**IST 687 HMO Data Analysis Report**

**Aruneema**

**Harshit Joshi**

**Danila Rozhevskii**

**Vaishnav Kanekar**

**Shweta Suhas Rane**

Contents

[**1. Problem Statement** 2](#_heading=h.gjdgxs)

[**2.** **The Data** 2](#_heading=h.30j0zll)

[**2.1 EDA** 2](#_heading=h.1fob9te)

[**2.1.1** **Univariate analysis** 3](#_heading=h.3znysh7)

[**2.1.3 Bivariate analysis** 5](#_heading=h.2et92p0)

[**3.** **EDA with expensive column** 14](#_heading=h.tyjcwt)

[**4. Linear Modelling** 17](#_heading=h.3dy6vkm)

[**4.1 Linear Regression with scaled data:** 19](#_heading=h.1t3h5sf)

[**4.2 Linear Regression with unscaled but encoded Test HMO\_data** 21](#_heading=h.4d34og8)

[**5. Supervised Learning** 22](#_heading=h.2s8eyo1)

[**5.1 TRAIN-TEST-SPLIT** 22](#_heading=h.17dp8vu)

[**5.2 MODEL TRAINING** 23](#_heading=h.3rdcrjn)

[**5.2.1 KSVM** 23](#_heading=h.26in1rg)

[**5.2.2 Radial SVM** 24](#_heading=h.lnxbz9)

[**5.2.3 K-fold Validation** 24](#_heading=h.35nkun2)

[**6. Unsupervised Learning** 27](#_heading=h.1ksv4uv)

# **Problem Statement**

We are given a case by the CEO of a hospital system to analyse the data on patients and provide insights based on the data analysis. The client is particularly interested in:

* Predict people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs)
* Provide actionable insight to the HMO, in terms of how to lower their total health cost

Additionally, they are interested to know what drives the healthcare cost.

To provide these insights, we will first clean out data set and perform exploratory data analysis to understand the data. This will be followed linear regression analysis that helps in predicting the cost of patients and find which variable will affects the cost more.

Finally, we will be using different versions of supervised and unsupervised learning to find which person will expensive or not.

# **The Data**

The HMO\_data provided to us has 7582 rows and 14 columns. The 14 columns are as follows:

DATA DICTIONARY

| Variable | Variable Description |
| --- | --- |
| 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 |

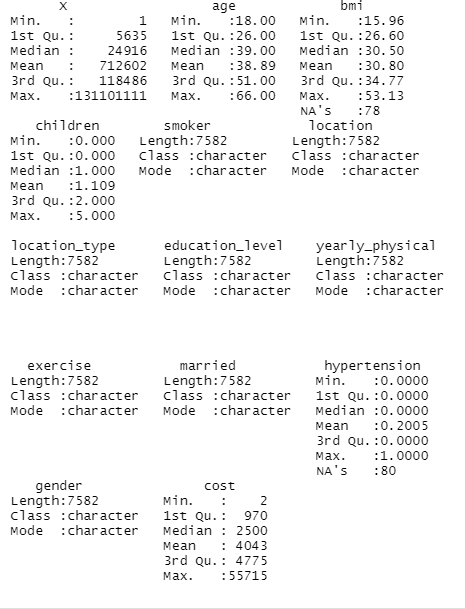
**Table 1: Data Description**

## **2.1 EDA**

The data was loaded from the following link-

<https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv>

On doing the summary analysis of each variable get the following result:



**Table 2: Data summary**

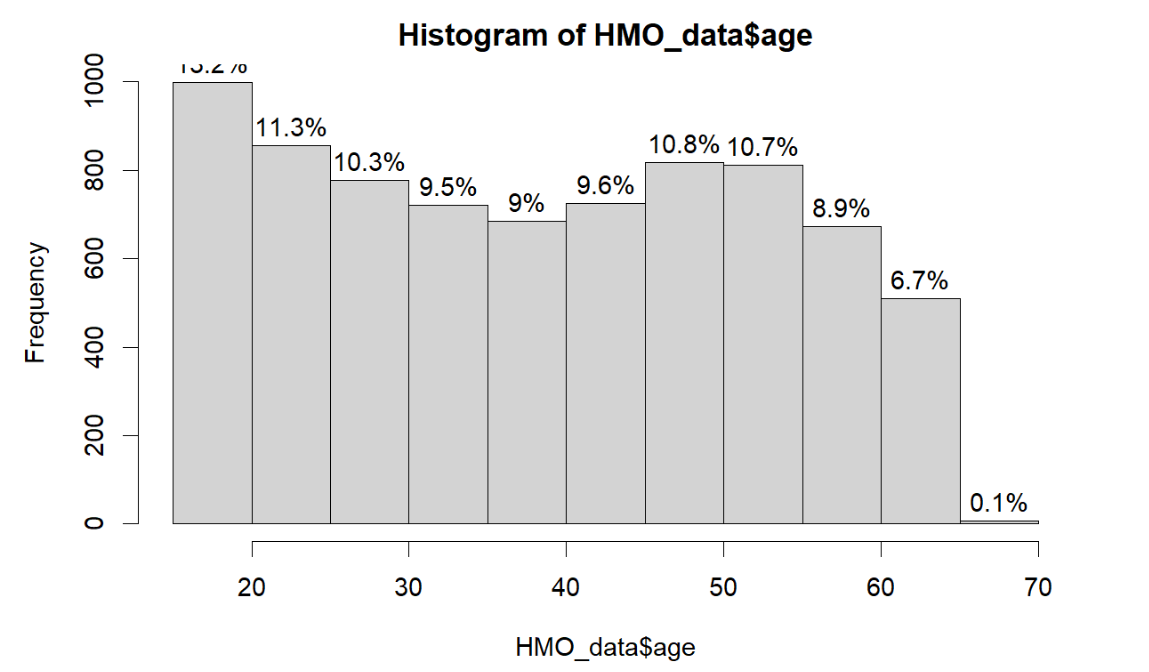
* The median value for Age is 39 where the mean value is 38.89.
* The median value for Bmi is 30.50 where the mean value is 30.80.
* The median value for cost is 2500 where the mean value is 4043.
* There are 78 NA’s in BMI and 80 NA’s in hypertension

For the NA’s we excute the na\_interpolation function to replace the missing values with the known values.

## **Univariate analysis**

Since the data has both numeric and non-numeric variables, the univariate and bivariate analysis was done separately for these two kind of variables. Combination of bar plots and histogram was used to plot the visualisation using the ggplot function.

**Plot:1 Frequency distribution of age**Chart, histogram

Description automatically generated

**Plot 2: Frequency distribution of cost**

Chart, bar chart

Description automatically generated

**Plot 3: Proportion distribution of gender**

Graphical user interface, application

Description automatically generated with medium confidence

**Insights:**

* From plot 1, the dataset has more proportion of people in their 20s (13.2%) followed by people in their late 40s to late 50s (10.8 and 10.7% respectively)
* From plot 2, the cost for most of the people lie below 10000 (84.6%). Very few proportion of people have cost above 10000
* From plot 3, it is evident that the proportion of male is slightly more than females in the given sample

## **Bivariate analysis**

A combination of scatter plots, histograms, box plots and bar plots were used to generate insights from bivariate analysis.

**Plot 4: Scatter plot of cost vs age**Chart, scatter chart

Description automatically generated

**Insights:**

* We plotted a scatterplot with the x axis as age and the y axis as cost
* This scatterplot gives us the insights of the cost with the intervals of age which ranges from 20-60

**HISTOGRAMS:**

Chart

Description automatically generated

**Plot 5: Distribution of BMI in 7 states**

Chart, histogram

Description automatically generated

**Plot 6: Distribution of BMI in Country and Urban location**

Chart

Description automatically generated

**Plot 7: Distribution of cost in 7 states**

**Insights:**

* Plot 5 shows that the BMI curve is normally distributed for all the states
* Plot 6 shows that the BMI of people living in country and urban areas follow similar distribution, with more proportion of people living in country
* Plot 7 shows that cost curve is skewed for all the seven states to the right
* Plot 8 shows that the cost for urban and country area is also skewed with count Country having higher proportion in the sample

Chart

Description automatically generated

**Plot 8: Distribution of cost in Country and urban region**

* Plot 9 shows that there are more proportion of non-smoker than smokers with the cost of both of the mostly lying below 20,000
* Plot 10 and 11 both shows that the cost of people from different educational background is skewed to the right with BMI being normally distributed. It also informs us that more people belong to Bachelor’s degree

