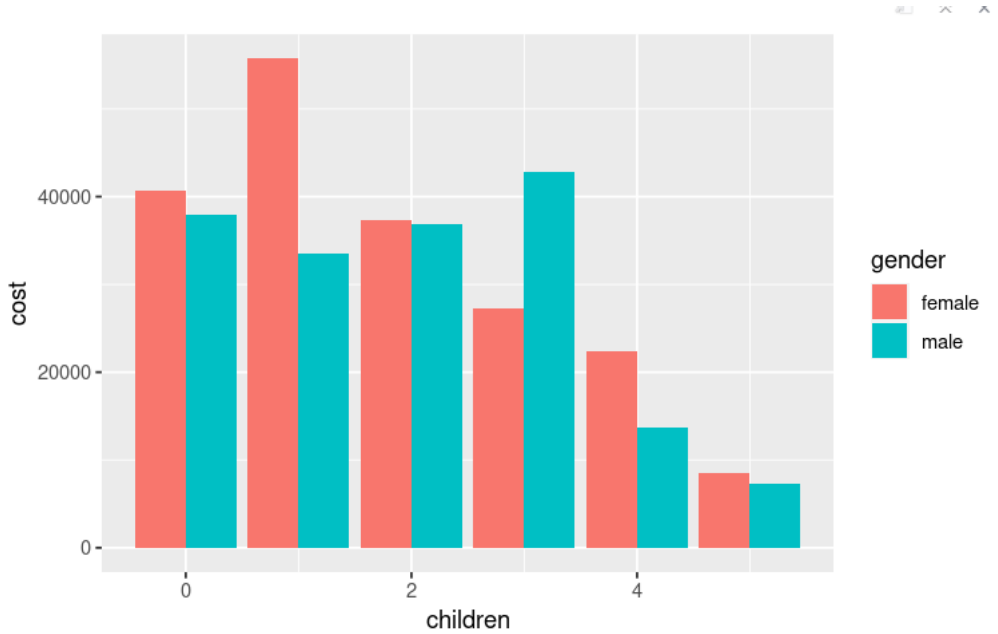
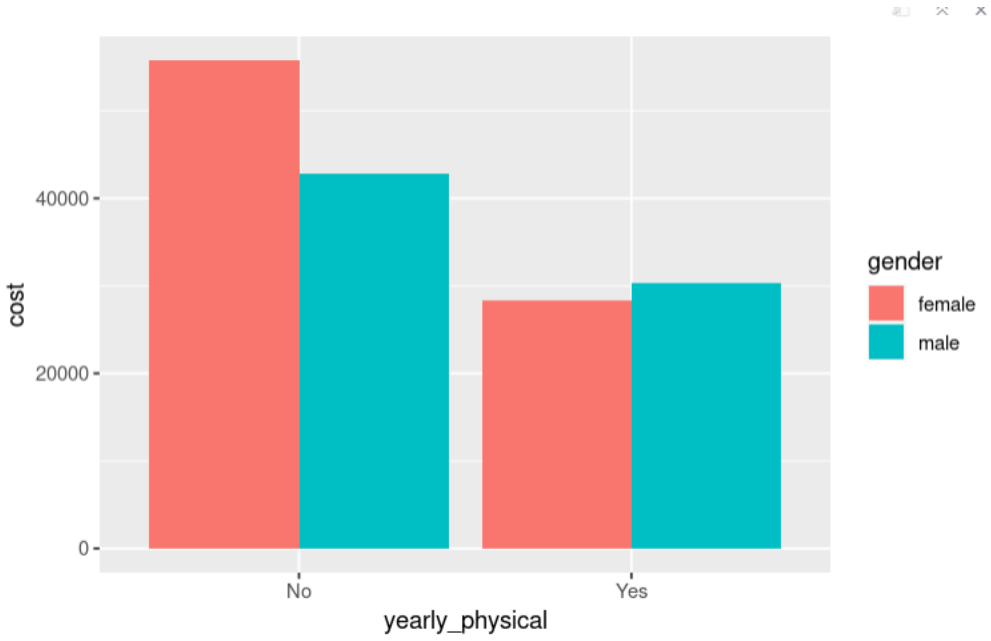


