IST 687 HMO Data Analysis Report

Aruneema

Harshit Joshi

Danila Rozhevskii

Vaishnav Kanekar

Shweta Suhas Rane

Contents

1. 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.

2. 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:

```
       X                    age             bmi       
 Min.   :        1   Min.   :18.00   Min.   :15.96  
 1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60  
 Median :    24916   Median :39.00   Median :30.50  
 Mean   :   712602   Mean   :38.89   Mean   :30.80  
 3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.77  
 Max.   :131101111   Max.   :66.00   Max.   :53.13  
                                     NA's   :78     
    children         smoker            location        
 Min.   :0.000   Length:7582        Length:7582       
 1st Qu.:0.000   Class :character   Class :character  
 Median :1.000   Mode  :character   Mode  :character  
 Mean   :1.109                                        
 3rd Qu.:2.000                                        
 Max.   :5.000                                        

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




   exercise           married           hypertension   
 Length:7582        Length:7582        Min.   :0.0000  
 Class :character   Class :character   1st Qu.:0.0000  
 Mode  :character   Mode  :character   Median :0.0000  
                                       Mean   :0.2005  
                                       3rd Qu.:0.0000  
                                       Max.   :1.0000  
                                       NA's   :80      
    gender               cost      
 Length:7582        Min.   :    2  
 Class :character   1st Qu.:  970  
 Mode  :character   Median : 2500  
                    Mean   : 4043  
                    3rd Qu.: 4775  
                    Max.   :55715  
```

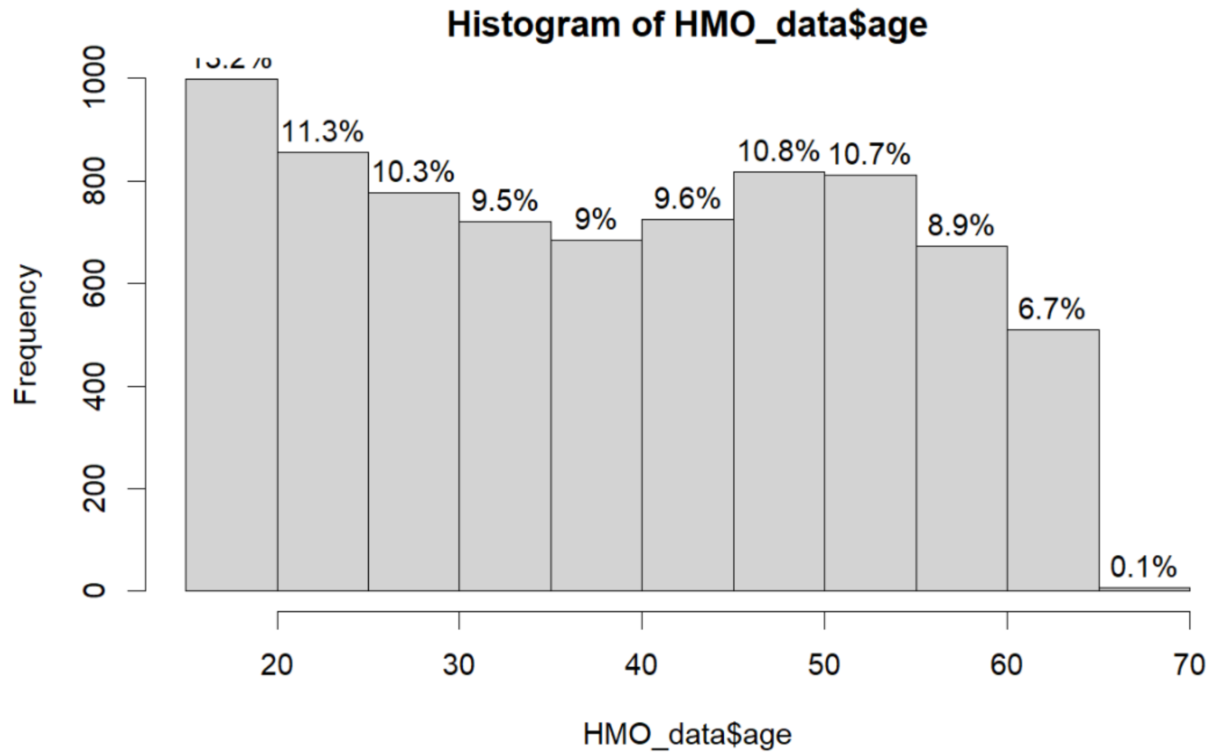
Table 2: Data summary

- 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

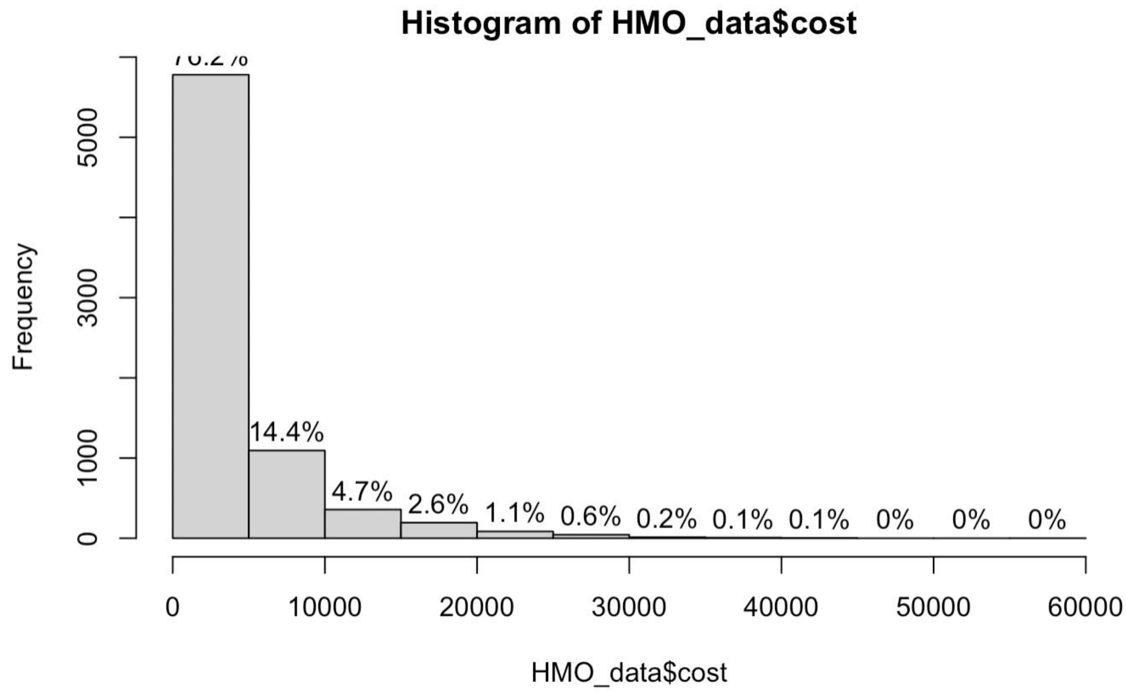
For the NA's we excute the na_interpolation function to replace the missing values with the known values.

2.1.1 Univariate analysis

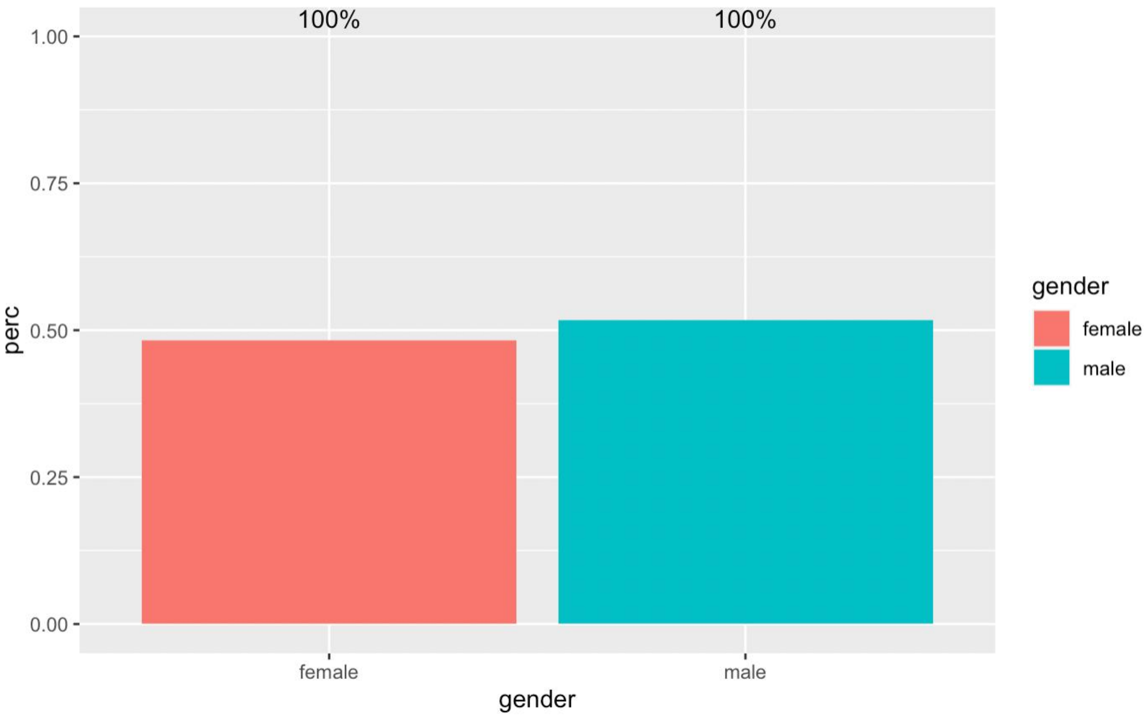
Since the data has both numeric and non-numeric variables, the univariate and bivariate analysis was done separately for these two kind of variables. Combination of bar plots and histogram was used to plot the visualisation using the ggplot function.