Chart

Description automatically generated

**Plot 9: Distribution of cost in among smokers and non-smokers**

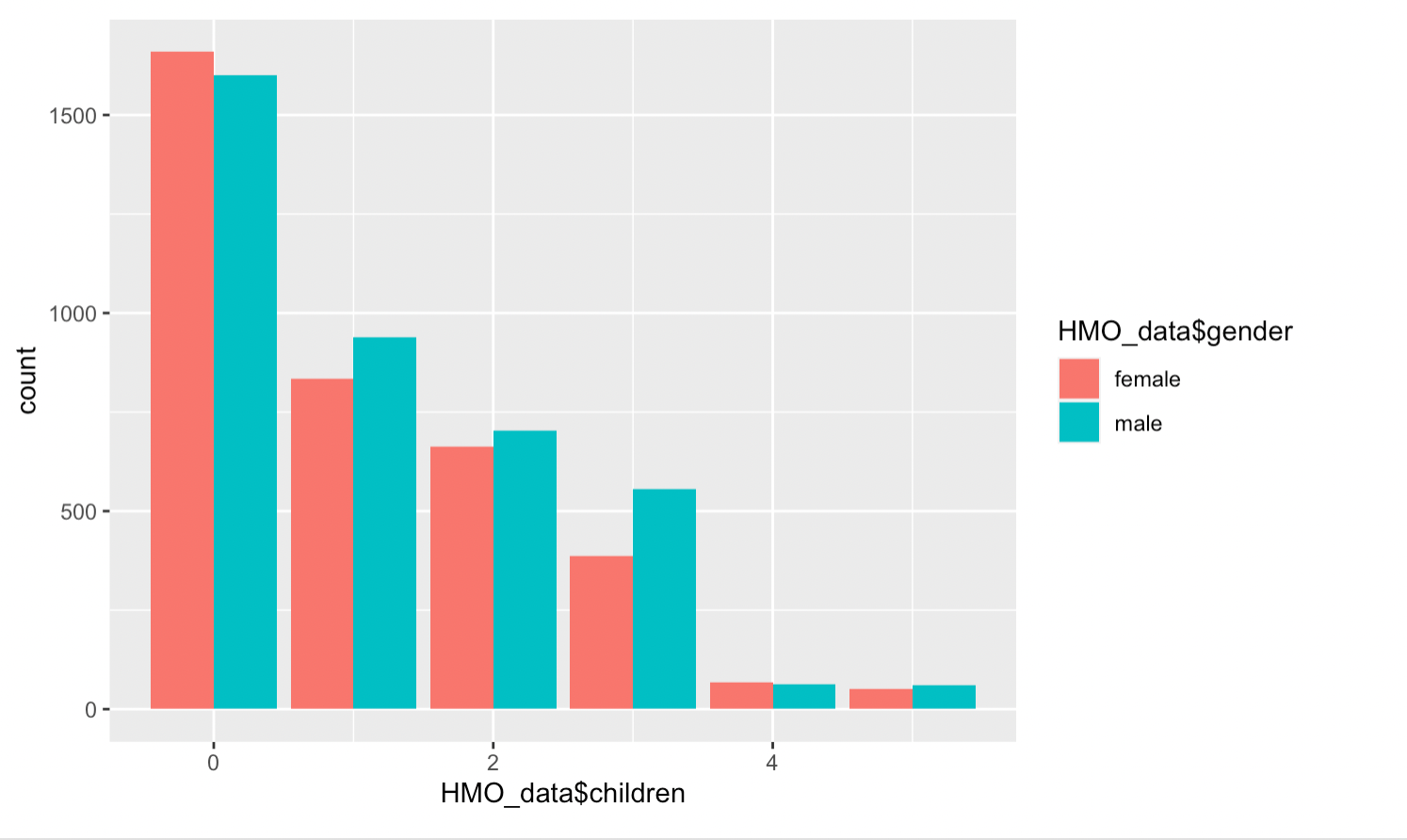
**Plot 10: Distribution of cost among different education level** Chart

Description automatically generatedChart, histogram

Description automatically generated

**Plot 11: Distribution of BMI among different education level**

**BAR GRAPHS:**



**Plot 12: Distribution of BMI among different education level**

* Plot 12 indicates that there are more proportion of males and females with no children and very few males and females have children greater than equal to 4
* Plot 13 shows that there is more proportion of people who are inactive with males relatively being more active or in-active then femalesChart, bar chart

  Description automatically generated

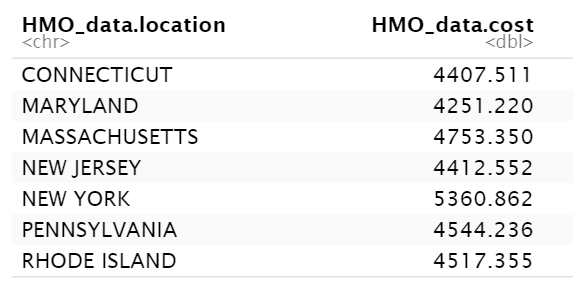
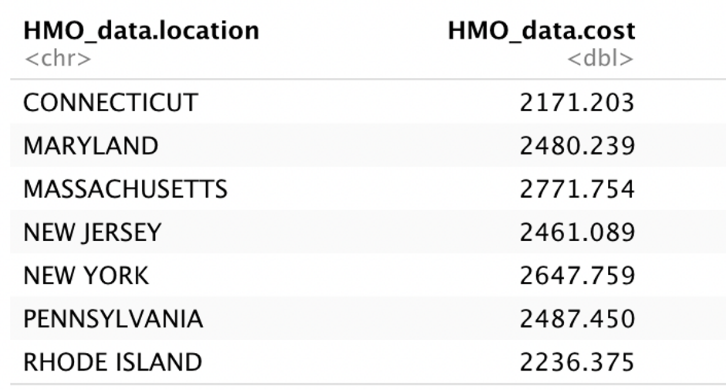
**Plot 13: Distribution of actively exercising people**

* Plot 14 shows that the sample has more proportion of non-smokers as compared to smokers with males being relatively more smokers than females

Chart, bar chart

Description automatically generated

**Plot 14: Proportion of smoker among the two genders**

**Active Non-Active**

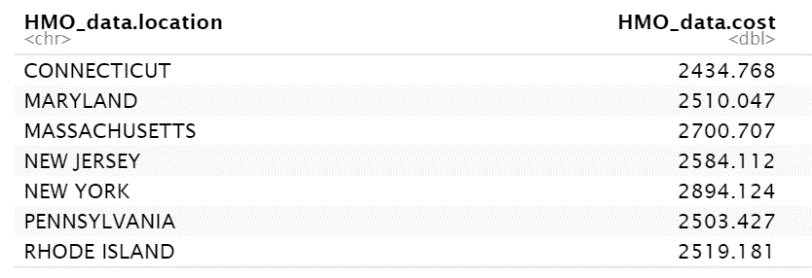
**Table 3: Average cost of people who exercise vs people who are non-active in the seven states**

* Table 3 shows that among people who are exercising, Massachusetts has the costliest patients while Connecticut has less costly patients
* While among people who are not exercising, New York has the most costly patients while Maryland has less costly patients
* Table 4 shows that Rhode Island has minimum mean age of 38 and Massachusetts has maximum mean age of 40
* We can derive from the figure that the mean cost is minimum for Maryland with 3784.174 and the maximum is 4661.506 for New York
* Table 5 shows that average cost of smokers is higher than for non-smokers, with state of New York has the most cost, while Maryland and Connecticut has less cost

Table

Description automatically generated

**Table 4: average cost and age across seven states**

**Smokers Non-Smokers**

**Table 5: average cost for smokers and non-smokers**

**BOXPLOTS:**Chart, diagram, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedDiagram

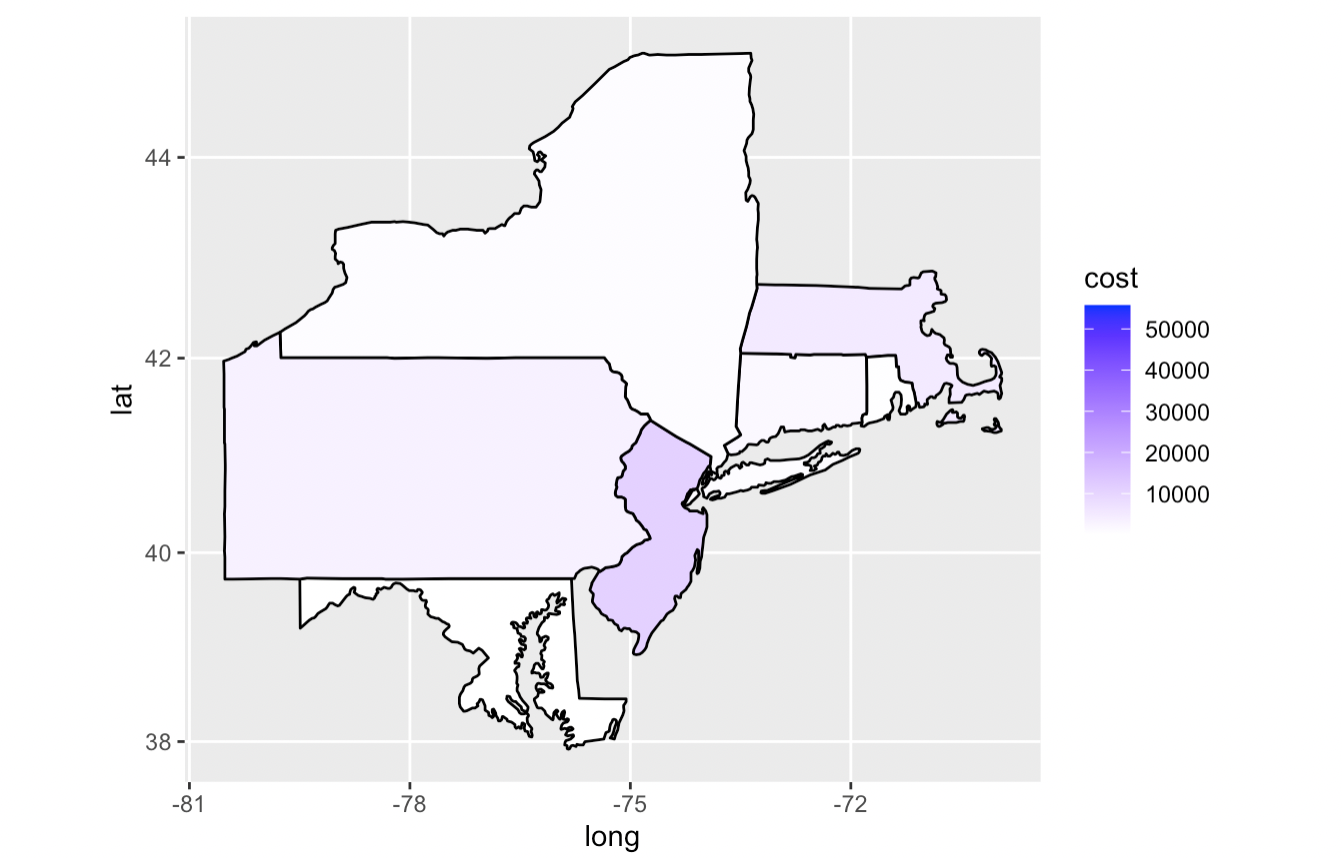
Description automatically generated with low confidenceChart, box and whisker chart