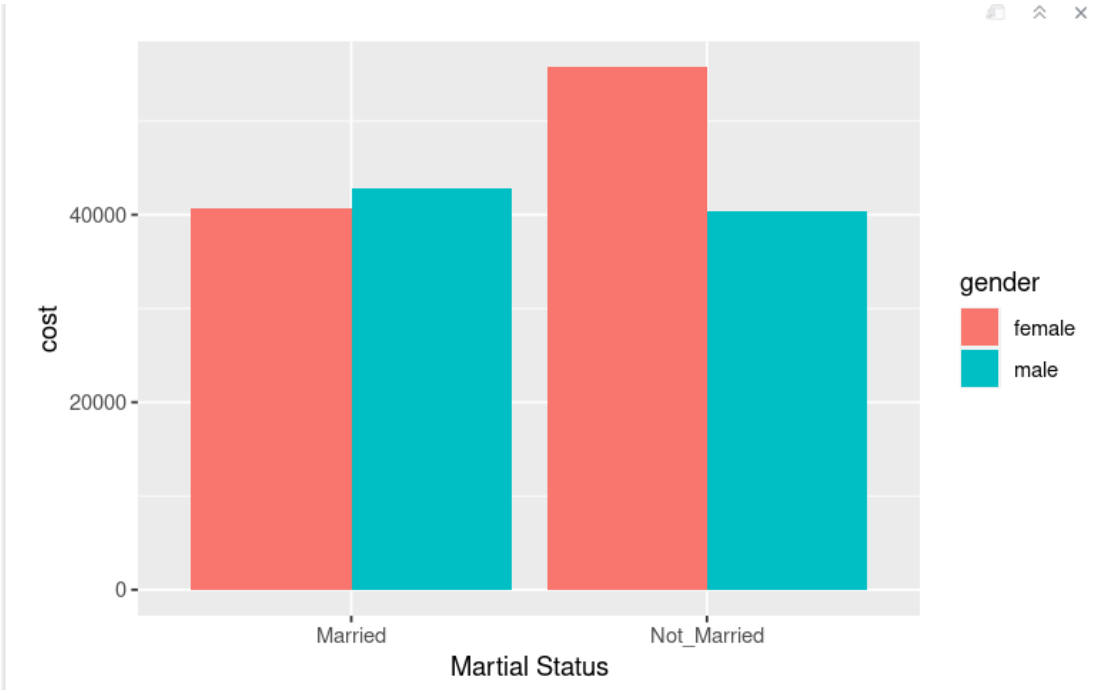
Fig: Children Vs Cost Vs Gender

Insight from the bar plot: The fifth graphs demonstrates the relationship between healthcare costs and the number of children in each family. Surprisingly as number of children increases the cost decreases.



Yearly_physical Vs Cost Vs Gender

Insight from the bar plot: The sixth graph also shows that the cost of healthcare is lower in groups that have a yearly checkup.



Marital Status Vs Cost Vs Gender

Insight from the bar plot: The seventh graph shows that the cost of healthcare is only higher for married females.