Plot:1 Frequency distribution of age



Plot 2: Frequency distribution of cost



Plot 3: Proportion distribution of gender

A tibble: 2 × 3

gender <chr>	n <int>	perc <dbl>
female	3662	0.482986
male	3920	0.517014

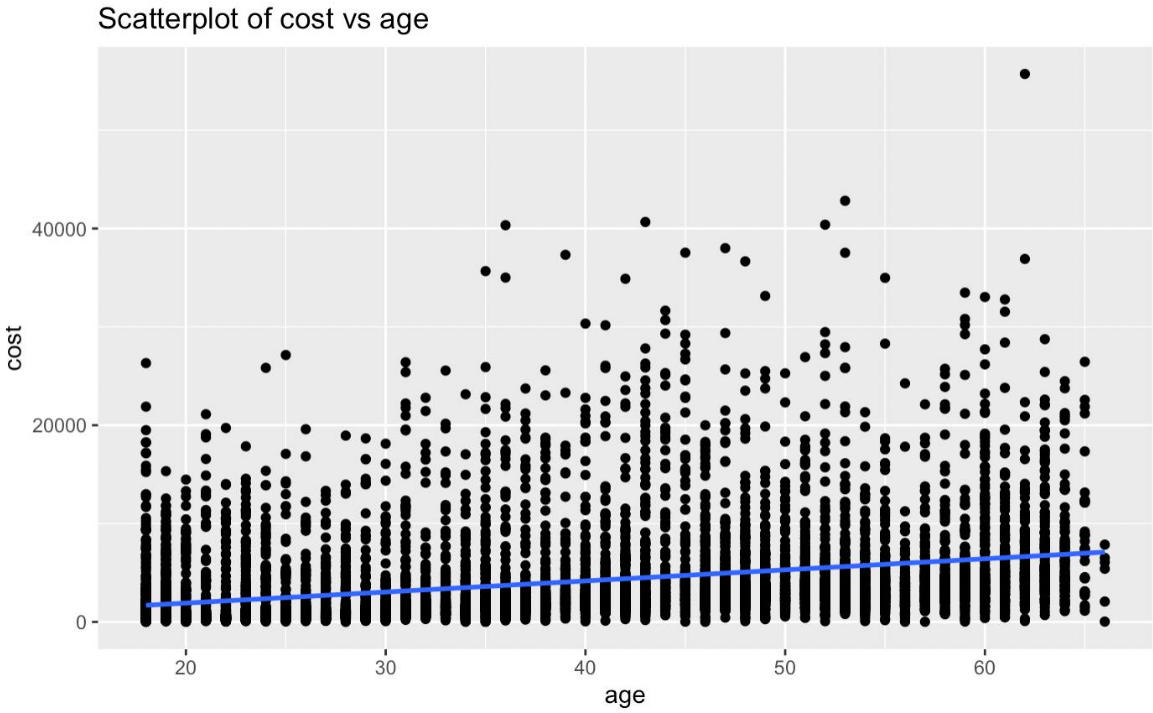
2 rows

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

2.1.3 Bivariate analysis

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

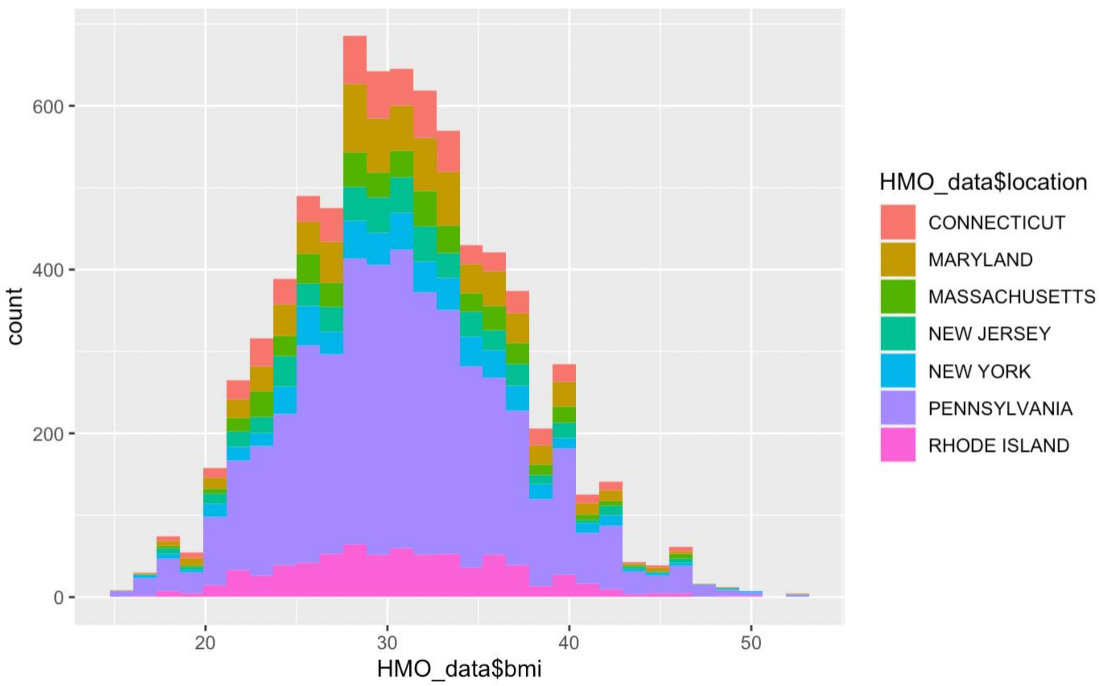


Plot 4: Scatter plot of cost vs age

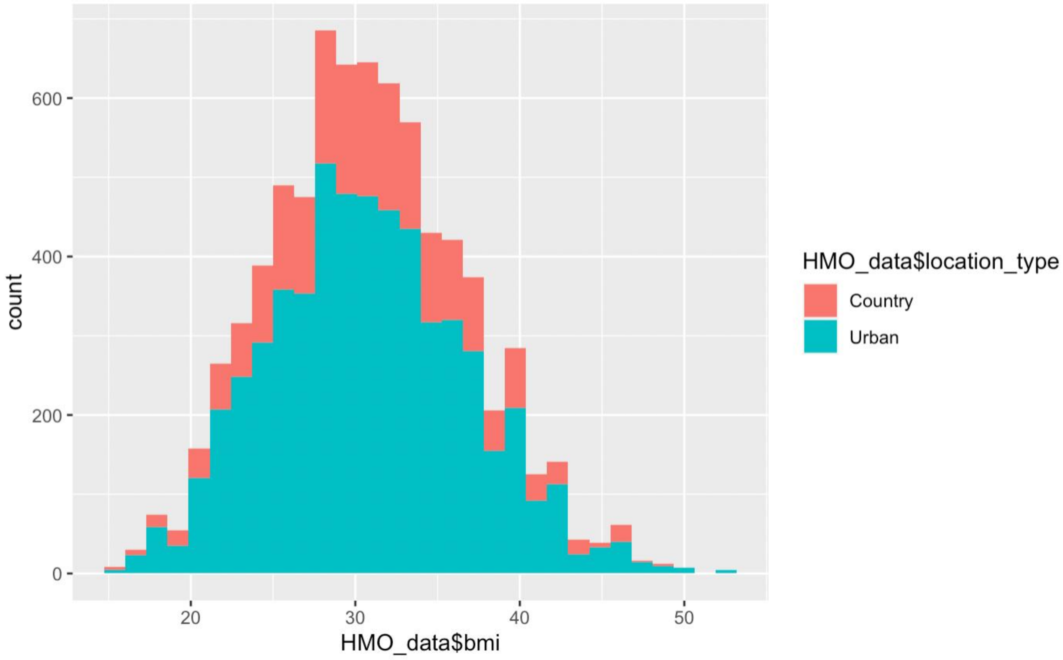
Insights:

- 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

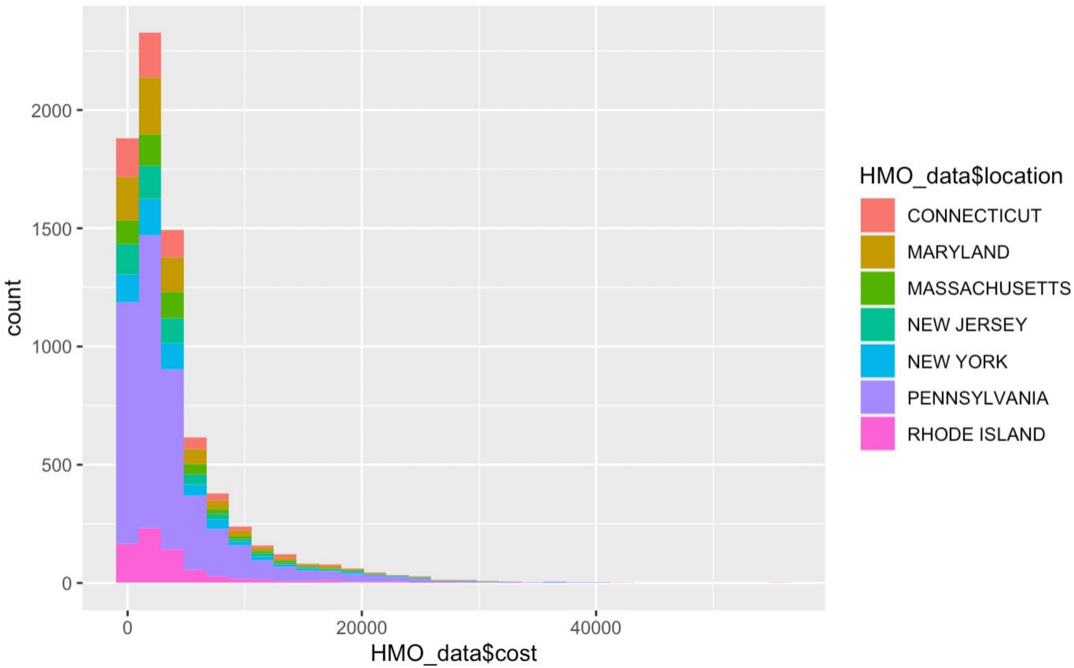
HISTOGRAMS:



Plot 5: Distribution of BMI in 7 states



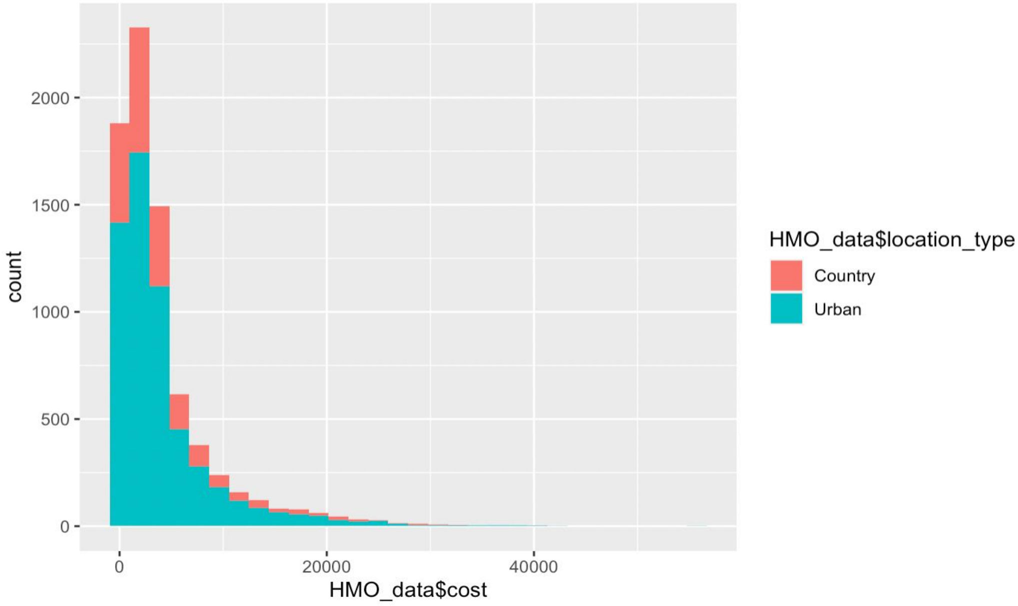
Plot 6: Distribution of BMI in Country and Urban location