Description automatically generated

**Plot 16: Boxplots of different variables against cost**

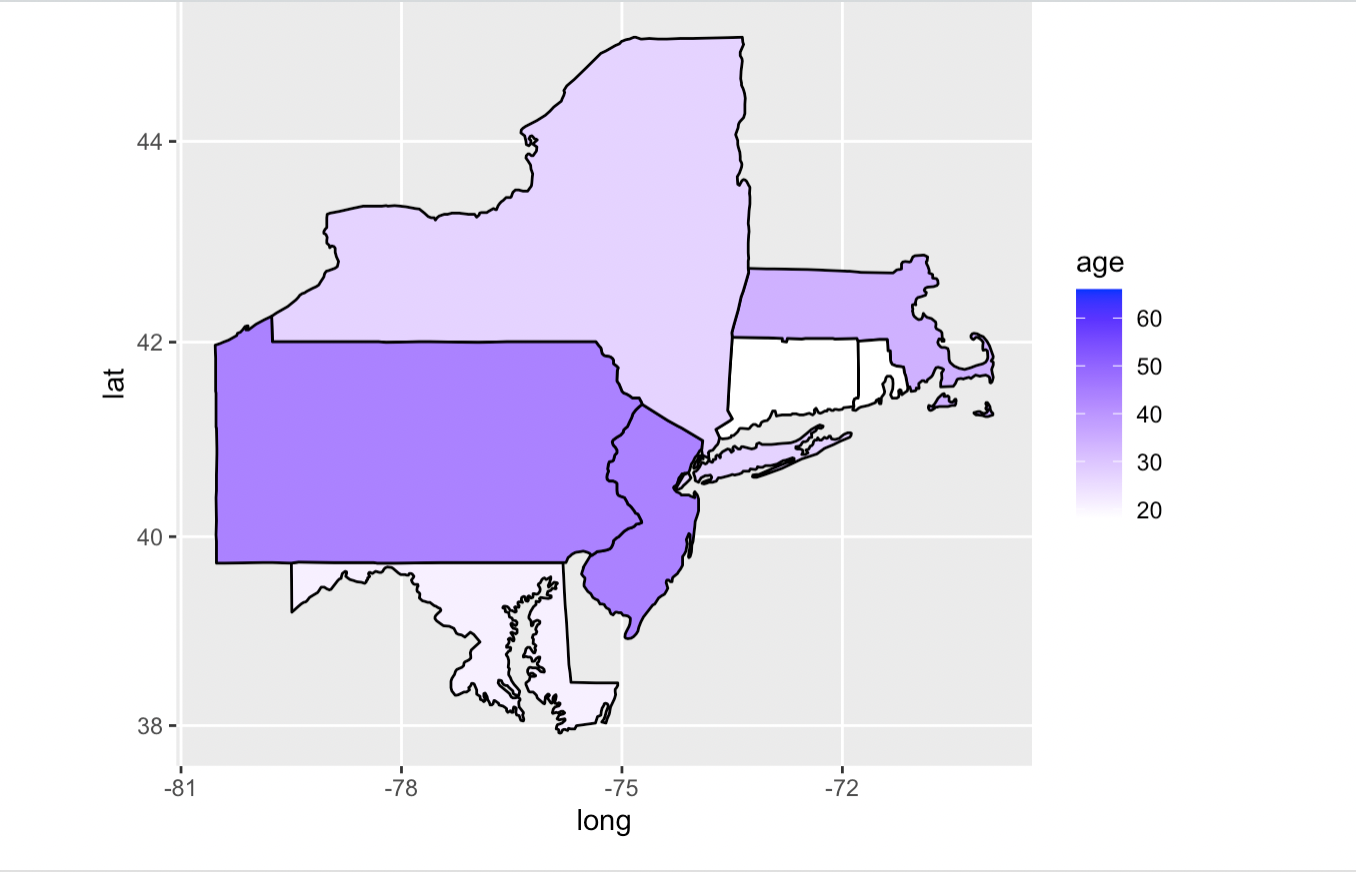
The box plot above relationship between gender, smoker, location and location\_type with cost. It seems there are many outliers present with higher cost range but for the majority, the cost lies below 5000.

**Maps:**



**Map 1**

* Map 1 tells us about the aggregate data cost for each location type.
* The regions with low cost are covered as white and with high cost are covered as blue.



**Map 2**

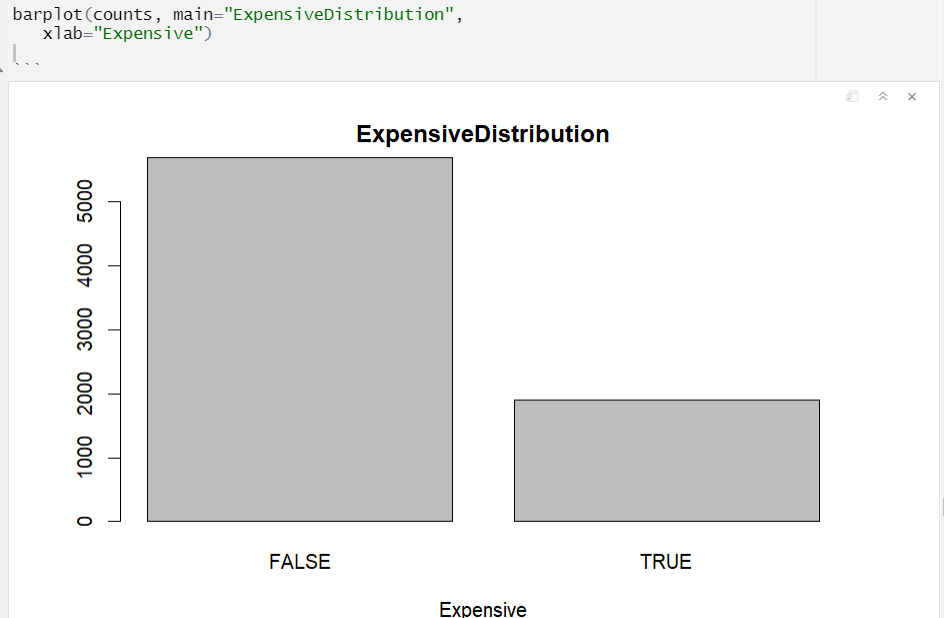
* This map tells us about the age distribution in the seven states
* The regions with low age are covered as white and with high cost are covered as shades of blue. Hence, Pennsylvania has much older sample population while Rhode Island has relatively younger population

# **EDA with expensive column**

We conducted a exploratory data analysis by creating an expensive column with a threshold value of $4775

This value is calculated from analyzing the cost column of our dataset, where 4775 is the 75th percentile value of the cost column. So the people above this threshold were labelled ‘expensive’.

To get a basic understanding of whether people are expensive or not, we created a barplot that depicts distribution of values TRUE FALSE indicating if the person is expensive or not based on our threshold.