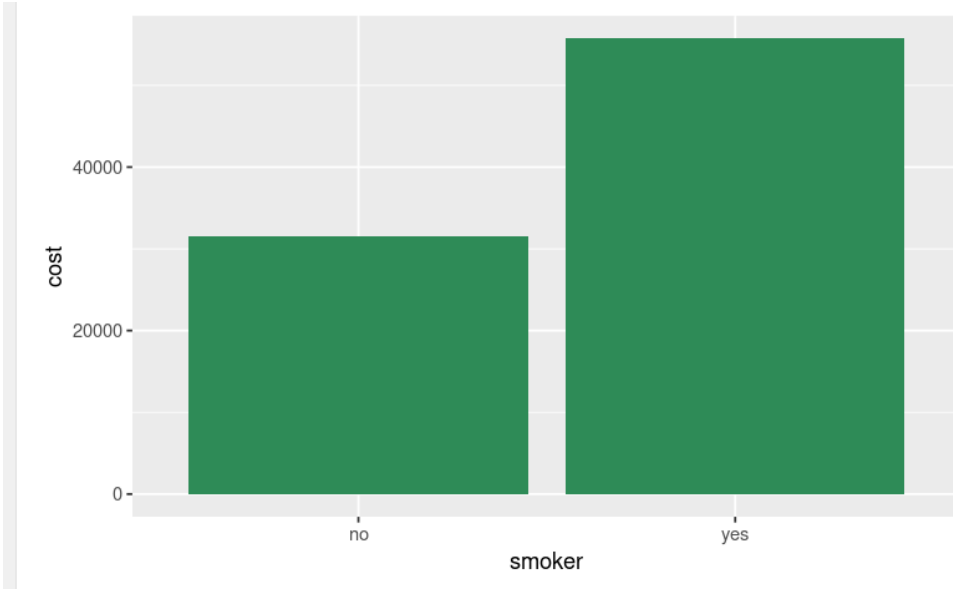
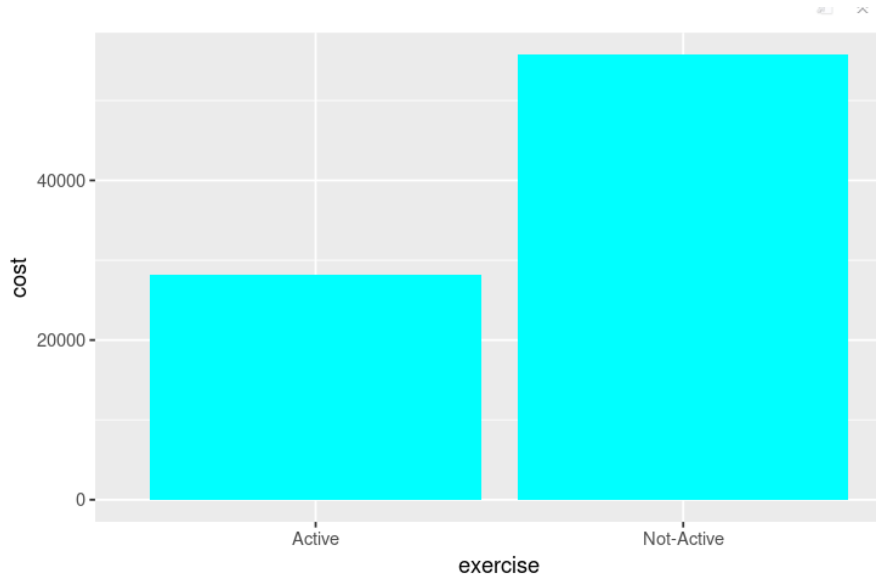


Fig: Smoker Vs Cost

Insight from the bar plot: The eighth and ninth graphs shows that whether people work out or smoke have a direct effect on their healthcare costs and people who smoke or workout less often face higher expenses in term of their health.