Plot 7: Distribution of cost in 7 states

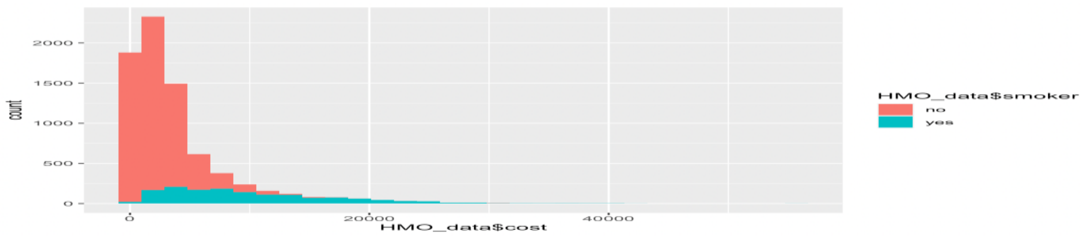
Insights:

- 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

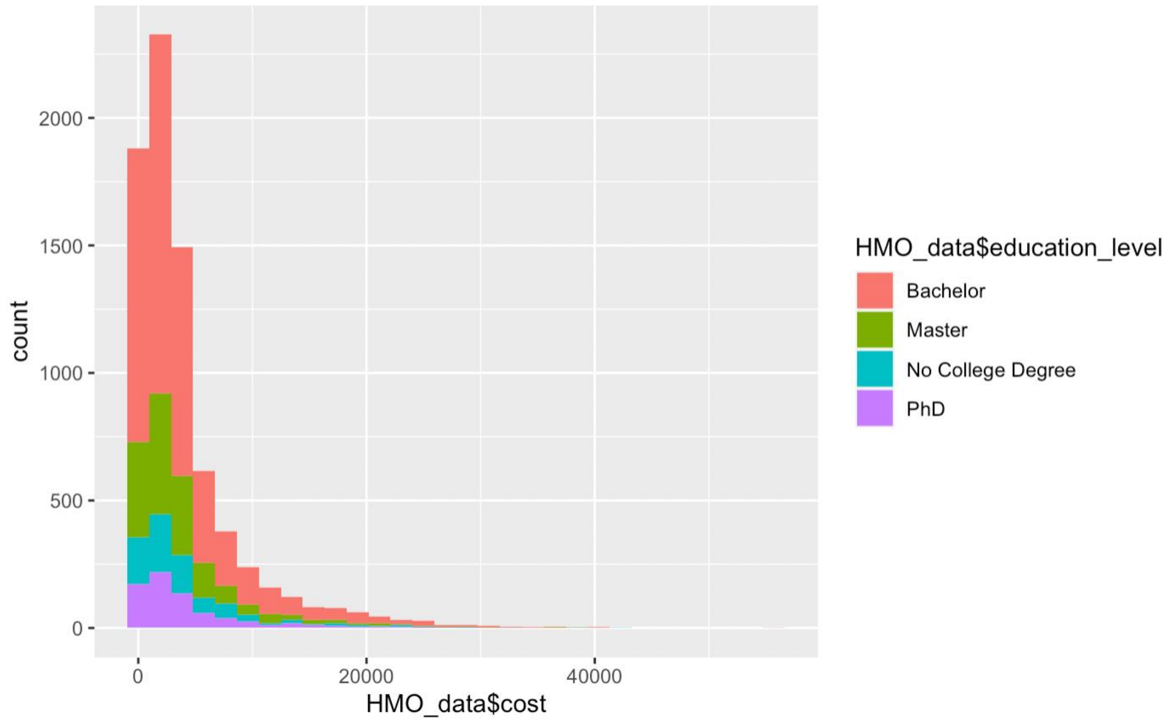


Plot 8: Distribution of cost in Country and urban region

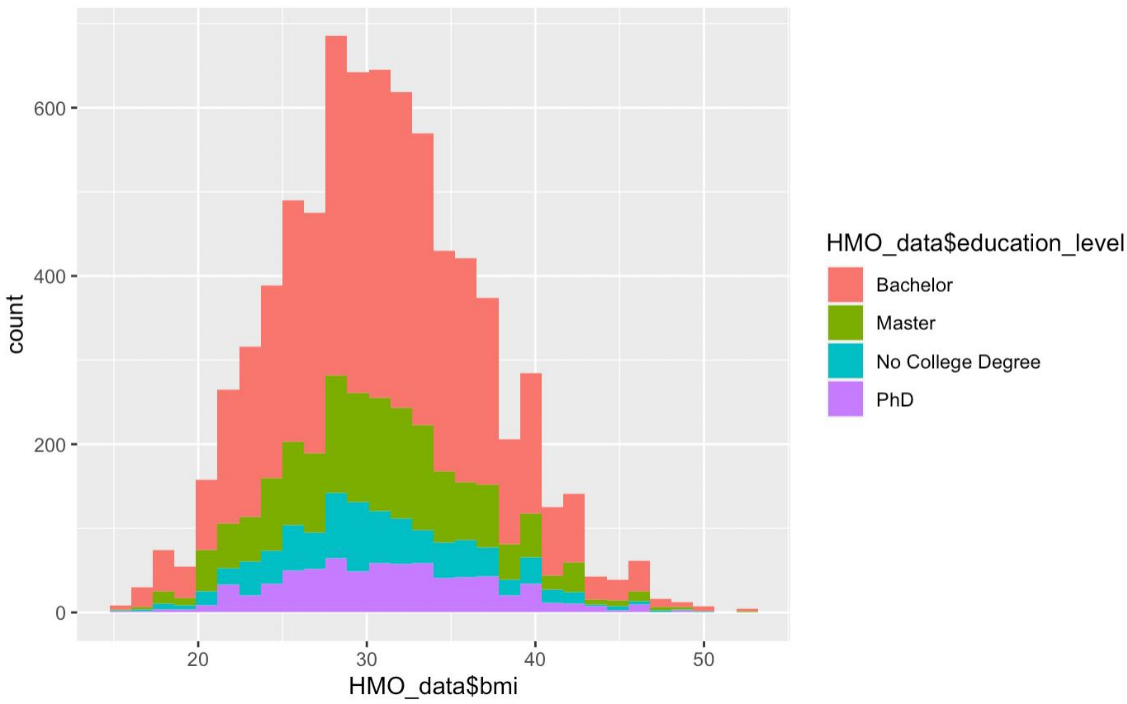
- Plot 9 shows that there are more proportion of non-smoker than smokers with the cost of both of the mostly lying below 20,000
- Plot 10 and 11 both shows that the cost of people from different educational background is skewed to the right with BMI being normally distributed. It also informs us that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