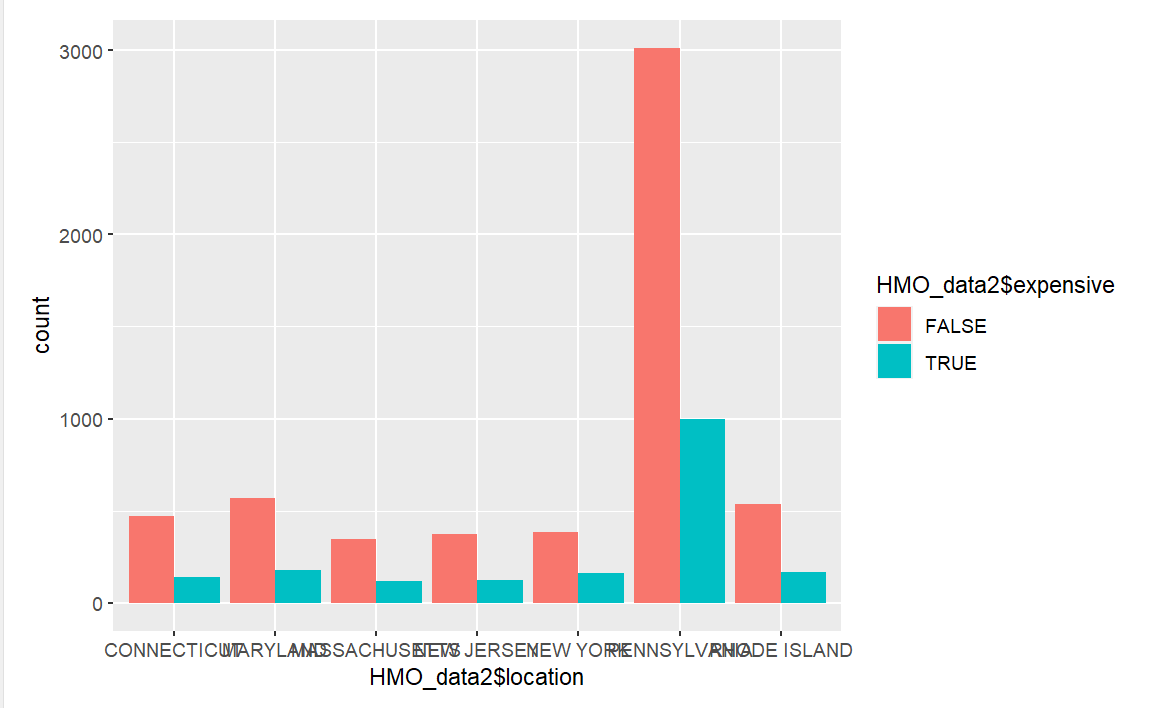


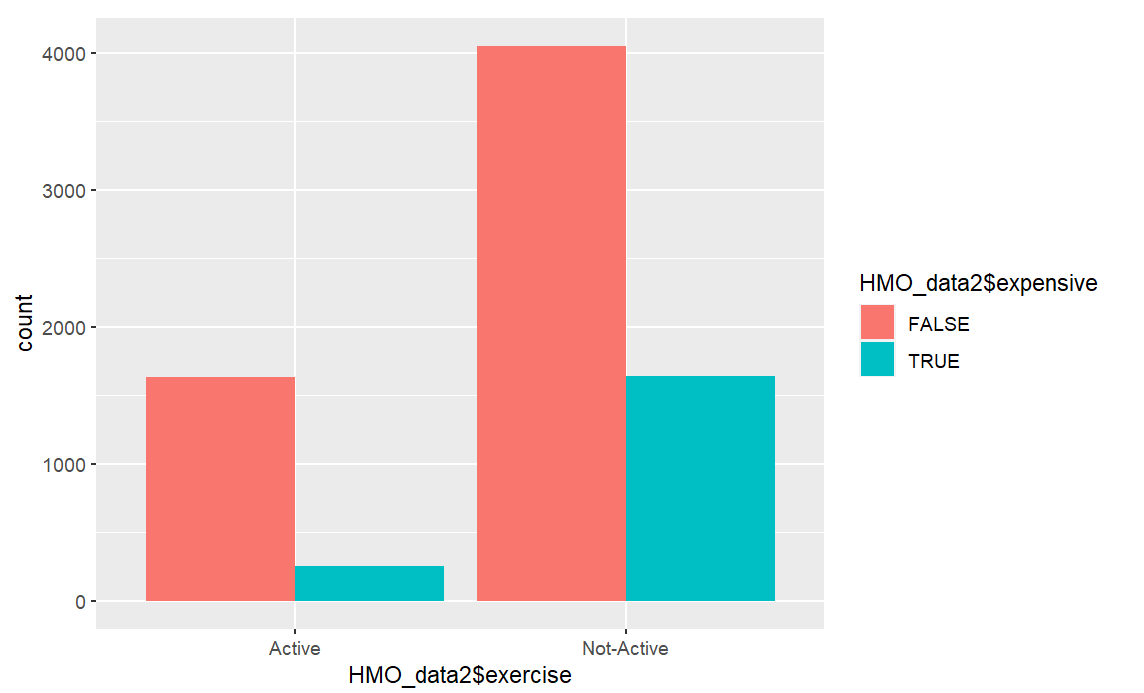
**Plot 19: Bar graph showing count of expensive (True) and not-expensive (False)**

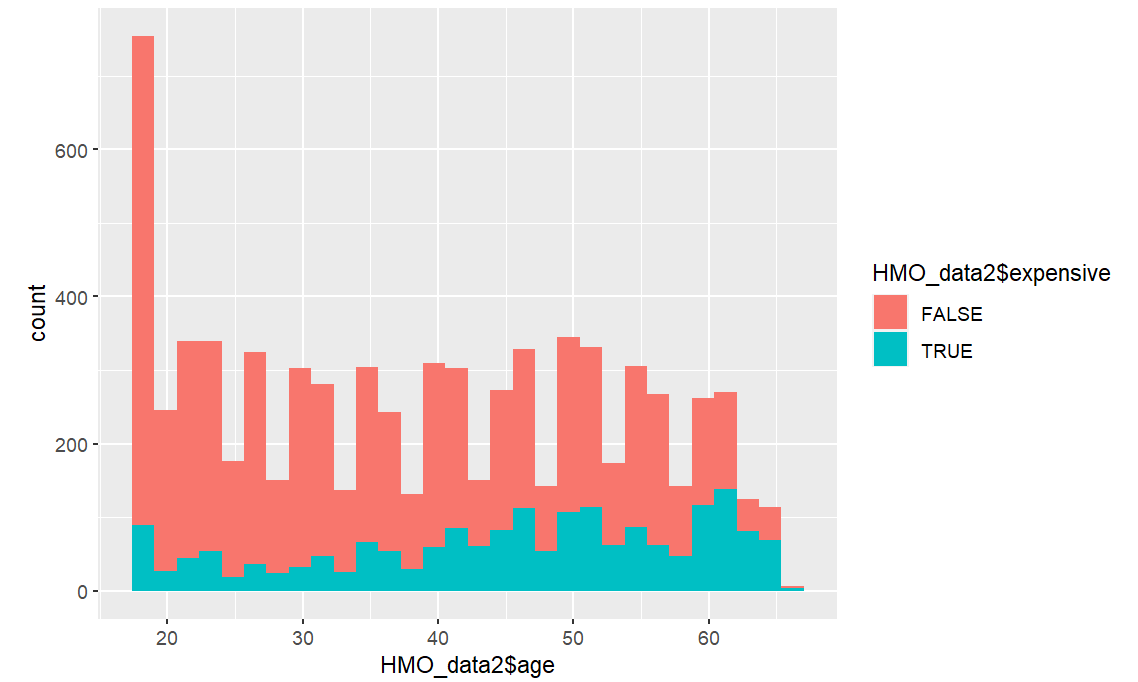
It looks like there’s very few people who are expensive, which makes sense as the threshold value was 75th percentile of the cost column.

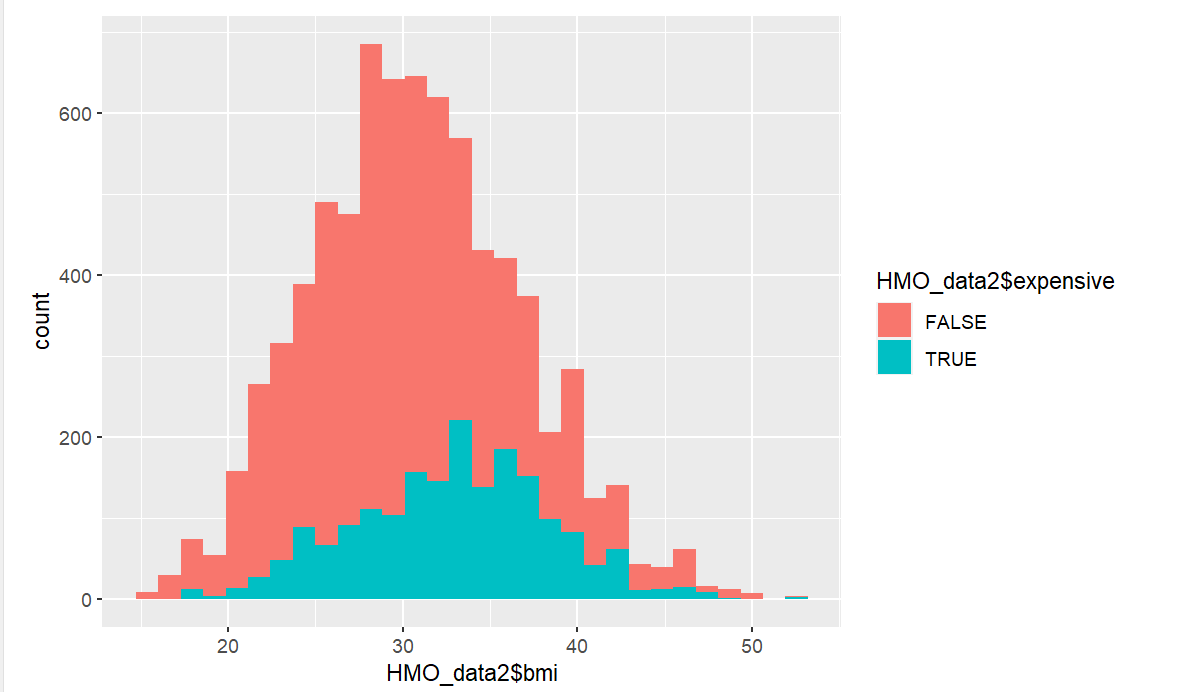
We then tried to find out if there’s any relationship between people being expensive with:

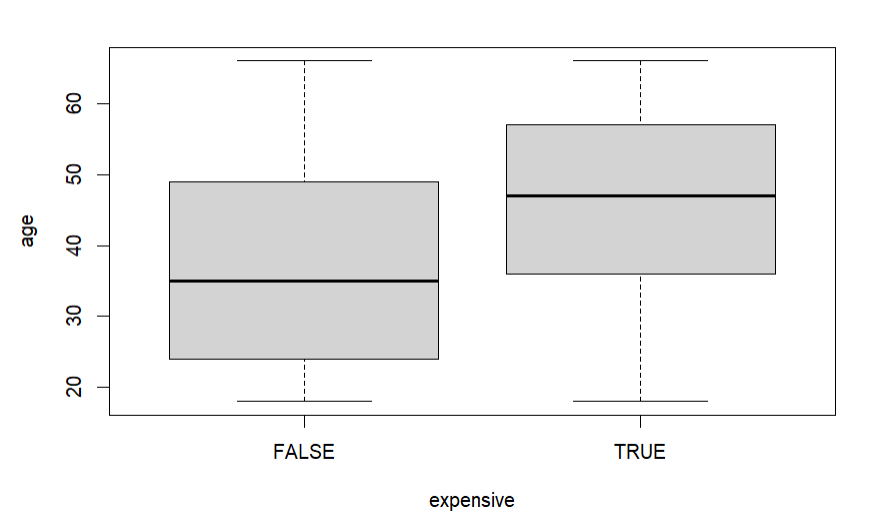
1. Number of children they have
2. Their educational level
3. If they are smoker or not
4. Their location
5. Whether they exercise or not
6. Their marital status
7. BMI index
8. Age

We did this with help of the ggplot library and visualization tools provided in R. Some of the graphs that were helpful were barplot, histogram and also boxplot.







**Plot 20: Bar graphs and histograms showing insights into relationship between expensive column and other variables**

**Plot 22: Box plot showing expensive distribution for age**

With help of these visualizations, it is evident that people who are expensive are/have:

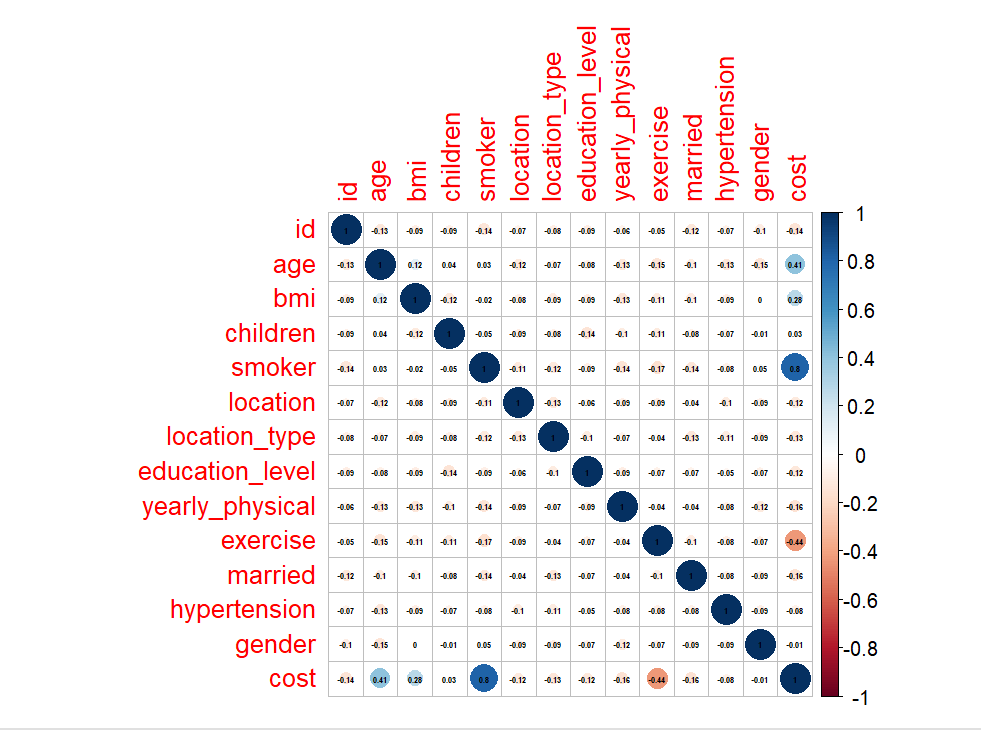
1. Older age in general as compared to less expensive people
2. Higher BMI as compared to less expensive folks
3. Exercise less

Pennsylvania seems to have higher proportion of expensive people, probably due to higher presence of them in the sample.

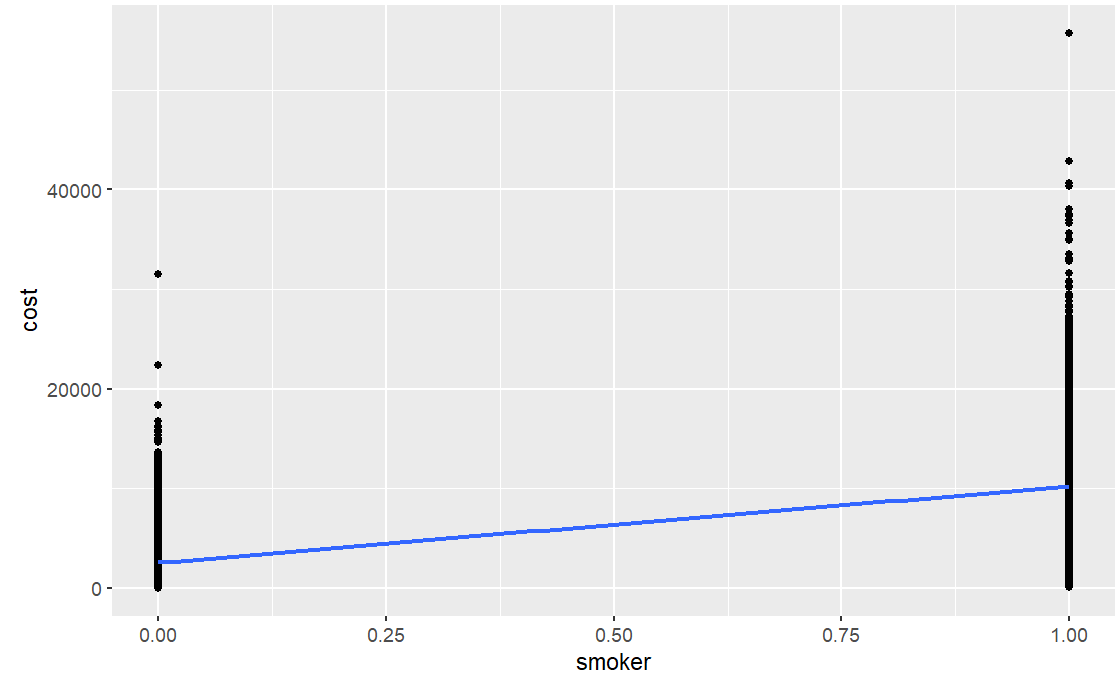
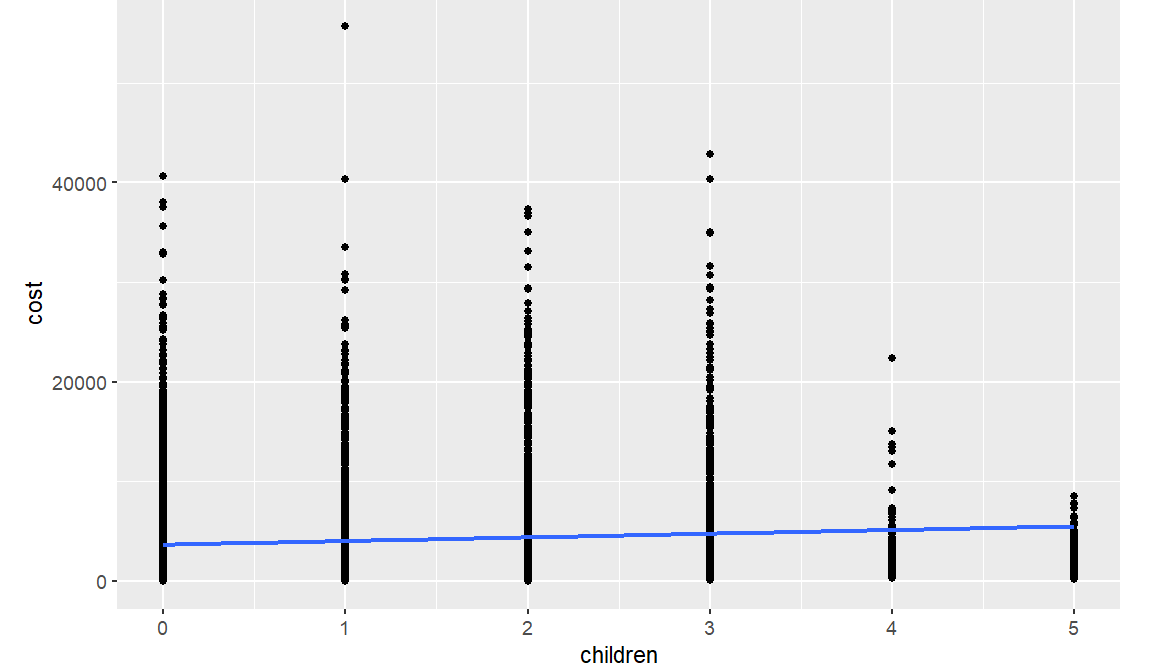
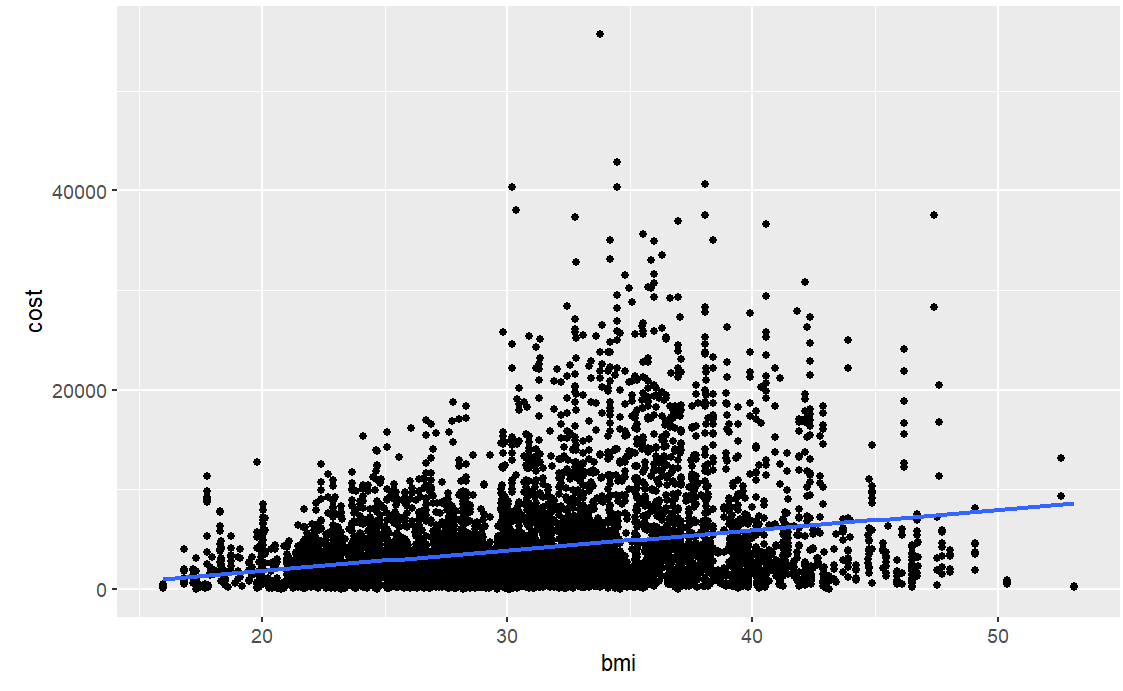
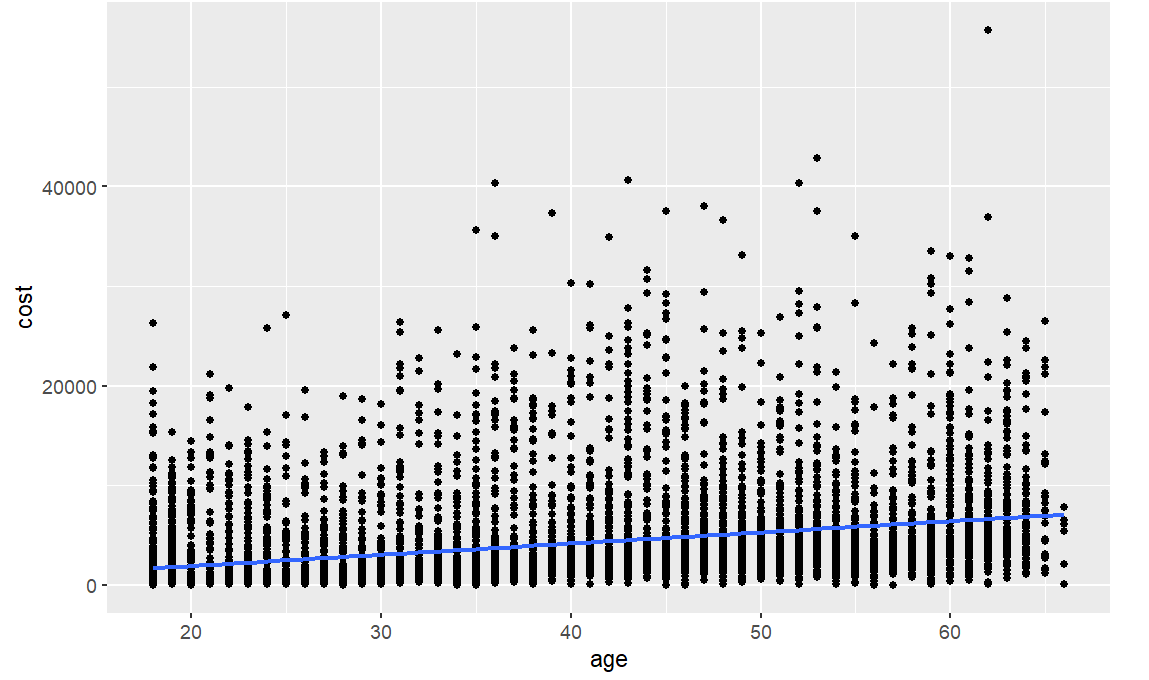
# **4. Linear Modelling**