Exercise Vs Cost

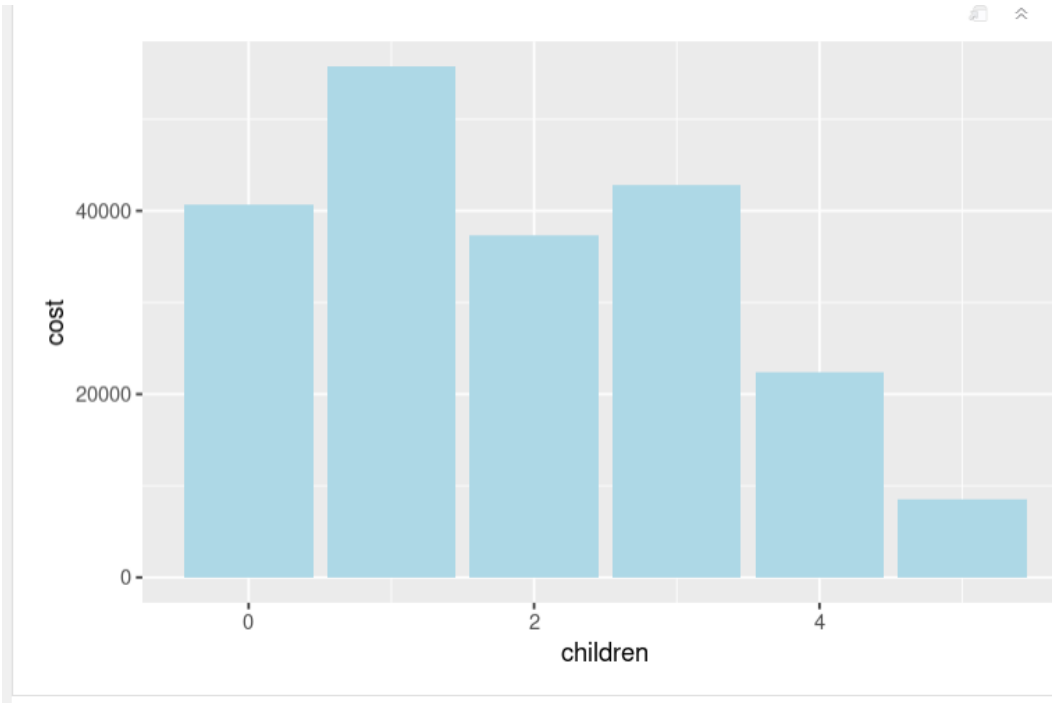


Fig: Children Vs Cost

Insight from the bar plot:

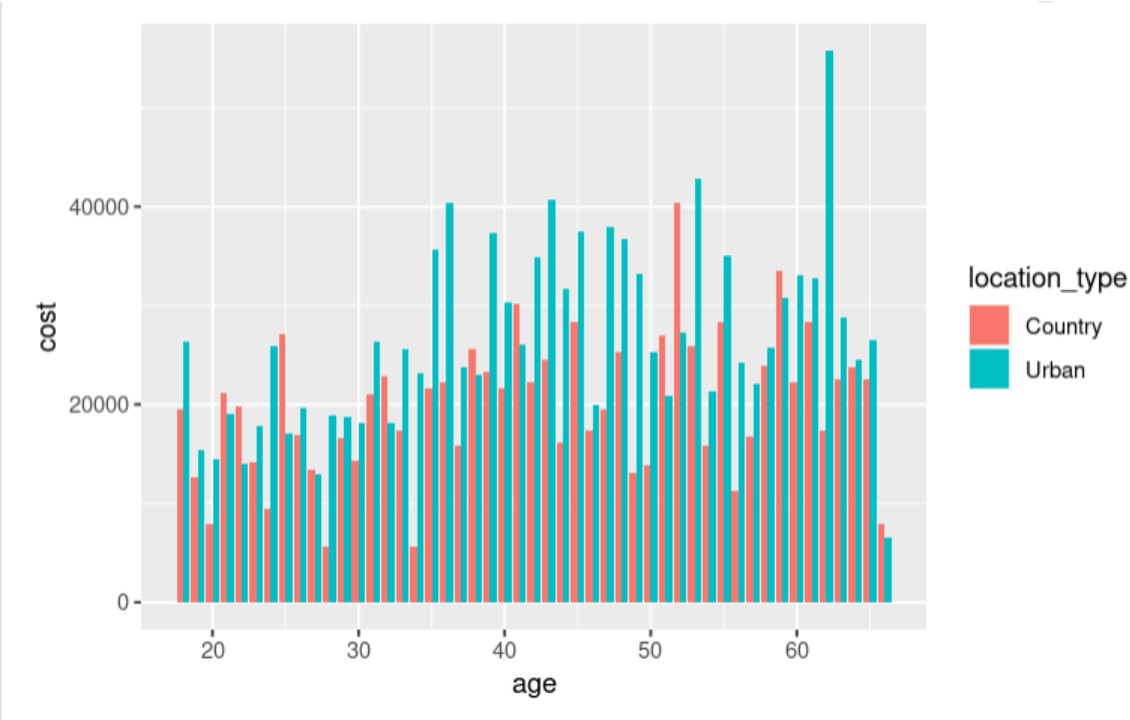


Fig: age Vs Cost Vs location_type

Insight from the bar plot: The tenth graph tries to compare healthcare costs with age and location and we can clearsee that generally as the age increases the costs relate to healthcare increases.