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

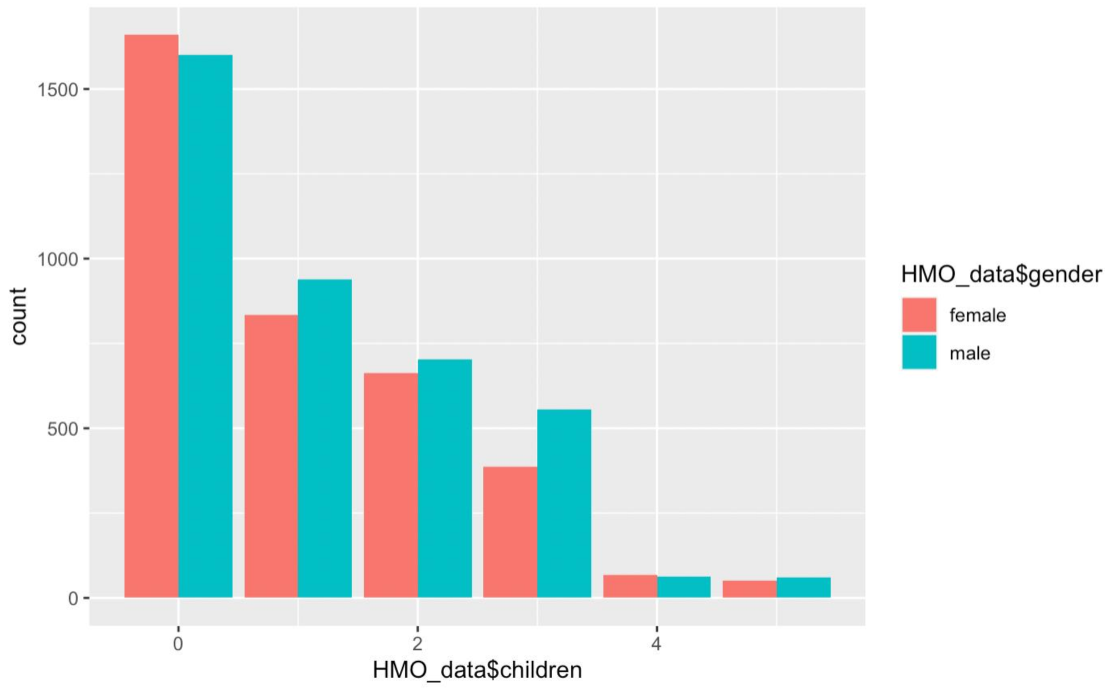


Plot 10: Distribution of cost among different education level



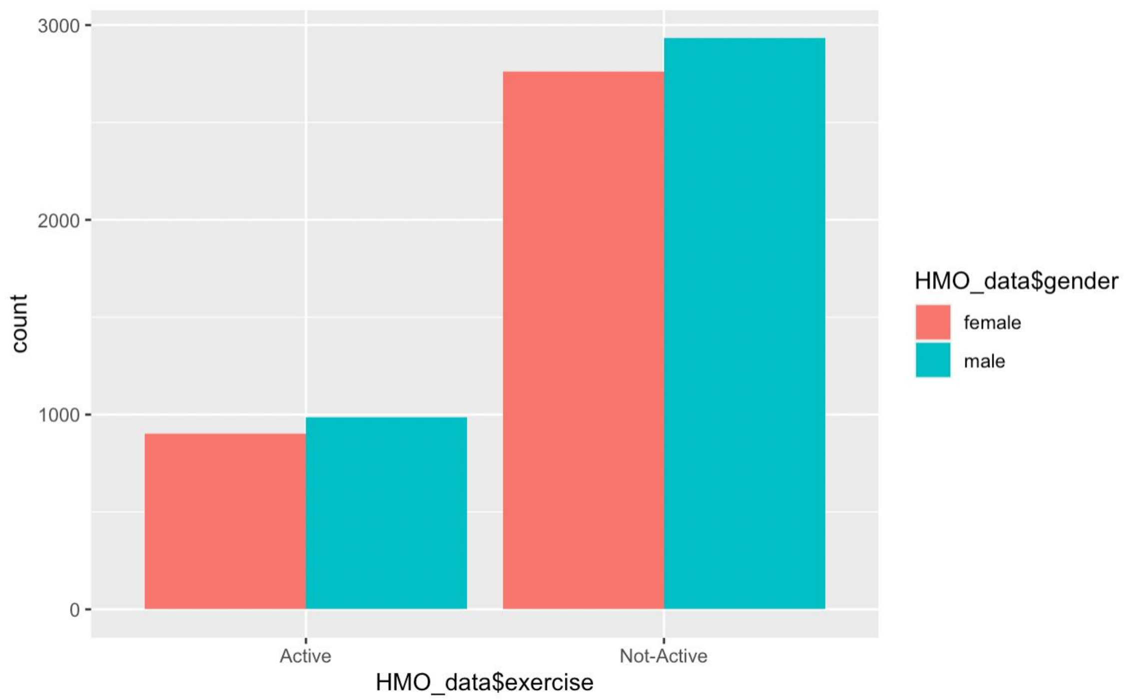
Plot 11: Distribution of BMI among different education level

BAR GRAPHS:



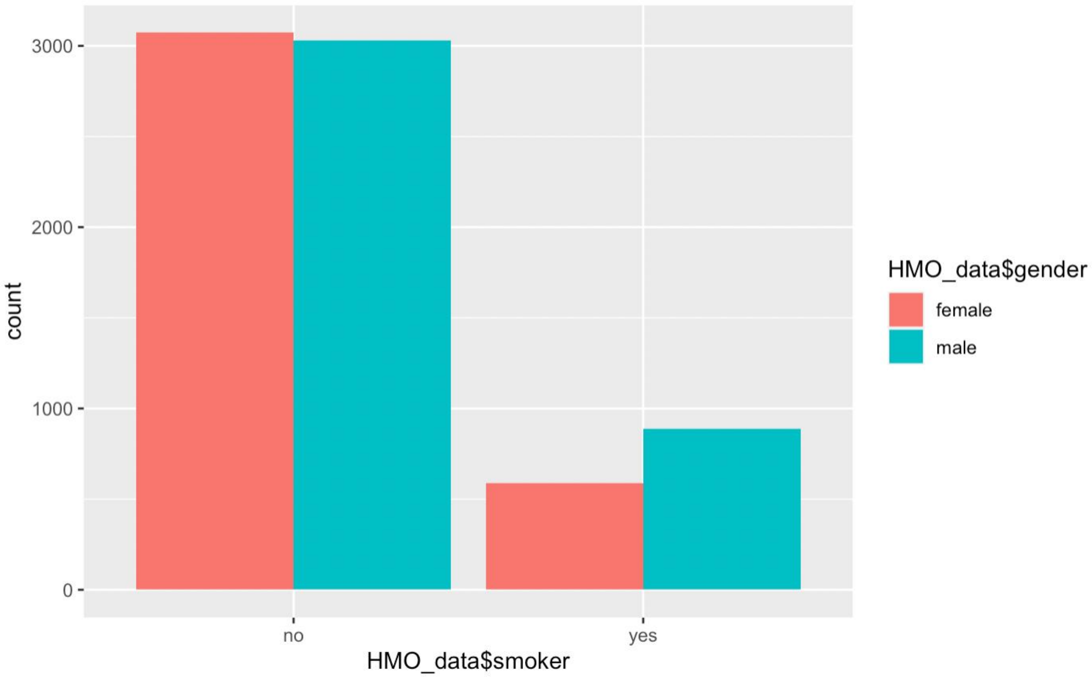
Plot 12: Distribution of BMI among different education level

- 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 females



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



Plot 14: Proportion of smoker among the two genders

Active

HMO_data.location <chr>	HMO_data.cost <dbl>
CONNECTICUT	2171.203
MARYLAND	2480.239
MASSACHUSETTS	2771.754
NEW JERSEY	2461.089
NEW YORK	2647.759
PENNSYLVANIA	2487.450
RHODE ISLAND	2236.375

Non-Active

HMO_data.location <chr>	HMO_data.cost <dbl>
CONNECTICUT	4407.511
MARYLAND	4251.220
MASSACHUSETTS	4753.350
NEW JERSEY	4412.552
NEW YORK	5360.862
PENNSYLVANIA	4544.236
RHODE ISLAND	4517.355

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

A tibble: 7 × 3

HMO_data$location <chr>	mean_age <dbl>	mean_cost <dbl>
CONNECTICUT	38.76268	3847.519
MARYLAND	38.46586	3784.174
MASSACHUSETTS	40.57204	4267.540
NEW JERSEY	38.65863	3930.564
NEW YORK	38.95795	4661.506
PENNSYLVANIA	38.89800	4023.115
RHODE ISLAND	38.35795	4050.791

7 rows

Table 4: average cost and age across seven states

Smokers

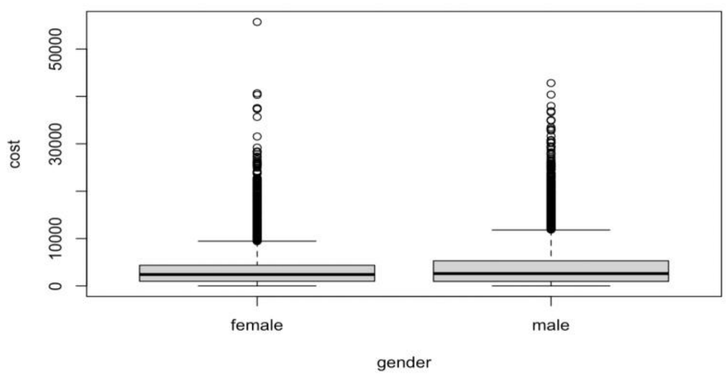
HMO_data.location <chr>	HMO_data.cost <dbl>
CONNECTICUT	10141.830
MARYLAND	8984.694
MASSACHUSETTS	10290.052
NEW JERSEY	10118.191
NEW YORK	10950.442
PENNSYLVANIA	10246.691
RHODE ISLAND	10943.039

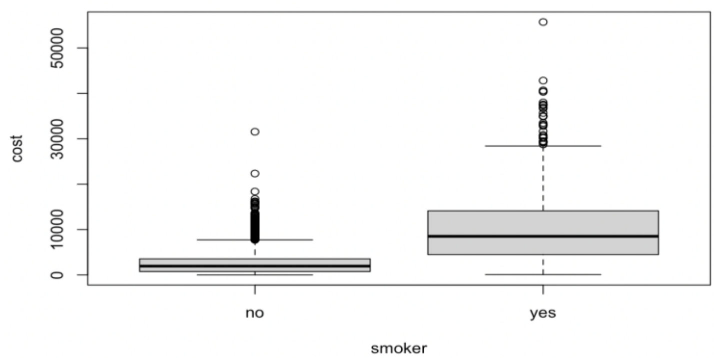
Non-Smokers

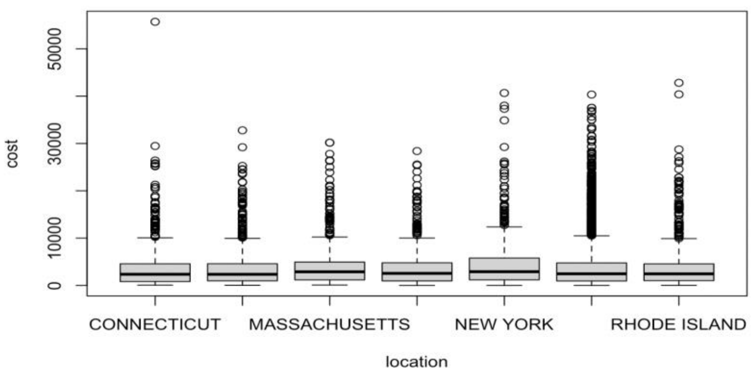
HMO_data.location <chr>	HMO_data.cost <dbl>
CONNECTICUT	2434.768
MARYLAND	2510.047
MASSACHUSETTS	2700.707
NEW JERSEY	2584.112
NEW YORK	2894.124
PENNSYLVANIA	2503.427
RHODE ISLAND	2519.181

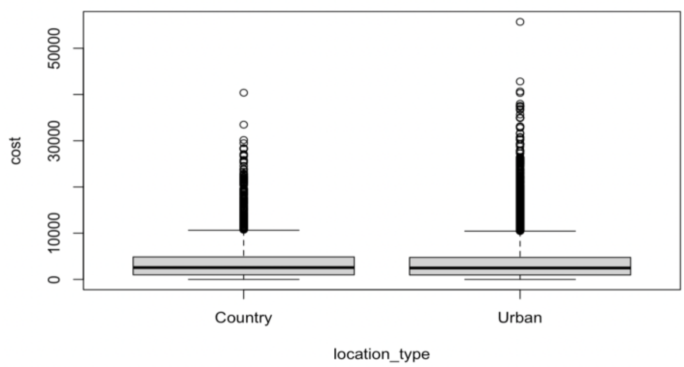
Table 5: average cost for smokers and non-smokers

BOXPLOTS:





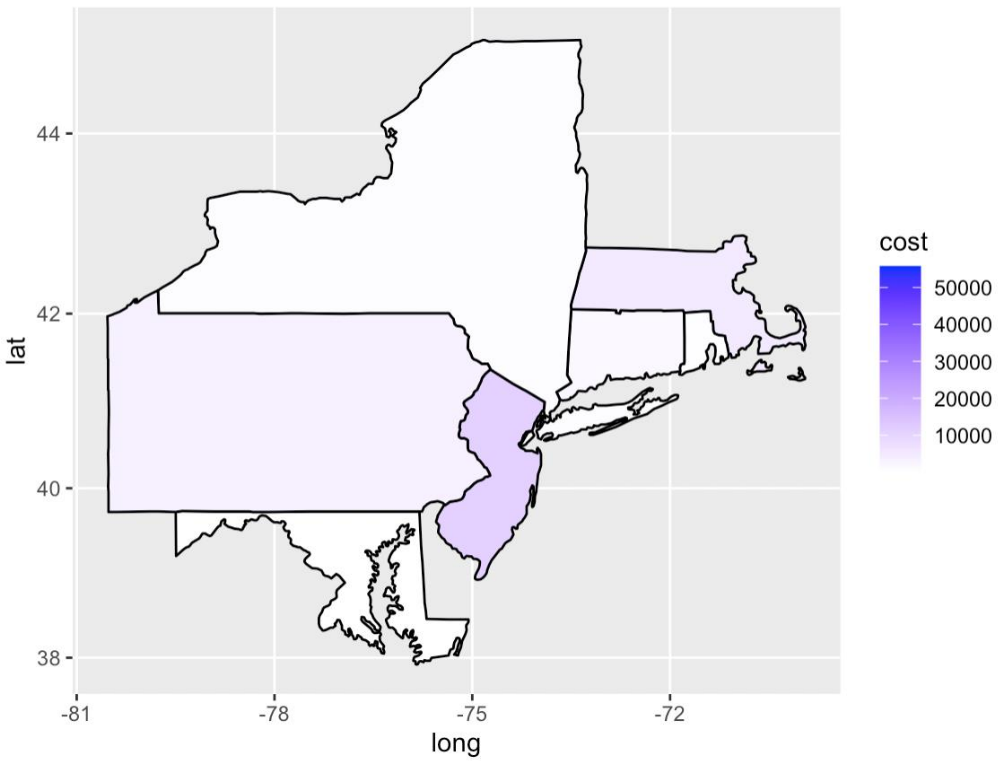




Plot 16: Boxplots of different variables against cost

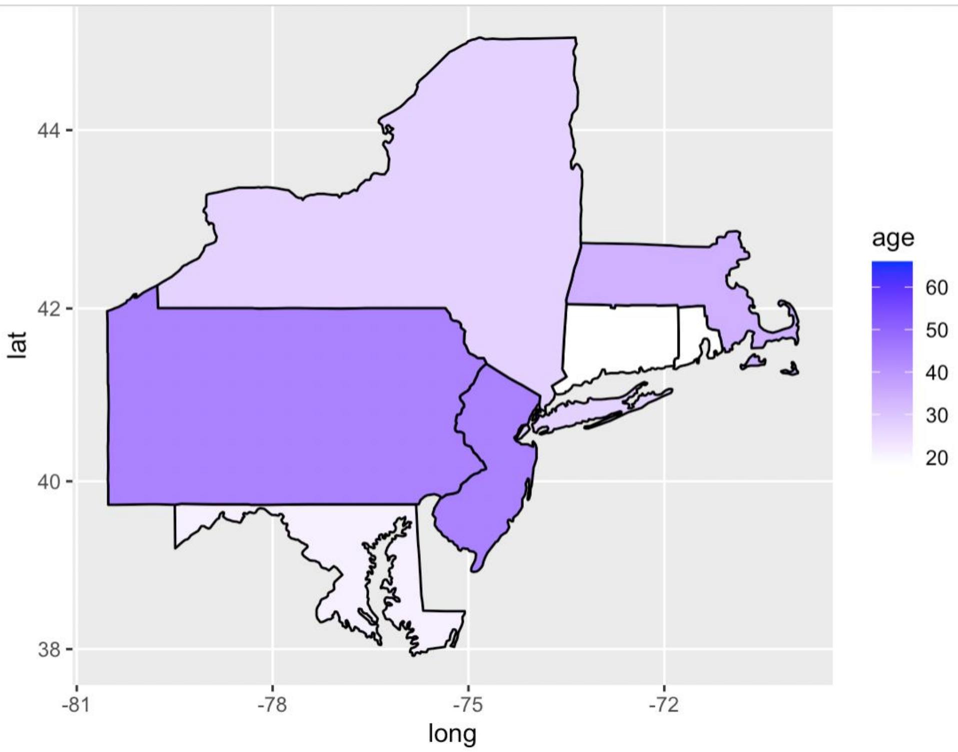
The box plot above relationship between gender, smoker, location and location_type with cost. It seems there are many outliers present with higher cost range but for the majority, the cost lies below 5000.

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

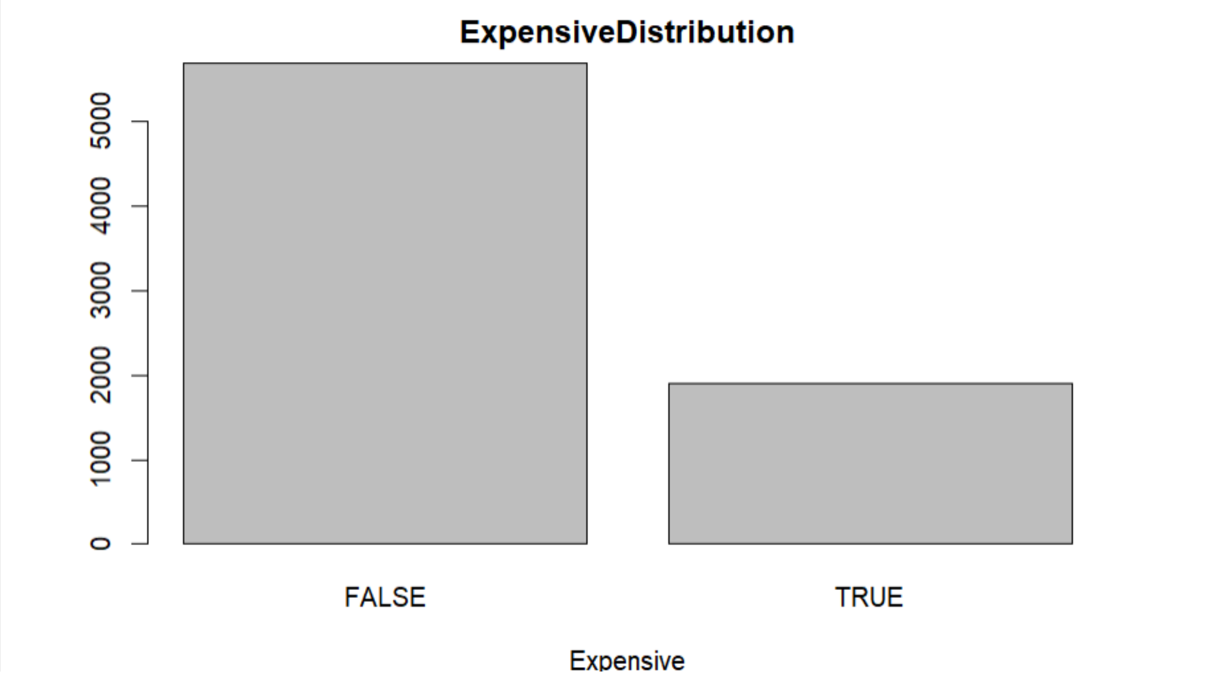
- This map tells us about the age distribution in the seven states
- The regions with low age are covered as white and with high cost are covered as shades of blue. Hence, Pennsylvania has much older sample population while Rhode Island has relatively younger population

3. 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.

```
barplot(counts, main="ExpensiveDistribution",
   xlab="Expensive")
```



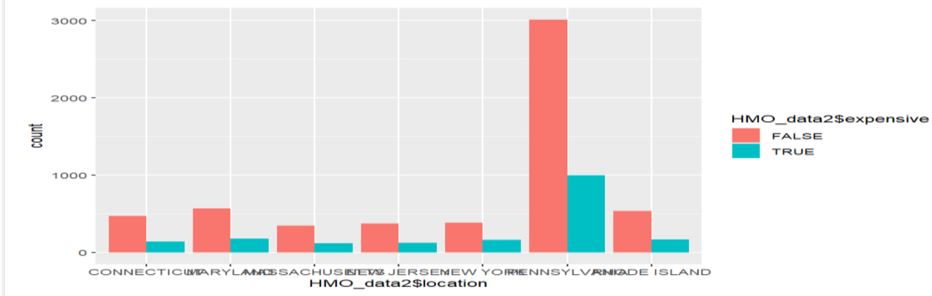
Plot 19: Bar graph showing count of expensive (True) and not-expensive (False)

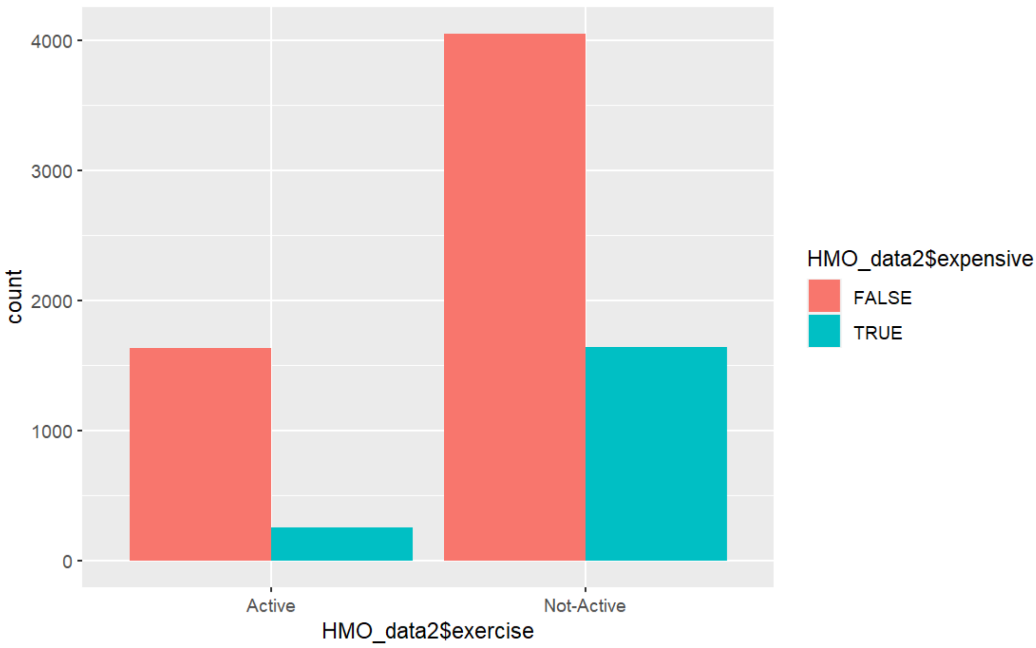
It looks like there’s very few people who are expensive, which makes sense as the threshold value was 75th percentile of the cost column.