Linear regression creates a predictive model showing trends in data. It is used to show a relation between two variables: a dependent (Y) and an independent (X) variable. The independent variable is known as the predictor variable and the dependent variable is known as the outcome or response variable. After scaling the HM0\_data we took predictor variables age, bmi, children, smoker, location, location type, education level, yearly physical, exercise, married, hypertension, and gender to predict the cost.

Before linear modelling we also created a correlation matrix to see the relationship between different variable:

**Plot 23: Correlation matrix**

The correlation matrix shows that cost has a decent to strong relationship with age and smoker variable. Additionally, we created some regression scatter plots to visualize the relationship between cost and some of the numerical variables:

Plot 24: Regression scatter plot between cost and other numerical variables 

The plots in plot 24 shows that the cost increases as age and bmi increases and if the person is smoker. The number of children does not seem to have much effect on the cost.

Now to linear regression:

* We have tried applying regression here to derive cost based on certain predictors. We have tried implementing individual parameters such as bmi, age, and smoker.
* For individual features, we see that the adjusted r-squared value is the highest for smokers (0.3813).
* We have tried implementing multiple regression using the above predictors and we see there’s a better relationship between these variables to determine cost because it works by considering the values of the available multiple independent variables and predicting the value of one dependent variable.
* There is a very minor difference between the r-squared values for model lmOut8 and lmOut9, but both seem strong. We can see that we don't need all the variables to sum up the data to 57%, so even if we have 6 variables (bmi, age, smoker, children, exercise, and hypertension) with an r-squared value of 0.57, i.e., 57%, it makes the same generalization, making it a better model than others to determine the value of cost. Based on this predicted result we can upscale the value to get the actual cost and compare it with the threshold set (i.e., 75th percentile) to eventually predict whether healthcare for an individual is expensive or not.

Table 6: Regression Matrix

## **4.1 Linear Regression with scaled data:**

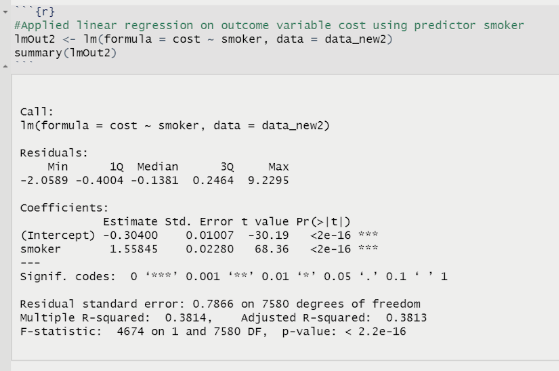


Table 7: cost ~ smoker

Table 8: cost ~ bmi + age + smoker

Table 9: cost with all the variablesGraphical user interface, text, application, email

Description automatically generatedText

Description automatically generated

Table 10: Cost with only 6 variablesGraphical user interface, text, application, email

Description automatically generated

lmOut9 gives us the best adjusted R-squared out of all the tested linear regression models and will be used to predict the cost.

Graphical user interface

Description automatically generated with low confidence

Table 11: Using predict() command

It can be seen that the predicted values from the lmOut7 model is coming in scaled format, which is not easily interpretable. Hence, in the next section we predict the cost on unscaled data to get more interpretable result.

## **4.2 Linear Regression with unscaled but encoded Test HMO\_data**

After performing linear regression on scaled data, we performed linear regression model lmOut9 on unscaled, endoded data set HMO\_test data so that our predicted values can come out in interpretable manner.

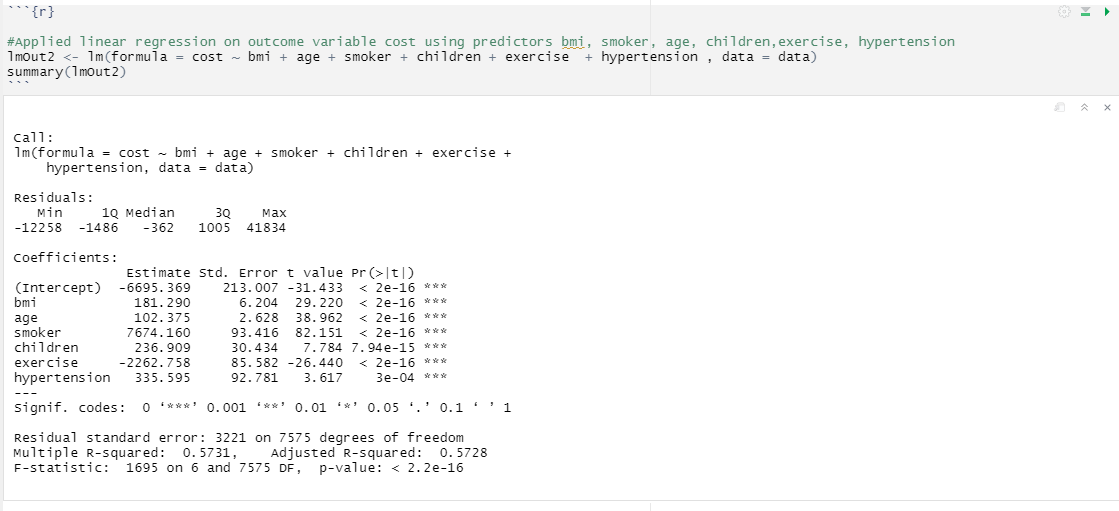


Table 12: cost ~ With all 6 variables

Using predict()

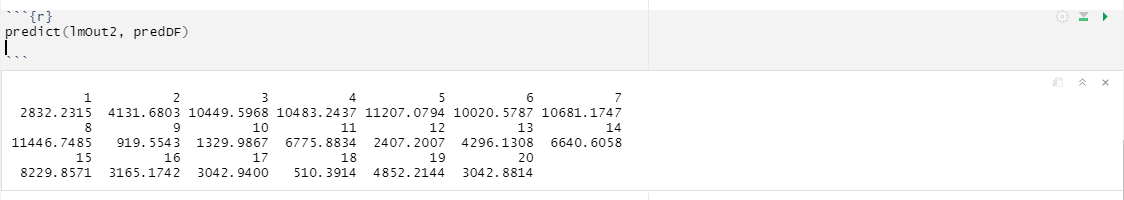


Table 13: predicted cost result

# **5. Supervised Learning**

To predict whether the given patient will be expensive or not with the given attributes, we took the supervised learning approach and built several SVM models to make the model learn how to differentiate between expensive and not expensive people.

SVM models used labelled information to learn how to separate points. In our case, we had to label ‘expensive’ and ‘not-‘expensive’ people beforehand, by putting the condition that if the cost is above 4773, that person is expensive. We took 4773 as threshold because, it represents the third quartile, i.e. 75% of the people have cost below this value. Now, we have the new column named ‘expensive’ in our dataset, that labels people as ‘FALSE’ – not expensive and ‘TRUE’- expensive.

Before training the data, the data was imported from original source, treated for missing values through interpolation and then categorical values were converted into factors and integers since SVM methods are distance-based algorithms and can only process numerical values to model them.

## **5.1 TRAIN-TEST-SPLIT**

After preparing the data, a training list was created that divides the data into 60:40 ratio for training and testing dataset respectively. Then a train and a test dataset were created from the training list. The dimensions of train set came out to be 4550 observations with 13 variables while the test set was 3032 observations with 13 variables.

## **5.2 MODEL TRAINING**

Since the requirement of the client was to find the model that labels people with best sensitivity, several different SVM models were tried to find the best suited model for the case.

## **5.2.1 KSVM**

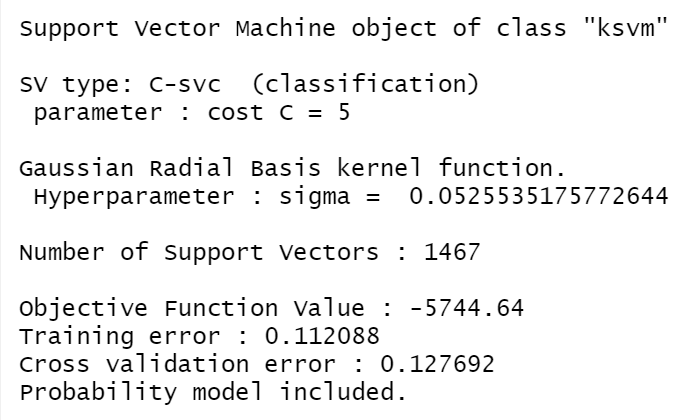
The first model implemented was ksvm. ‘Kernlab’ was loaded, and seed was set at ‘111’ so that we get the same result on running the model again. ‘Expensive’ was made the dependent variable, dependent on all the independent variables present in the dataset. ‘C’ of 5 was chosen with a cross-validation of 5, where the probabilty of predictions are stored in a matrix (prob.model=TRUE). 

Table 14: Result of SVM

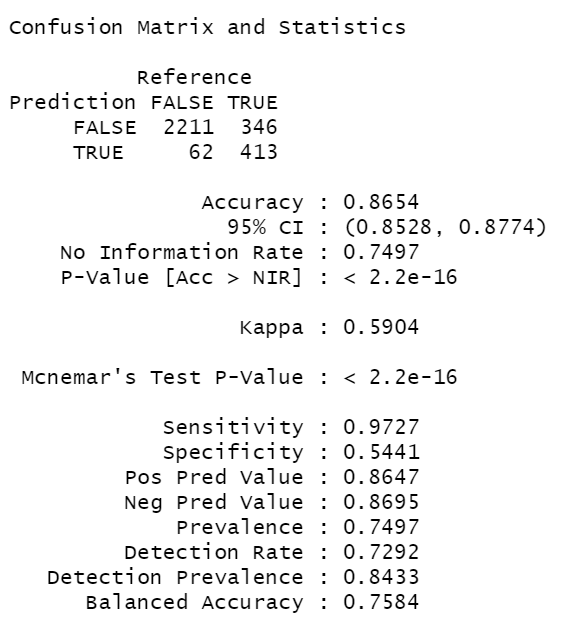
This model was tested on test set and was assessed with the help of confusion matrix, the result of which is as follows: 

Table 15: Confusion matrix for ksvm

While the accuracy if the model is 86.54% the sensitivity is 97.27%

## **5.2.2 Radial SVM**

Second method used to conduct SVM modelling was using ‘svmRadial’ method. Here, again ‘expensive’ was made dependent on all the independent variables, with method= svmRadial, and no trControl argument since we using simple partitioning approach. We pre-process the data, standardizing it to have mean of zero and standard deviation of 1, by putting in the argument ‘center’ and ‘scale’.

This model was tested on test data and the result of confusion matrix is as follows:

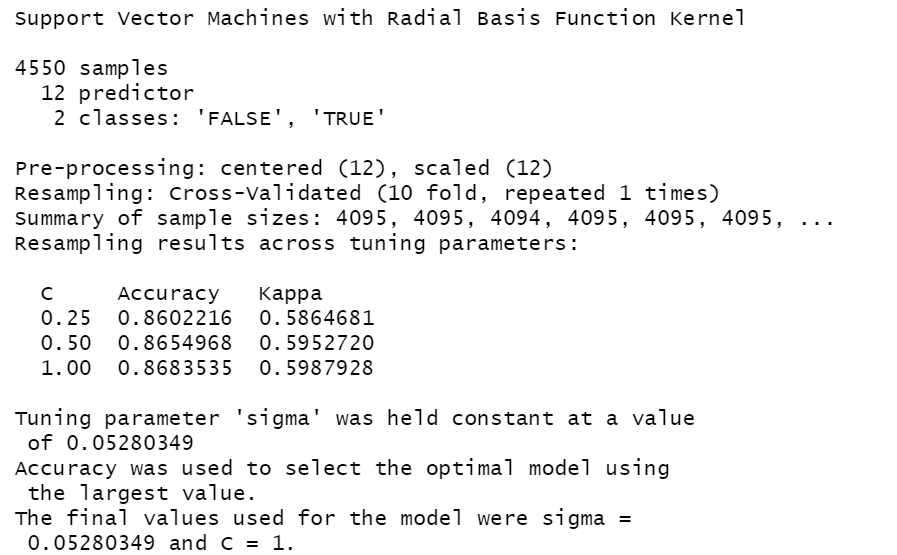


Table 16: Confusion matrix for Radial\_svm

While the accuracy if the model is 84.70% the sensitivity is 95.96%. The sensitivity is lower than ksvm model.

## **5.2.3 K-fold Validation**

Third method used to conduct svm was using k-fold validation. This method uses radial svm but also splits the data ‘k-times’ so that the model will run ‘k’ times. Repetition help in avoiding the problem of overfitting specially of large dataset. So, we thought of implementing the k-fold validation, with the repetition value of 10, no controls and pre-processing the data to normalise it. The dependent variable ‘expensive’ was dependent on the 13 independent variables. The accuracy on train data came out to be 86.83%

Table 17: result of k-fold validation with all variables

We tried the same k-fold model, but this time with 6 independent variable i.e.

expensive ~ bmi + age + smoker + children + exercise + hypertension

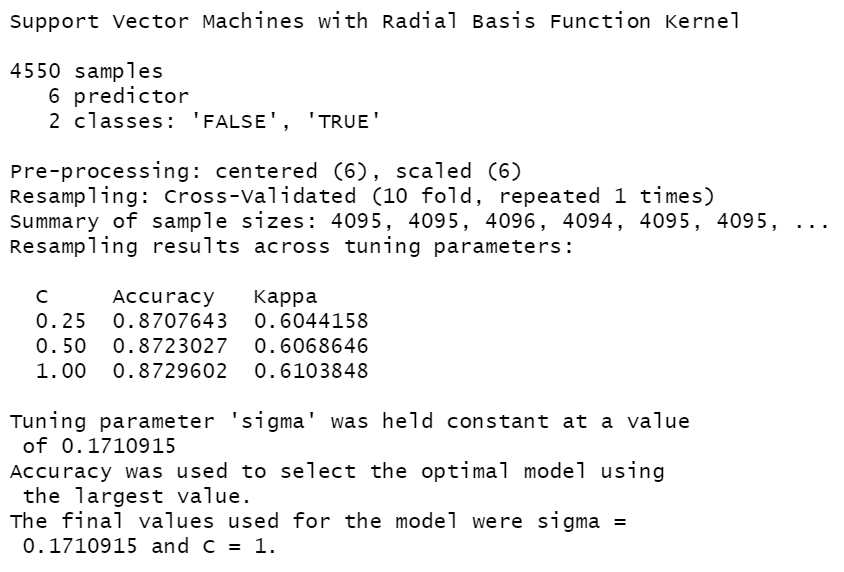
we chose these variables because these 6 variables seem to have strong relationship with cost in linear regression modelling, so it might be the case that they help in better prediction of expensive/not-expensive people. 