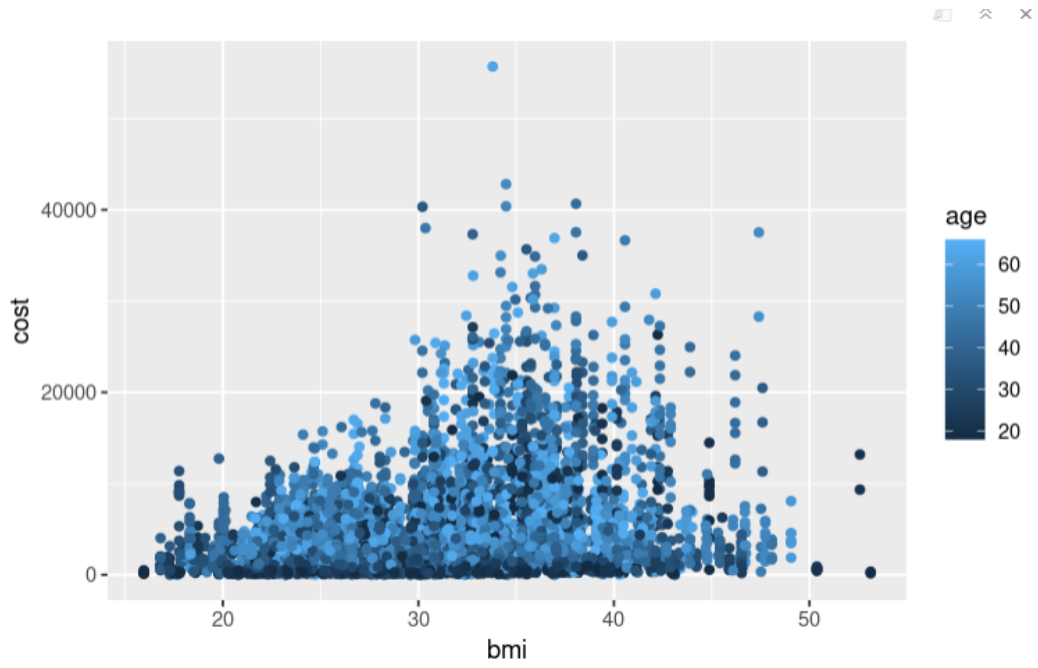


Fig: Scatter plot of bmi, cost and age

Insight from the scatter plot: And finally the last graph compares health care prices with bmi and age. Here we can see a normal distribution in bmi and healthcare costs which implies that at first there is a positive relationship between bmi and cost and after bmi reaches 35, a negative correlation is observed. This data, however does not align with the predominant belief in medical science as it predicts that there should be a positive and linear relationship between bmi and health problems which leads to higher expense

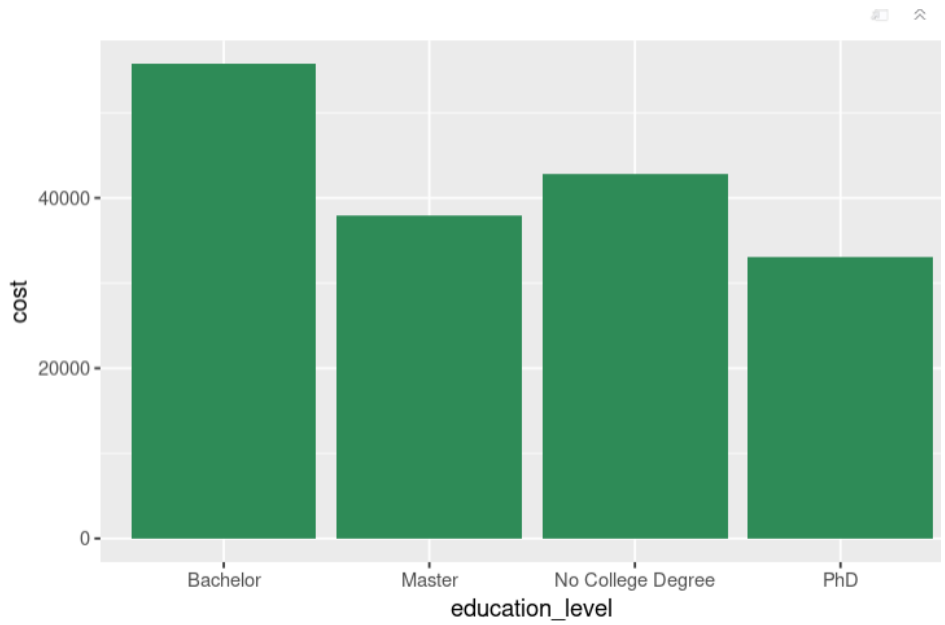


Fig: education_level Vs cost

Insight from the bar plot:

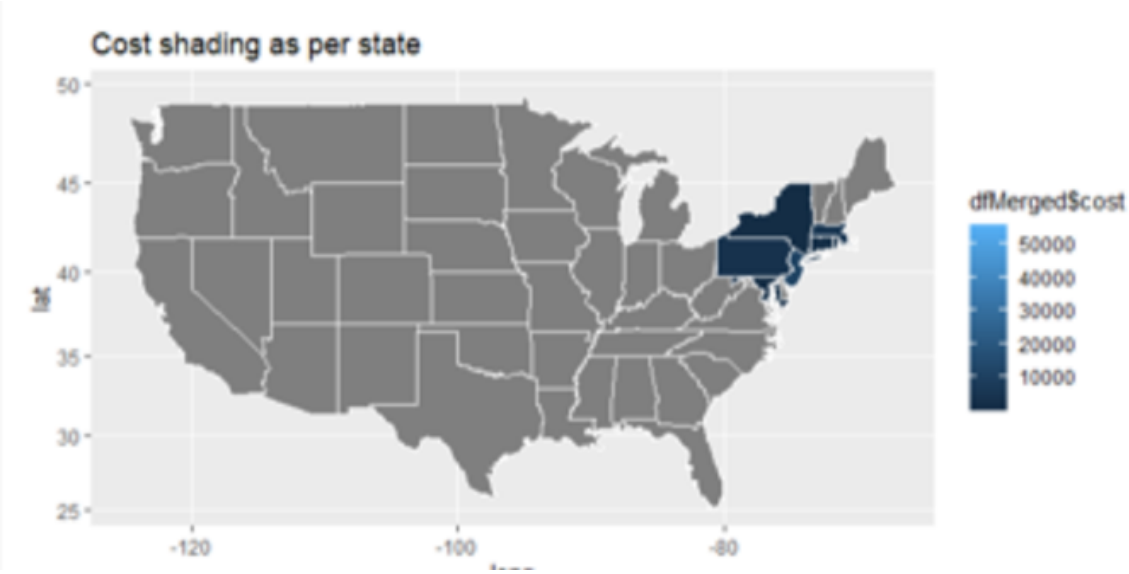


Fig: Cost shading as per state

Insight from the map: The map shows that this data set includes only a few states in the north east of the United States. Among these states New York has the highest cost of healthcare.