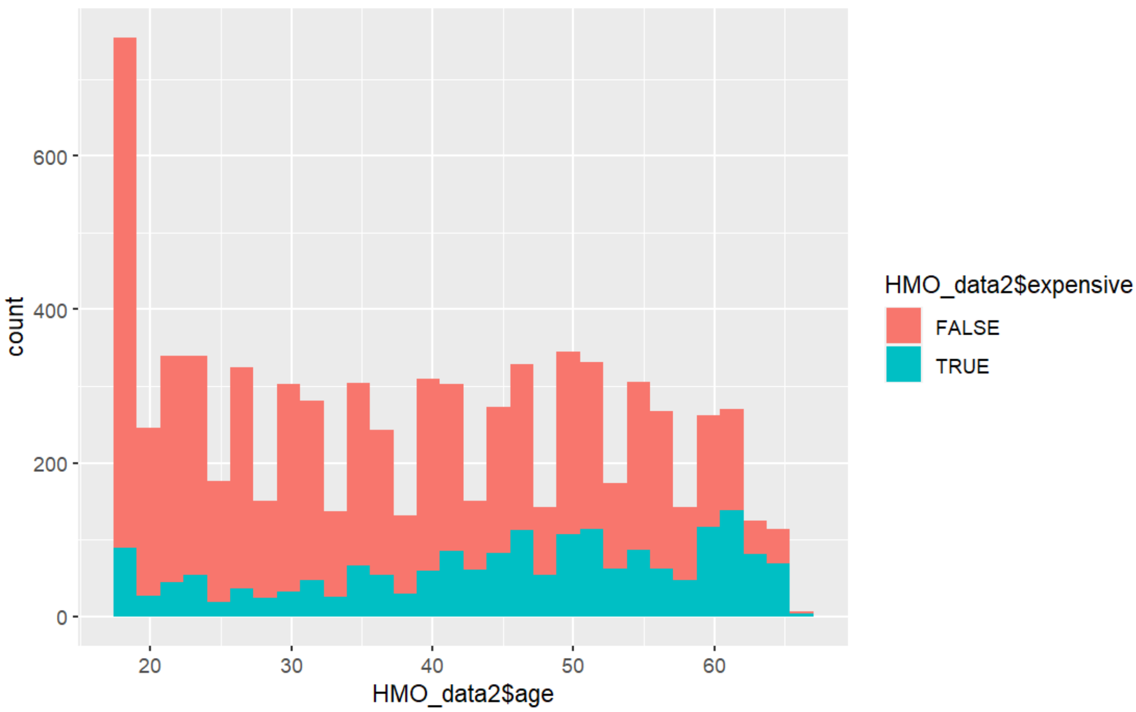
We then tried to find out if there’s any relationship between people being expensive with:

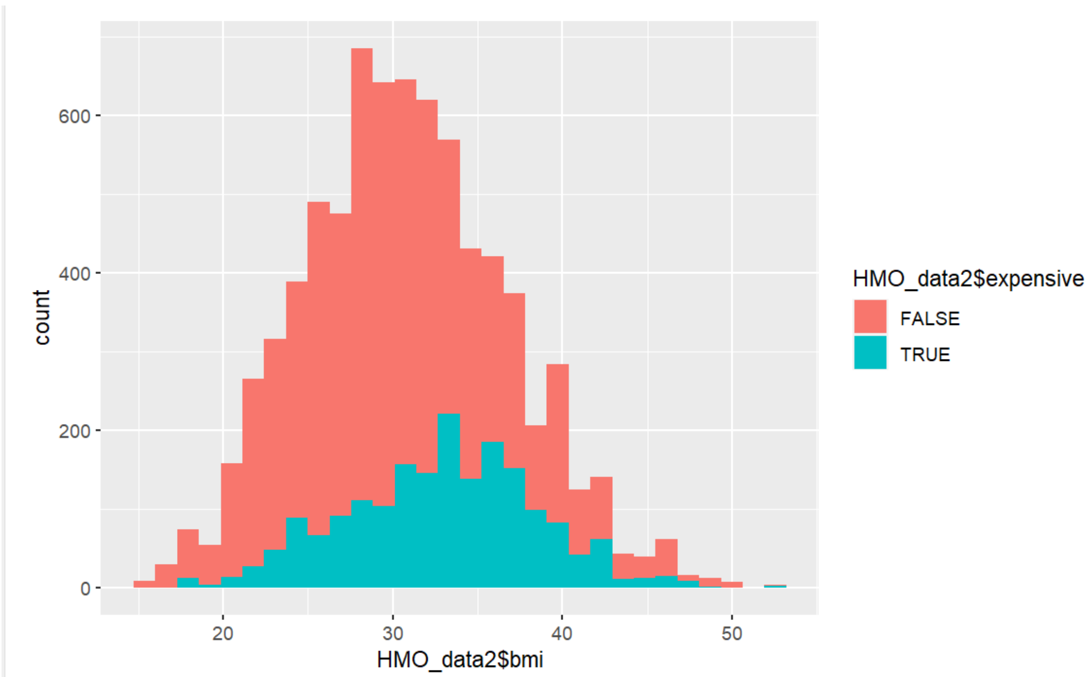
1) Number of children they have
2) Their educational level
3) If they are smoker or not
4) Their location
5) Whether they exercise or not
6) Their marital status
7) BMI index
8) Age

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.

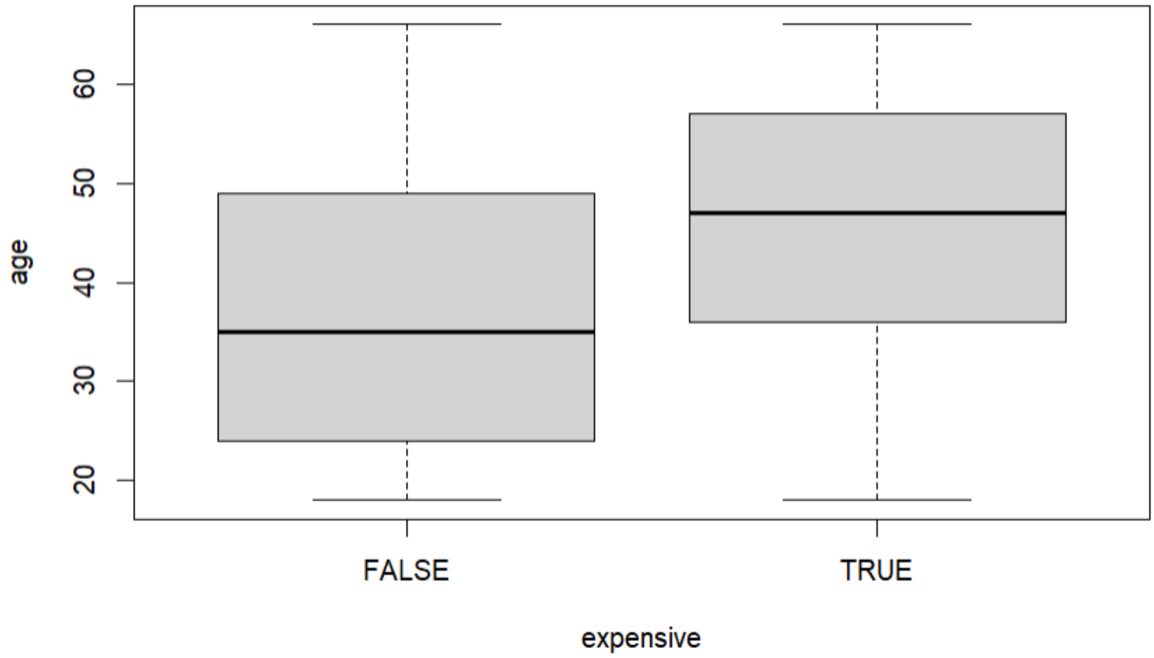





600
400
200
0
count
20
30
40
50
60
HMO_data2$age
HMO_data2$expensive
FALSE
TRUE


600
400
200
0
count
20
30
40
50
HMO_data2$bmi
HMO_data2$expensive
FALSE
TRUE

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:

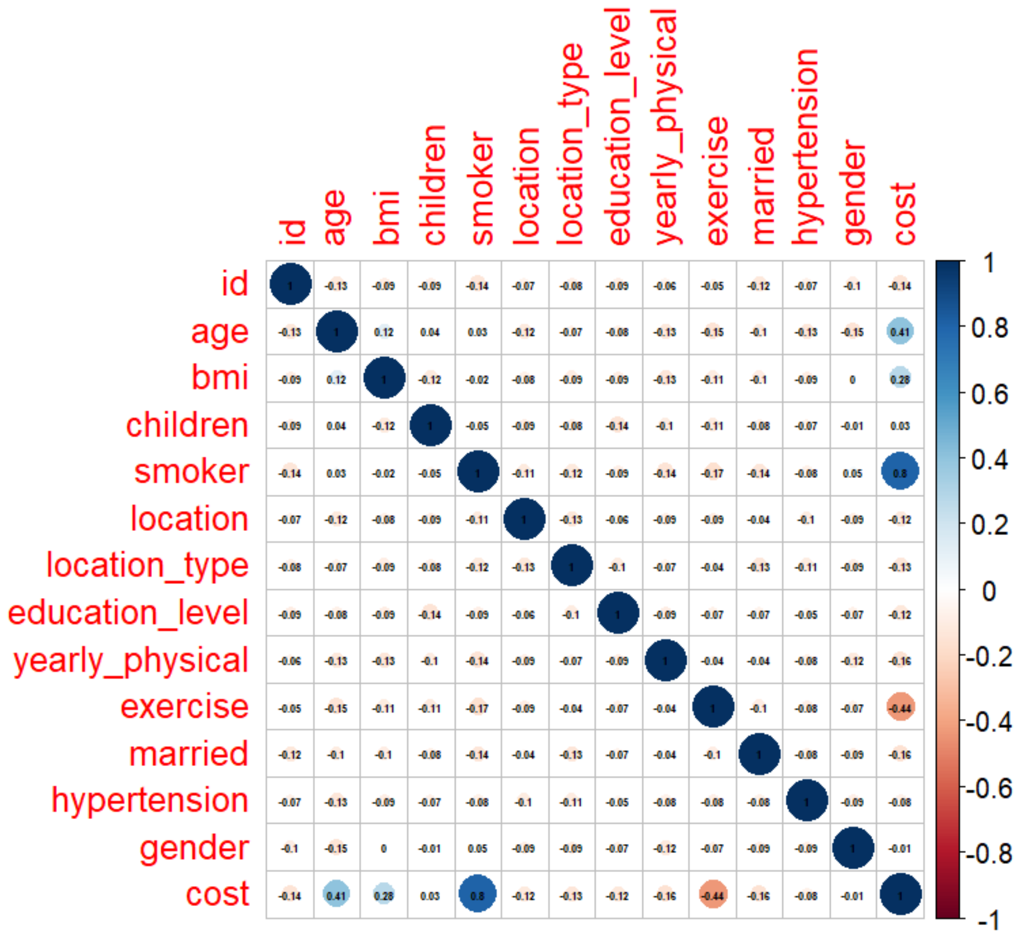
1) Older age in general as compared to less expensive people
2) Higher BMI as compared to less expensive folks
3) Exercise less

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

4. Linear Modelling

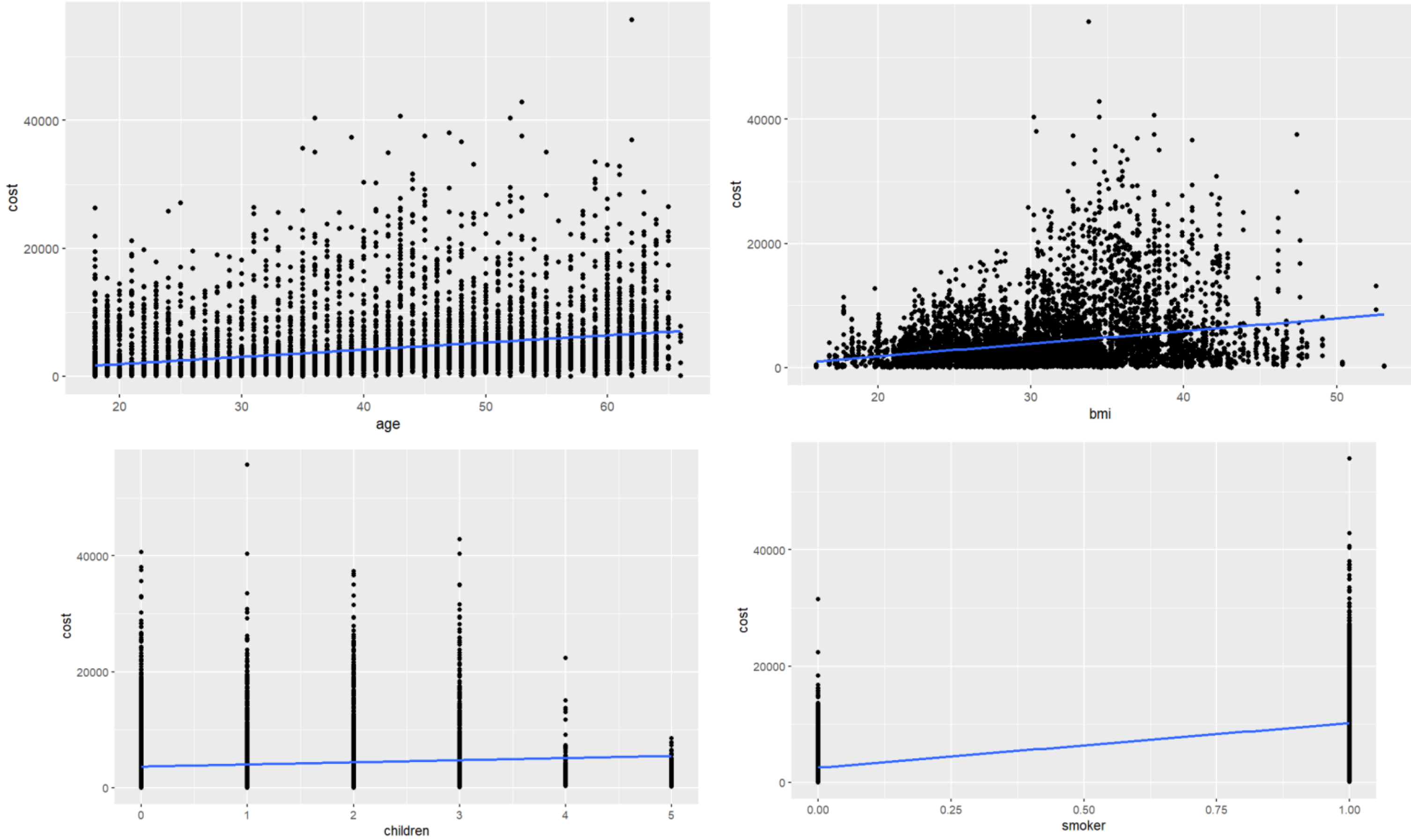
Linear regression creates a predictive model showing trends in data. It is used to show a relation between two variables: a dependent (Y) and an independent (X) variable. The independent variable is known as the predictor variable and the dependent variable is known as the outcome or response variable. 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.

Model No	age	bmi	smoker	children	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	Adjusted R-squared
lmOut1	N	Y	N	N	N	N	N	N	N	N	N	N	0.06061
lmOut2	N	N	Y	N	N	N	N	N	N	N	N	N	0.3813
lmOut3	Y	N	N	N	N	N	N	N	N	N	N	N	0.1048
lmOut4	Y	Y	N	N	N	N	N	N	N	N	N	N	0.1517
lmOut5	N	Y	Y	N	N	N	N	N	N	N	N	N	0.4412
lmOut6	Y	N	Y	N	N	N	N	N	N	N	N	N	0.4825
lmOut7	Y	Y	Y	N	N	N	N	N	N	N	N	N	0.529
lmOut8	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	0.5727
lmOut9	Y	Y	Y	Y	N	N	N	N	Y	N	Y	N	0.5726

Table 6: Regression Matrix

4.1 Linear Regression with scaled data:

```
```{r}
#Applied linear regression on outcome variable cost using predictor smoker
lmOut2 <- lm(formula = cost ~ smoker, data = data_new2)
summary(lmOut2)
```


Call:
lm(formula = cost ~ smoker, data = data_new2)

Residuals:
    Min   1Q Median   3Q   Max
-2.0589 -0.4004 -0.1381 0.2464 9.2295

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.30400   0.01007 -30.19   <2e-16 ***
smoker   1.55845   0.02280   68.36   <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7866 on 7580 degrees of freedom
Multiple R-squared: 0.3814,   Adjusted R-squared: 0.3813
F-statistic: 4674 on 1 and 7580 DF, p-value: < 2.2e-16
```

Table 7: cost ~ smoker

Table 8: cost ~ bmi + age + smoker

Table 9: cost with all the variables

Table 10: Cost with only 6 variables

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

```
```{r}
#Applied linear regression on outcome variable cost using predictors bmi, smoker, age, children, location, location_type, education_level,
exercise, married, hypertension, gender, yearly_physical, id
lmOut8 <- lm(formula = cost ~ bmi + age + smoker + children +location + location_type + exercise + married + hypertension + gender +
education_level + yearly_physical , data = data_new2)
summary(lmOut8)
```
```

```
Call:
lm(formula = cost ~ bmi + age + smoker + children + location +
    location_type + exercise + married + hypertension + gender +
    education_level + yearly_physical, data = data_new2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4576 -0.3018 -0.0740  0.2051  8.4787

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.196243   0.017686 -11.096  < 2e-16 ***
bmi              0.219813   0.007557  29.086  < 2e-16 ***
age              0.294726   0.007565  38.958  < 2e-16 ***
smoker           1.555444   0.019024  81.762  < 2e-16 ***
children         0.057861   0.007539   7.675 1.86e-14 ***
location         0.007214   0.007514   0.960 0.337065
location_type    0.002535   0.017333   0.146 0.883747
exercise        -0.459239   0.017372 -26.435  < 2e-16 ***
married         -0.026577   0.015949  -1.666 0.095684 .
hypertension     0.068255   0.018821   3.627 0.000289 ***
gender           0.006487   0.015121   0.429 0.667921
education_level -0.010642   0.007515  -1.416 0.156807
yearly_physical  0.029174   0.017382   1.678 0.093322 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6537 on 7569 degrees of freedom
Multiple R-squared:  0.5734,	Adjusted R-squared:  0.5727
F-statistic: 847.7 on 12 and 7569 DF,  p-value: < 2.2e-16
```

```
```{r}
#Applie
lmOut9
summary(
```
```

```
Call:
lm(formula = cost ~ bmi + age + smoker + children + exercise +
    hypertension, data = data_new2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4729 -0.3018 -0.0724  0.2052  8.4786

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.202822   0.010108 -20.065  < 2e-16 ***
bmi           0.219870   0.007544  29.143  < 2e-16 ***
age           0.294266   0.007561  38.917  < 2e-16 ***
smoker        1.555525   0.018959  82.048  < 2e-16 ***
children      0.058149   0.007529   7.723 1.28e-14 ***
exercise     -0.458482   0.017369 -26.397  < 2e-16 ***
hypertension  0.067623   0.018819   3.593 0.000328 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6538 on 7575 degrees of freedom
Multiple R-squared:  0.5729,	Adjusted R-squared:  0.5726
F-statistic:  1693 on 6 and 7575 DF,  p-value: < 2.2e-16
```

```
```{r}
#Creating a dataframe using the suggested values
predDF <- data.frame(age = 0.5723449, bmi=0.4422956, smoker=0)
predict(lmOut7, predDF)
```
```

```
          1
 -0.03611013
```

```
We tried applying linear model using various parameters and combinations using  bmi, smoker, age, children, location, location_type,
education_level, exercise, married, hypertension, gender, yearly_physical, id to predict cost. By adding the independent variable i.e bmi,
age, smoker, children, exercise and hypertension we are increasing the overall prediction capacity of our model which is 57%.
```

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.

```{r}
#Applied linear regression on outcome variable cost using predictors bmi, smoker, age, children,exercise, hypertension
lmOut2 <- lm(formula = cost ~ bmi + age + smoker + children + exercise  + hypertension , data = data)
summary(lmOut2)
```

```
Call:
lm(formula = cost ~ bmi + age + smoker + children + exercise + 
    hypertension, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-12258  -1486   -362   1005  41834 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -6695.369    213.007 -31.433  < 2e-16 ***
bmi            181.290      6.204  29.220  < 2e-16 ***
age            102.375      2.628  38.962  < 2e-16 ***
smoker        7674.160     93.416  82.151  < 2e-16 ***
children       236.909     30.434   7.784 7.94e-15 ***
exercise     -2262.758     85.582 -26.440  < 2e-16 ***
hypertension   335.595     92.781   3.617    3e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3221 on 7575 degrees of freedom
Multiple R-squared:  0.5731,	Adjusted R-squared:  0.5728 
F-statistic:  1695 on 6 and 7575 DF,  p-value: < 2.2e-16
```

Table 12: cost ~ With all 6 variables

Using predict()

```{r}
predict(lmOut2, predDF)
```

```
         1          2          3          4          5          6          7 
 2832.2315  4131.6803 10449.5968 10483.2437 11207.0794 10020.5787 10681.1747 
         8          9         10         11         12         13         14 
11446.7485   919.5543  1329.9867  6775.8834  2407.2007  4296.1308  6640.6058 
        15         16         17         18         19         20 
 8229.8571  3165.1742  3042.9400   510.3914  4852.2144  3042.8814 
```

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).

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

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0525535175772644

Number of Support Vectors : 1467

Objective Function Value : -5744.64
Training error : 0.112088
Cross validation error : 0.127692
Probability model included.
```

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:

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  2211  346
     TRUE     62  413

               Accuracy : 0.8654
                 95% CI : (0.8528, 0.8774)
    No Information Rate : 0.7497
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5904

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

            Sensitivity : 0.9727
            Specificity : 0.5441
         Pos Pred Value : 0.8647
         Neg Pred Value : 0.8695
             Prevalence : 0.7497
         Detection Rate : 0.7292
   Detection Prevalence : 0.8433
      Balanced Accuracy : 0.7584
```

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:

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  2163  354
     TRUE    110  405

               Accuracy : 0.847
                 95% CI : (0.8337, 0.8596)
    No Information Rate : 0.7497
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5434

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

            Sensitivity : 0.9516
            Specificity : 0.5336
         Pos Pred Value : 0.8594
         Neg Pred Value : 0.7864
             Prevalence : 0.7497
         Detection Rate : 0.7134
   Detection Prevalence : 0.8301
      Balanced Accuracy : 0.7426
```

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%

```
Support Vector Machines with Radial Basis Function Kernel

4550 samples
  12 predictor
   2 classes: 'FALSE', 'TRUE'

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 4095, 4095, 4094, 4095, 4095, 4095, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.8602216  0.5864681
  0.50  0.8654968  0.5952720
  1.00  0.8683535  0.5987928

Tuning parameter 'sigma' was held constant at a value
 of 0.05280349
Accuracy was used to select the optimal model using
 the largest value.
The final values used for the model were sigma =
 0.05280349 and C = 1.
```

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.

```
Support Vector Machines with Radial Basis Function Kernel

4550 samples
   6 predictor
   2 classes: 'FALSE', 'TRUE'

Pre-processing: centered (6), scaled (6)
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 4095, 4095, 4096, 4094, 4095, 4095, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.8707643  0.6044158
  0.50  0.8723027  0.6068646
  1.00  0.8729602  0.6103848

Tuning parameter 'sigma' was held constant at a value
 of 0.1710915
Accuracy was used to select the optimal model using
 the largest value.
The final values used for the model were sigma =
 0.1710915 and C = 1.
```

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:

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  2226  359
     TRUE     47  400

               Accuracy : 0.8661
                 95% CI : (0.8535, 0.878)
    No Information Rate : 0.7497
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5866

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

            Sensitivity : 0.9793
            Specificity : 0.5270
         Pos Pred Value : 0.8611
         Neg Pred Value : 0.8949
             Prevalence : 0.7497
         Detection Rate : 0.7342
   Detection Prevalence : 0.8526
      Balanced Accuracy : 0.7532

       'Positive' Class : FALSE
```

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:

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE     9    5
     TRUE      3    3

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

                  Kappa : 0.1304

 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.7500
            Specificity : 0.3750
         Pos Pred Value : 0.6429
         Neg Pred Value : 0.5000
             Prevalence : 0.6000
         Detection Rate : 0.4500
   Detection Prevalence : 0.7000
      Balanced Accuracy : 0.5625

       'Positive' Class : FALSE
```

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.

	id <dbl>	age <dbl>	bmi <dbl>	children <dbl>	smoker <dbl>	location <dbl>	location_type <dbl>	education_level <dbl>
1	1	18	27.900	0	1	1	0	1
2	2	19	33.770	1	0	7	0	1
3	3	27	33.000	3	0	3	0	2
4	4	34	22.705	0	0	6	1	2
5	5	32	28.880	0	0	6	1	4
6	7	47	33.440	1	0	6	0	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.

```
```{r}
kmodel
```

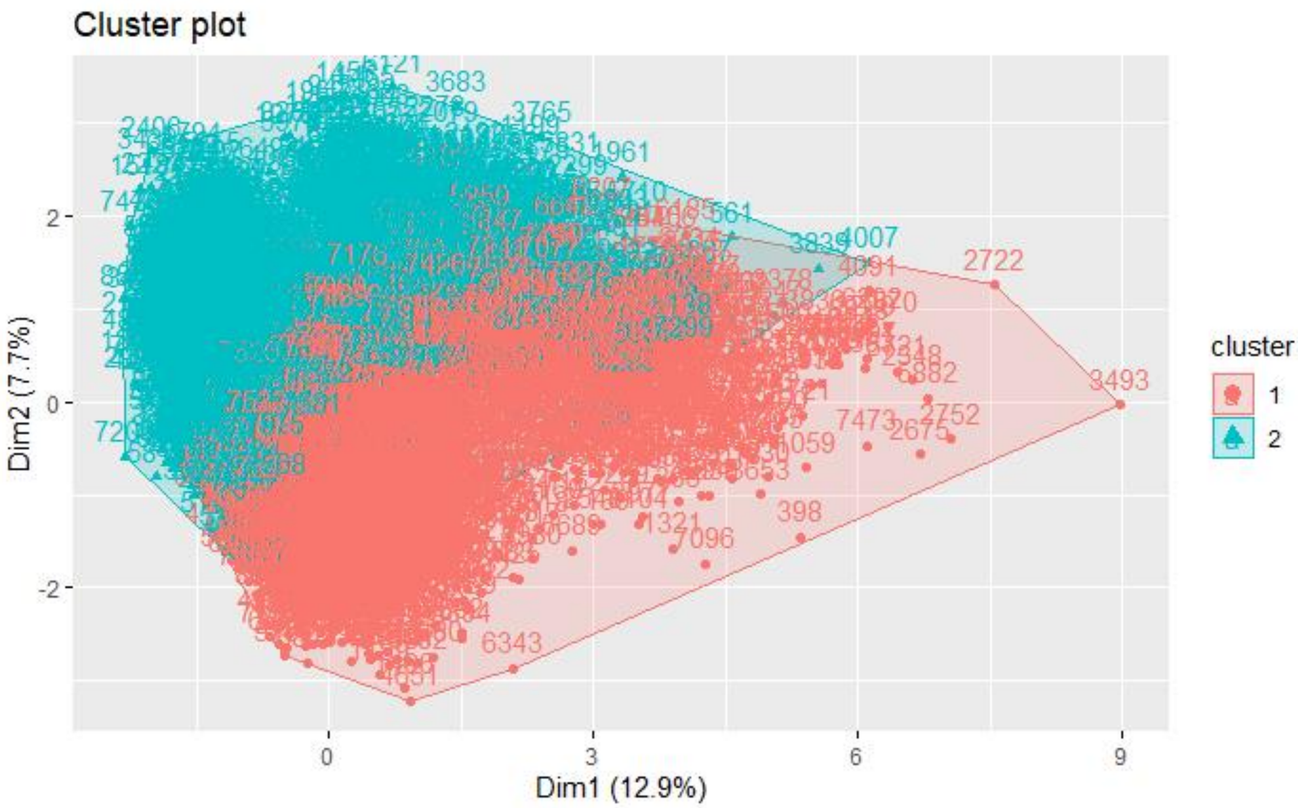
K-means clustering with 2 clusters of sizes 3786, 3796

Cluster means:
         age   bmi   smoker exercise
1 -0.8597969 -0.2270099 0.1973059 0.2485473
2 0.8575319 0.2264119 0.1928346 0.2494731
```

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

```{r}
data[c('age', 'bmi','cost','cluster')] %>%
  group_by(cluster) %>%
  summarise_all("mean")
```

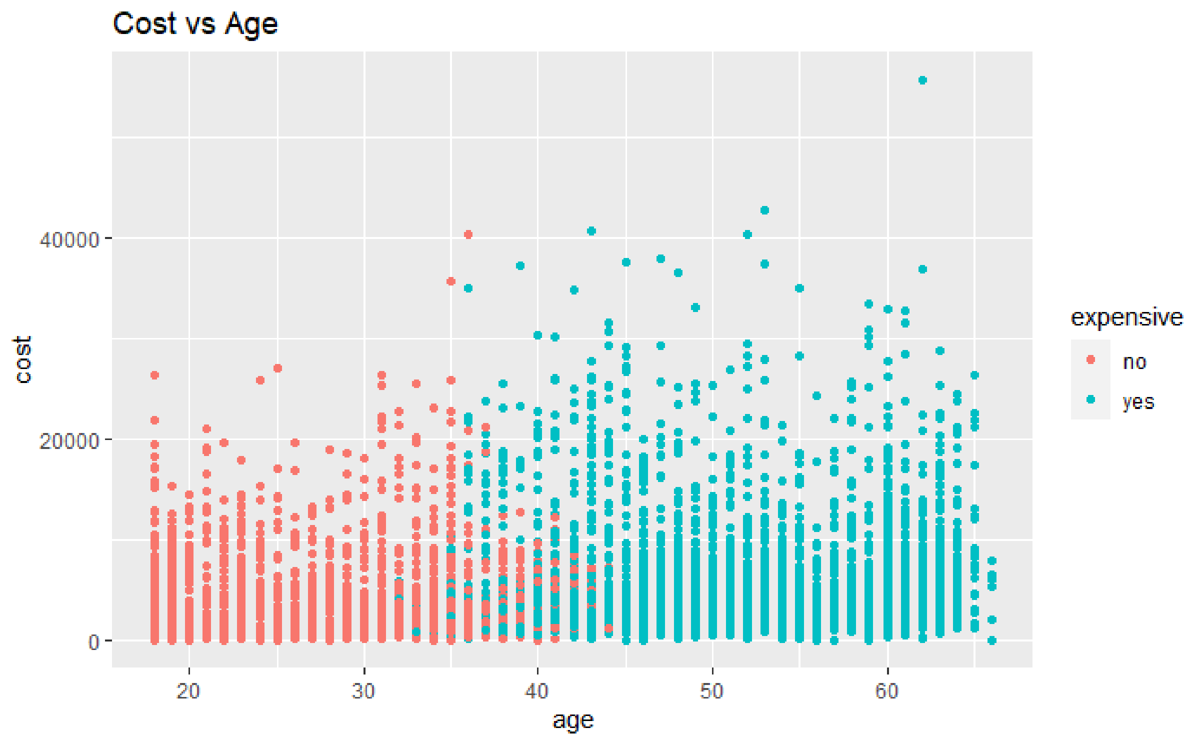
A tibble: 2 x 4

cluster <int>	age <dbl>	bmi <dbl>	cost <dbl>
1	51.04320	32.14985	5607.202
2	26.69599	29.44013	2474.589

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