Table 18: Result of k-fold validation with 6 variables

The accuracy of training model came out to be 87.29%, which is better than the first k-fold model.

Implementing this model on test data, we get the following confusion matrix:

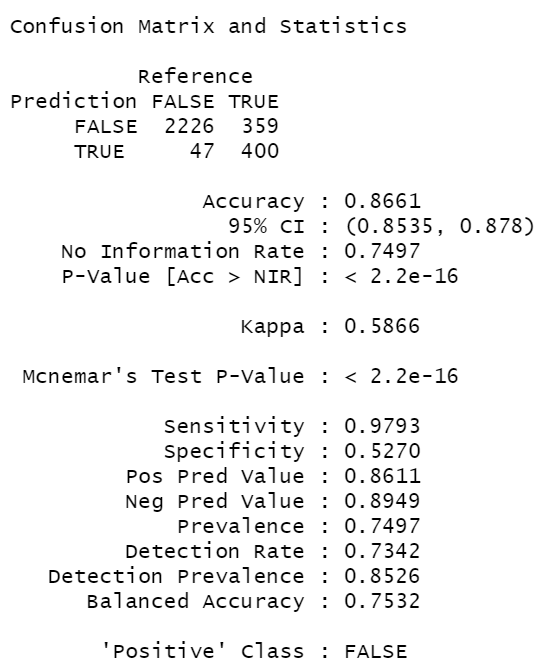


Table 19: Confusion matrix for k-fold validation with 6 variables

The accuracy came out to be 86.61% with sensitivity of 97.93%

| Model | Accuracy | Sensitivity |
| --- | --- | --- |
| SVM | 86.54% | 97.27% |
| Radial SVM | 84.70% | 95.96% |
| K-fold validation with Radial SVM | 86.61% | 97.93% |

Table 20: Comparative model performance table

On comparing all the models, k-fold SVM is giving us the best result for sensitivity.

This model was also tested on the small test data HMO\_test\_sample and the confusiton matrix result is as follows:

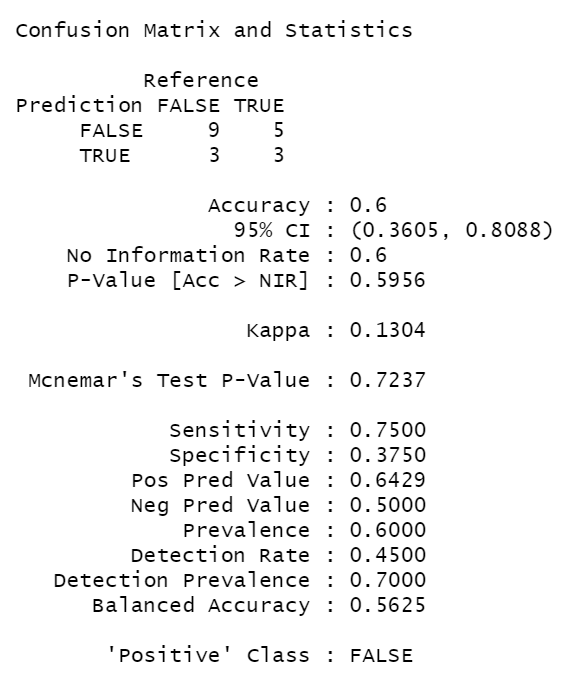


Table 20: Confusion matrix for k-fold validation with 6 variables on test data

# **6. Unsupervised Learning**

We tried to implement k-means clustering as an unsupervised learning technique to test out how well it can identify and label expensive and non-expensive people.

## **6.1 Data cleaning**

We used cleaned dataset from previous step and changed all column types to numeric. We also turned categorical variables with more than one categorical value into inidivudual numbers. Categorical variables with exactly two values were turned into 0 and 1.

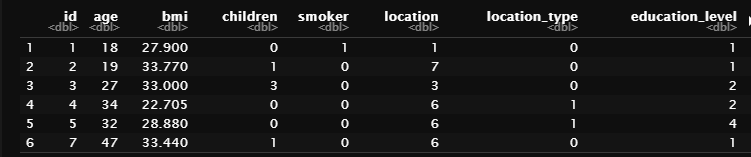


Table 21: Dataset with new column types

## **6.2 Columns scaling**

To perform unsupervised classification, we had to scale all column variables with more than two values. Once all those variables are scaled, they have the same normal distribution, which means the algorithm can use them for column comparison and dataset clustering. As result, we scaled Age, bmi, children , location, education\_level, cost columns.

## **6.3 Picking variables for classification**

While we tested a number of variables combination to use for clustering the dataset into expensive and non-expensive. We decided to pick 4 variables with highest correlation scores from the correlation matrix (Plot 23): Age, bmi, Smoker, Exercise. The reason why we don’t include the cost is because we want to derive the average cost from the clustered data by these variables.

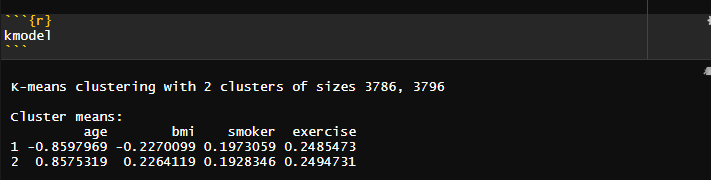
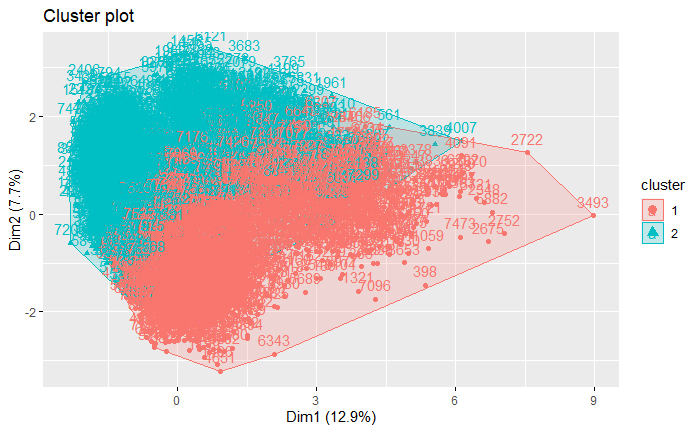


Table 22: KMeans Clustering Results

## **6.4 KMeans Clustering**

The resulting 2 clusters were 3786 and 3796 rows respectively. The distinction was pretty even with the expected overlap, since we can consider extreme cases with disabled people and people whose healthcare costs depend on natural disability. The next step is to identify which cluster is expensive and which one is non-expensive.



Plot 25: KMeans split result

## **6.5 Cluster Identification**

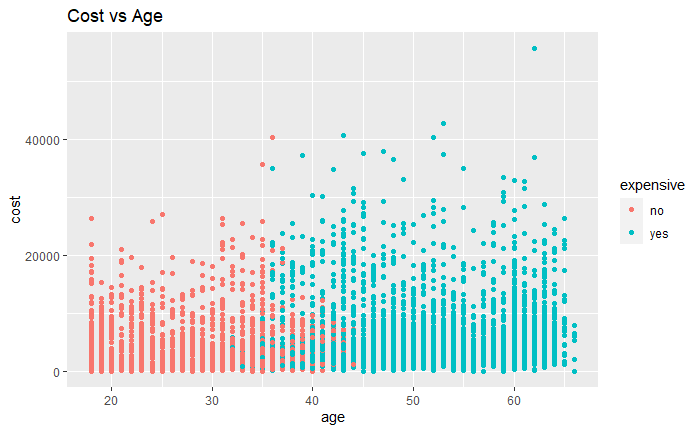
From the average values of age, bmi, and cost, we were able to infer that the cluster 1 is expensive and cluster 2 is non-expensive. The average age for expensive cluster is 51 and non-expensive is 27. The bmi are 32 and 29 respectively. The final variable is cost which also shows a big difference, more than twice. These results can prove



Table 23: Cluster average values

## **6.6 Expensive & Non-Expensive**

After plotting Cost vs Age on the updated clustered dataset, we can observe the split into expensive and non-expensive. The borderline is around 40 years old, which makes sense because after that people’s health starts to decrease due to age, so they are in a more risky group with predicted increase in healthcare costs. We believe that this clustering provides a good insight and further proves the point that the 4 variables we used are effective in deciding whether someone is expensive or not.



Plot 26: Expensive and Non-Expensive distributions