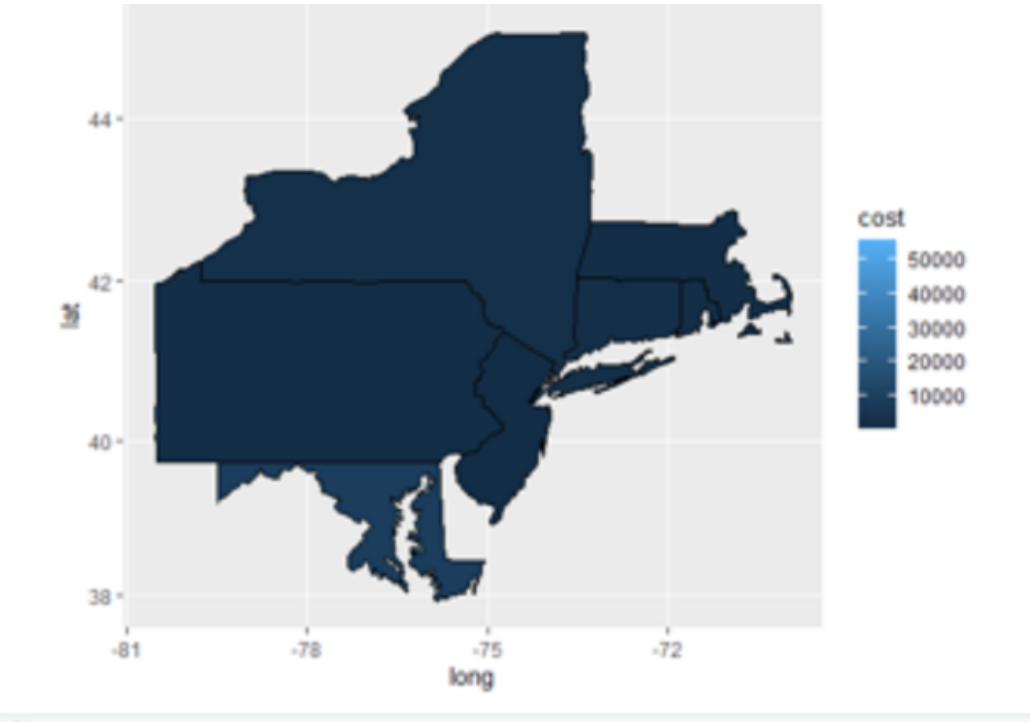


Fig: Cost per states zoomed in

Insight from the map:

5. **Applying Machine Learning:** In order to forecast new output whether the cost spent on the health care is expensive or not, machine learning algorithms use historical data from the Health management Organization as input. Here, we have used two types of model- Support Vector Machine and Decision tree.

Support Vector machine: SVMs, or support vector machines, are effective methods that enable you to verify your results and their accuracy. To check the validity, we divided the dataset into training and testing groups.

Code:

```
library(caret)
set.seed(111)
```

```
trainList2 <- createDataPartition(y=data$expensive,p=.30,list=FALSE)

trainset2 <- data[trainList2,]

testset2 <- data[-trainList2,]

library(kernlab)

svmModel2 <- ksvm(expensive ~ .,data=dataFinal2,C=4,cross=2,prob.model=TRUE)

svmModel2
```

```
□[1mindexing□[0m □[34mHMO_TEST_data_sample_solution.csv□[0m [==============

Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE     8    4
     TRUE      4    4

               Accuracy : 0.6
                 95% CI : (0.3605, 0.8088)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.5956

                  Kappa : 0.1667

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.6667
            Specificity : 0.5000
         Pos Pred Value : 0.6667
         Neg Pred Value : 0.5000
             Prevalence : 0.6000
         Detection Rate : 0.4000
   Detection Prevalence : 0.6000
      Balanced Accuracy : 0.5833

       'Positive' Class : FALSE
```

Accuracy: 60%

Sensitivity: 0.667

Decision tree: An internal node represents a feature (or property), a branch represents a decision rule, and each leaf node indicates the conclusion in a decision tree, which resembles a flowchart. The root node in a decision tree is the first node from the top. It

gains the ability to divide data according to attribute values. Recursive partitioning is the process of repeatedly dividing a tree. This framework, which resembles a flowchart, aids in decision-making. It is a flowchart-like representation that perfectly replicates how people think. Decision trees are simple to grasp and interpret because of this.

Code:

```
trainL<-createDataPartition(y=data$expensive, p=.70, list=FALSE)
train<-data[trainL,]
test<-data[-trainL,]
grid <- expand.grid(.cp=c(0.01,0.05,0.10,0.15,0.20,0.25))
dt_model <- train(expensive ~ ., data = train, metric = "Precision", method =
"rpart",tuneGrid = grid )
dt_predict <- predict(dt_model, newdata = test )
head(dt_predict, 11)


dt_model_preprune <- train(Cost ~ ., data = train, method = "rpart",
                           metric = "Precision",
                           tuneLength = 8,
                           control = rpart.control(minsplit = 50, minbucket = 20, maxdepth = 6))
```

```
□[1mindexing□[0m □[34mHMO_TEST_data_sample_solution.csv□[0m [==========

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
```

Accuracy: 65 %

Sensitivity: 0.75

6. **Providing the recommendation:**
    - Sponsoring the non smoking campaign: As a insurance company we can tie up with the non profit organization who are conducting various activities and campaigns for making people aware of side-effects of smoking to prevent people from falling prey to disease related to respiratory issue

- Maintaining a healthy diet by consulting a Doctor. By consulting the doctor on regularly basis can help person to stay fit by eating and including nutrients and essentials in their daily diet
- Fitness Campaign on Monthly Basis: The insurance company should conduct the campaign for making people physically fit by organization marathons, game day, matches etc.
- Covering Coverage Addiction Therapy.

Output on shiny App:

Website: https://riya-sawant.shinyapps.io/Predicting_accuracy_of_model/

Interface: