IST 687 - DATA ANALYSIS ON HEALTHCARE COST

BY

HIRAL PAGHADAL

KABIR THAKUR

SAGNIK DAS

SOWMEYA M

Contents:

Project Description:

- This project is about healthcare expenses and affecting factors.
- Everybody's life is centered around health. Our lives are moving so quickly that we are forming bad habits which affect our health. There are many factors that result in some people paying more in the hospital. We need to analyze key drivers of why some people pay more in the hospital and why they are termed "expensive" and find key insights on why some people are termed "inexpensive".
- So, a predictive model is used to understand the factors affecting health and creating more medical bills and costs, further helping to identify the key drivers affecting cost.

Project Technical Details:

- Dataset consists of data that gives information about individuals along with their details such as health conditions and health care costs. The dataset consists of 7582 observations and 14 variables.
- Checked if there are any null values and discrepancies or missing values in the dataset.
- The columns BMI and Hypertension had NA values which were removed.
- Identifying the categorical and numerical data separately to ease the conversion of categorical data to numerical data.

Goal:

To analyze and provide insight, based on the data from the given dataset. Predict who will spend more on healthcare next year. Moreover, provide actionable insight to HMO on how to lower the cost of people who are termed “expensive”

Objectives:

1. Determine key factors/drivers on why some individuals’ healthcare cost is expensive.
2. Predict which people will be expensive in terms of health care costs based on the given data.
3. We will address our goal in the aforementioned phases:
   A. Data Preprocessing and Cleaning

      First, we will check if there is any column with NA values. If yes, we will eliminate the missing value with approximate values using na_interpolation function.

   B. Data visualization

      We will plot the histogram, boxplot, scatterplot, map, etc to get the visual representation of data and interpret certain factors which affect healthcare costs.

   C. Preparing predictive models

      Few predictive models will be used to predict whether the expense of an individual is high or low based on their habits and health conditions. For this, the entire dataset will be divided into 2, test data and train data.

      We will use linear and multiple linear regression, Tree Bag, and Support Vector Machine models to predict the output.

Packages used:

- tidyverse
- imputeTS

- ggplot2
- ggmap
- kernlab
- caret
- rpart
- rpart.plot
- usmap
- dplyr

Data Preprocessing and Cleaning:

We have checked for empty values in all the columns in the given dataset.

There were two columns that had “NA” values and they were BMI and Hypertension.

````
```{r}
sum(is.na(data$X))
sum(is.na(data$age))
sum(is.na(data$bmi))
sum(is.na(data$children))
sum(is.na(data$smoker))
sum(is.na(data$location))
sum(is.na(data$location_type))
sum(is.na(data$education_level))
sum(is.na(data$yearly_physical))
sum(is.na(data$exercise))
sum(is.na(data$hypertension))


#We can see that bmi and hypertension are the only two variables which have NA values. Both variables are numerical so we can run NA interpolation to get an approximate value for the data Correcting NA values
```
````

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

Finding NA values

````
```{r}
sum(is.na(data))
```
````

```
[1] 158
```

“na_interpolation” was used to remove the NA from the above numerical variables.

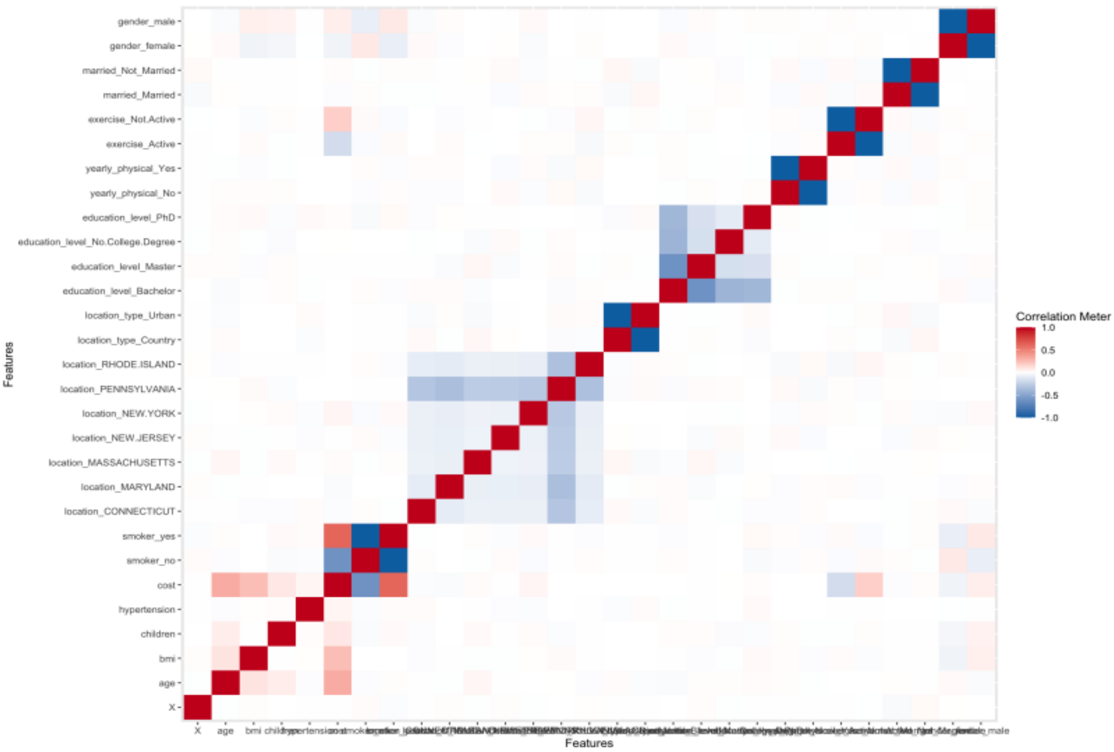
```{r}
# using na_interpolation to replace the variables having NA with approximate values
data$bmi=na_interpolation(data$bmi)
data$hypertension=na_interpolation(data$hypertension)
sum(is.na(data))
# now the sum is zero, hence no NA values
```

```
[1] 0
```

We convert the categorical variables to numerical variables for model prediction

Correlation Analysis

Correlation is used to determine the association between two or more variables.



We find the correlation between variables correlation ranges from -1 to 1.

The cost of healthcare of an individual is highly correlated with smoker as if the individual is smoker, then the health expense for that individual would be more

	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise
age	1.000000000	0.094426327	0.0671967382	0.009277045	-0.0045929431	-1.346684e-02	0.017164197	-0.0088217019	-0.001869411
bmi	0.094426327	1.000000000	-0.0086844242	0.002729460	0.0156260823	-3.884952e-03	0.008898199	-0.0101230710	-0.013057213
children	0.067196738	-0.008684424	1.0000000000	0.020778740	0.0009466184	-6.685550e-04	-0.021624791	-0.0057930236	0.009678072
smoker	0.009277045	0.002729460	0.0207787399	1.000000000	0.0015521082	4.769516e-03	0.017548553	-0.0117957558	-0.019021459
location	-0.004592943	0.015626082	0.0009466184	0.001552108	1.0000000000	2.607389e-02	0.010834334	-0.0047156116	0.012765463
location_type	-0.013466839	-0.003884952	-0.0006685550	0.004769516	0.0260738941	1.000000e+00	0.013485358	-0.0009767752	0.009237691
education_level	0.017164197	0.008898199	-0.0216247905	0.017548553	0.0108343343	1.348536e-02	1.000000000	-0.0056124603	0.003272052
yearly_physical	-0.008821702	-0.010123071	-0.0057930236	-0.011795756	-0.0047156116	-9.767752e-04	-0.005612460	1.0000000000	-0.007844524
exercise	-0.001869411	-0.013057213	0.0096780718	-0.019021459	0.0127654629	9.237691e-03	0.003272052	-0.0078445245	1.000000000
married	-0.005047340	-0.004892709	-0.0050714305	0.007799440	-0.0174858369	-2.970218e-02	-0.004624324	-0.0170672015	-0.020063839
hypertension	-0.013897310	0.007555837	0.0111440319	0.014592339	-0.0094686094	1.486472e-02	0.016070792	0.0003869623	0.006110619
gender	-0.017235495	0.055414351	0.0463114424	0.082817826	0.0037798087	-7.450586e-05	0.013932009	-0.0101023487	-0.007249378
cost	0.264633852	0.196795555	0.0611157923	0.546075464	0.0067452954	-9.397062e-03	0.015261885	0.0094293139	0.155587738

	married	hypertension	gender	cost
age	-0.0050473403	-0.0138973101	-1.723550e-02	0.264633852
bmi	-0.0048927090	0.0075558365	5.541435e-02	0.196795555
children	-0.0050714305	0.0111440319	4.631144e-02	0.061115792
smoker	0.0077994395	0.0145923390	8.281783e-02	0.546075464
location	-0.0174858369	-0.0094686094	3.779809e-03	0.006745295
location_type	-0.0297021775	0.0148647246	-7.450586e-05	-0.009397062
education_level	-0.0046243245	0.0160707922	1.393201e-02	0.015261885
yearly_physical	-0.0170672015	0.0003869623	-1.010235e-02	0.009429314
exercise	-0.0200638389	0.0061106186	-7.249378e-03	0.155587738
married	1.0000000000	-0.0004190491	-3.310296e-03	0.007542641
hypertension	-0.0004190491	1.0000000000	3.728884e-03	0.039382543
gender	-0.0033102957	0.0037288842	1.000000e+00	0.067821686
cost	0.0075426413	0.0393825433	6.782169e-02	1.000000000

DATASET AND VARIABLE

X: Integer, Unique identifier 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 good 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 healthcare for that person, during the past year.

Exploratory Data Analysis:

We explore the data by using the str() function which gives the internal structure of the data frame. It gives few of the contents of the columns and their data type.

```
str(data) #str gives us the structure of the data
```

```
spec_tbl_df [7,582 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ X              : num [1:7582] 1 2 3 4 5 7 9 10 11 12 ...
 $ age            : num [1:7582] 18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num [1:7582] 27.9 33.8 33 22.7 28.9 ...
 $ children       : num [1:7582] 0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : chr [1:7582] "yes" "no" "no" "no" ...
 $ location       : chr [1:7582] "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...
 $ location_type  : chr [1:7582] "Urban" "Urban" "Urban" "Country" ...
 $ education_level: chr [1:7582] "Bachelor" "Bachelor" "Master" "Master" ...
 $ yearly_physical: chr [1:7582] "No" "No" "No" "No" ...
 $ exercise       : chr [1:7582] "Active" "Not-Active" "Active" "Not-Active" ...
 $ married        : chr [1:7582] "Married" "Married" "Married" "Married" ...
 $ hypertension   : num [1:7582] 0 0 0 1 0 0 0 1 0 0 ...
 $ gender         : chr [1:7582] "female" "male" "male" "male" ...
 $ cost           : num [1:7582] 1746 602 576 5562 836 ...
 - attr(*, "spec")=
  .. cols(
  ..   X = col_double(),
  ..   age = col_double(),
  ..   bmi = col_double(),
  ..   children = col_double(),
  ..   smoker = col_character(),
  ..   location = col_character(),
  ..   location_type = col_character(),
  ..   education_level = col_character(),
  ..   yearly_physical = col_character(),
  ..   exercise = col_character(),
  ..   married = col_character(),
  ..   hypertension = col_double(),
  ..   gender = col_character(),
  ..   cost = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

summary( ) function gives us the Each column's summary result. If the column is of the numerical type, the summary will include information such as minimum, that is the minimum value in the column , maximum that is the maximum value in the column , median that is the middlemost value in the column , mean is the average value in the column, 3rd quartile which is 75% of the data, If the column is of the char type, the summary would include details such as length, class, and mode.

Insights:

➔ The minimum age is 18 and the maximum age is 66 years, mean age is 38

➔ The minimum bmi is 15 and the maximum bmi is 53 years which is way over than the accepted value, bmi over 25 is termed obese

➔ People have maximum of 5 children and minimum 0 children and mean 1 children

➔ the minimum cots is 2 and the maximum is 55715 and the mean is 4043

```
summary(data) #summary gives us the statistical description of the data
```

```
       X                  age             bmi           children        smoker            location         location_type
 Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000   Length:7582        Length:7582        Length:7582
 1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000   Class :character   Class :character   Class :character
 Median :    24916   Median :39.00   Median :30.50   Median :1.000   Mode  :character   Mode  :character   Mode  :character
 Mean   :   712602   Mean   :38.89   Mean   :30.80   Mean   :1.109
 3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.:2.000
 Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000
                                     NA's   :78
 education_level    yearly_physical      exercise            married           hypertension       gender               cost
 Length:7582        Length:7582        Length:7582        Length:7582        Min.   :0.0000   Length:7582        Min.   :    2
 Class :character   Class :character   Class :character   Class :character   1st Qu.:0.0000   Class :character   1st Qu.:  970
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Median : 2500
                                                                             Mean   :0.2005                      Mean   : 4043
                                                                             3rd Qu.:0.0000                      3rd Qu.: 4775
                                                                             Max.   :1.0000                      Max.   :55715
                                                                             NA's   :80
```

unique() function displays the unique values within the columns and eliminates the duplicate ones

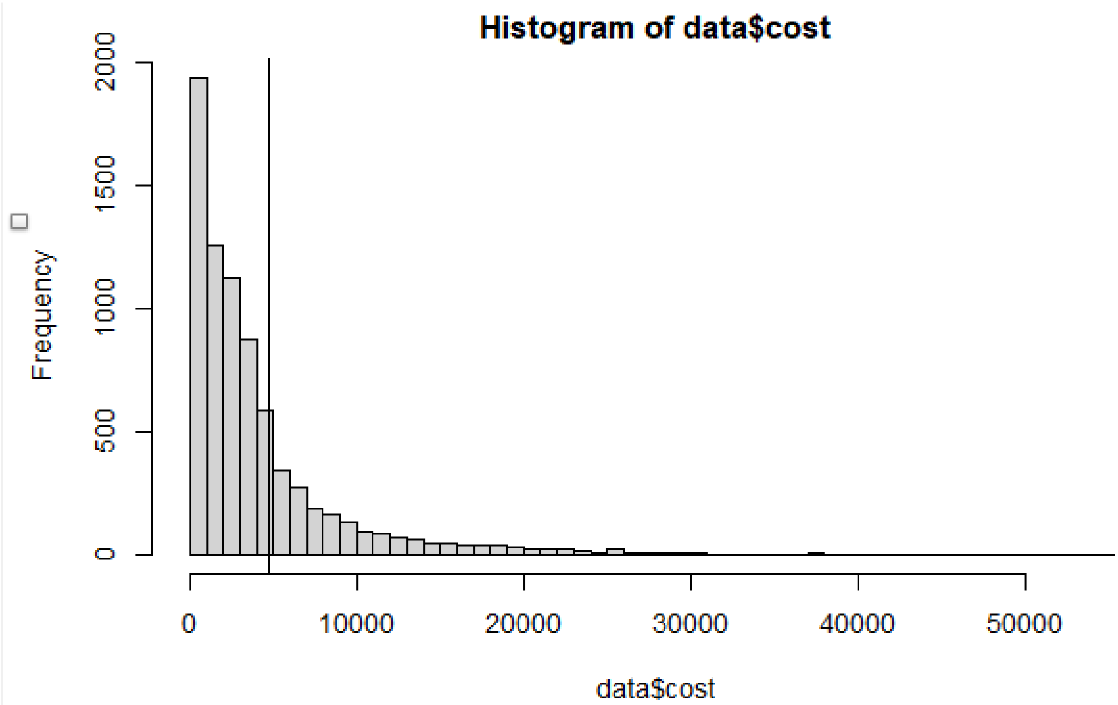
```
unique(data$location) #unique is used to eliminate the duplicate values
unique(data$location_type)
unique(data$education_level)
unique(data$gender)
unique(data$children)
```

```
[1] "CONNECTICUT"   "RHODE ISLAND"  "MASSACHUSETTS" "PENNSYLVANIA"  "MARYLAND"      "NEW JERSEY"    "NEW YORK"
[1] "Urban"   "Country"
[1] "Bachelor"           "Master"             "PhD"                "No College Degree"
[1] "female" "male"
[1] 0 1 3 2 5 4
```

Data Visualization

1. Univariate Analysis

Histogram of cost distribution:



A histogram was generated w.r.t to cost as we wanted to decide the point where we can divide the dataset into train set and test data.

We decided to have the splitting line at 75 % quantile of the cost

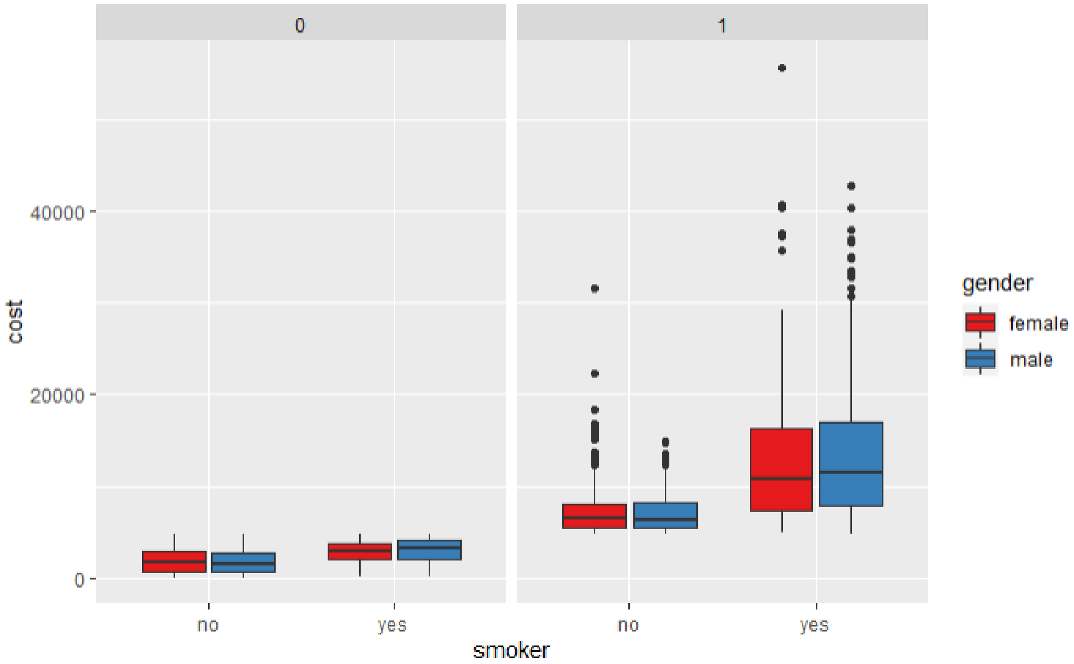
```
 75%
4775
```

The individual is termed “expensive” when their healthcare cost is greater than $4775.

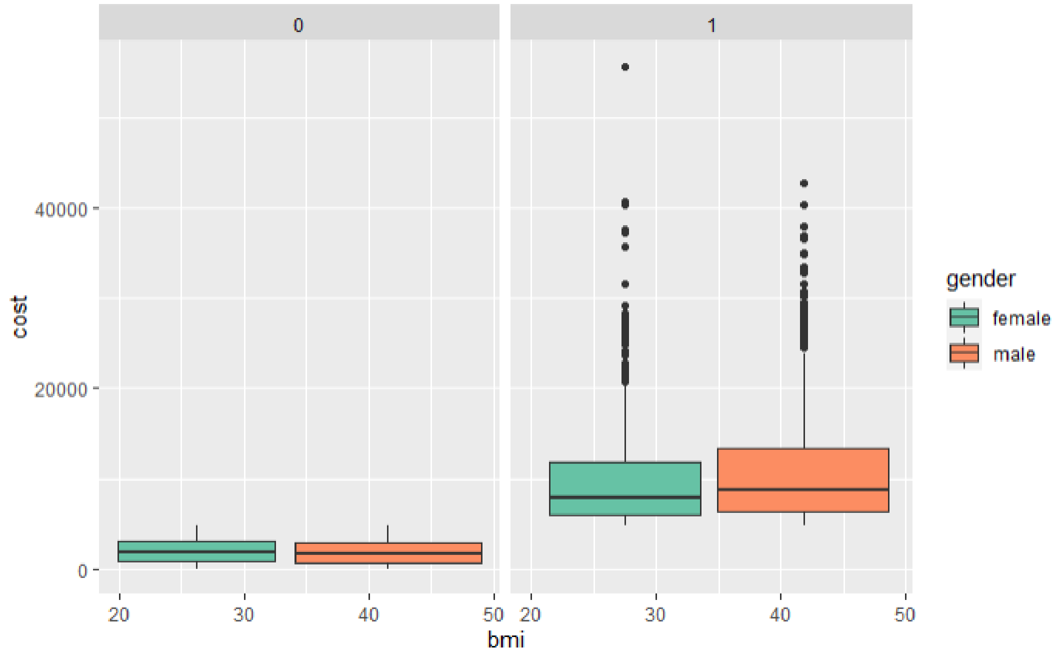
Termed “non-expensive” when their healthcare cost is lower than $4775.

We created a new column cost_new which has a binary value of 1 or 0 depending on whether the cost is higher than 4775 or lesser than 4775. If the cost is higher than 4775 then it is 1, if it is less than 4775 it is 0

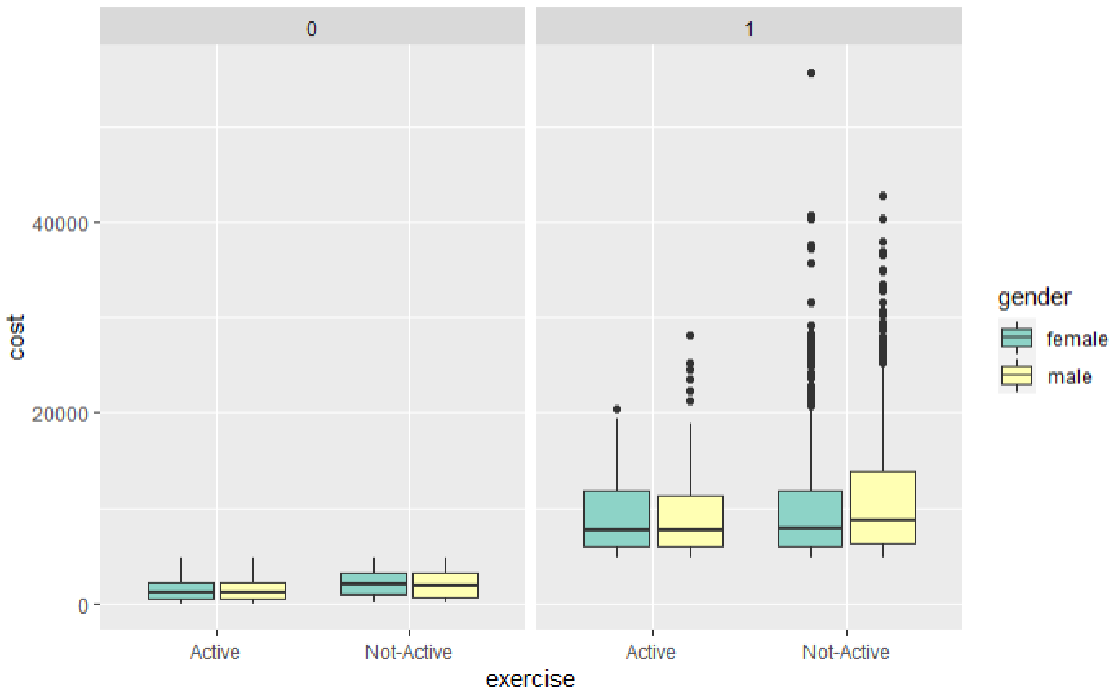
Boxplot



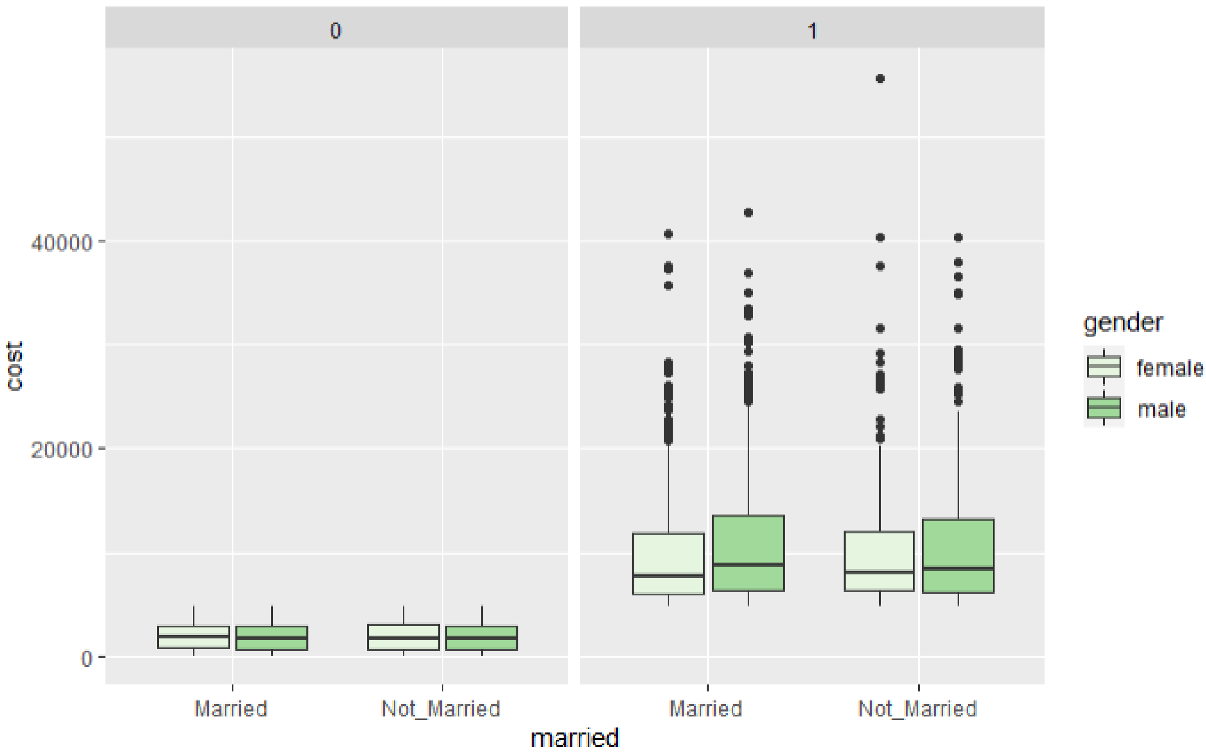
We first plot boxplot for people who smoke vs cost variable. We find that males who smoke have a higher cost and therefore are termed “expensive” than females. The plot has a high number of outliers for people who are expensive. There is an outlier which is above 40,000 cost. People who don’t smoke have similar costs, irrespective of their gender.



Second boxplot, we plotted the BMI with the cost variable. The Males had a higher bmi than females for people who were expensive and as the bmi increased the cost increased.

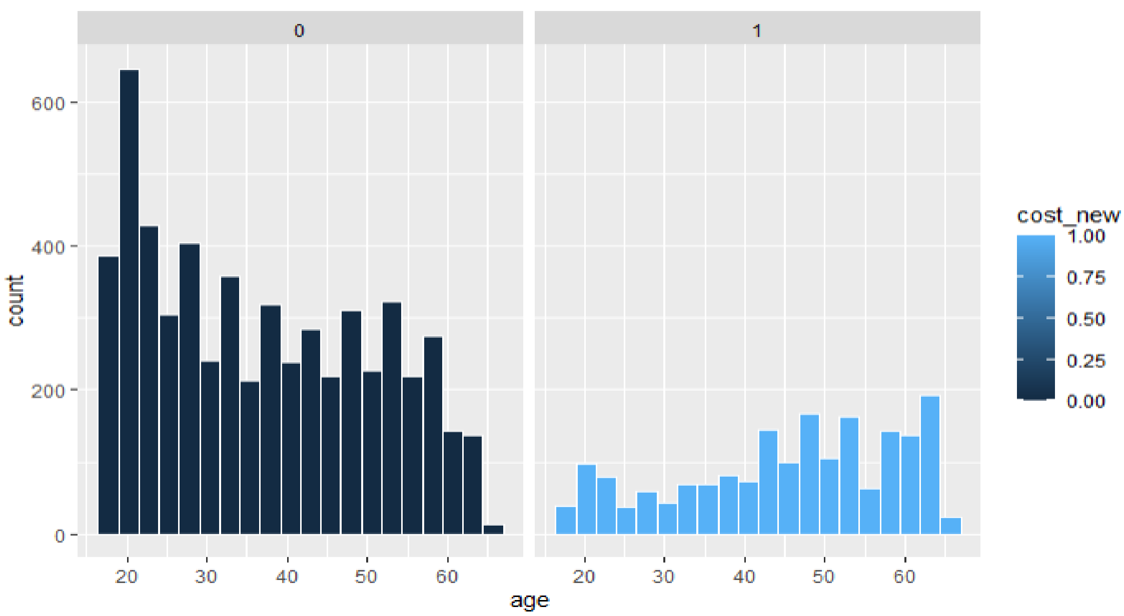


Thirdly, we plot an exercise variable with the cost and find a key insight that people who are not active have a higher cost than people who are active.

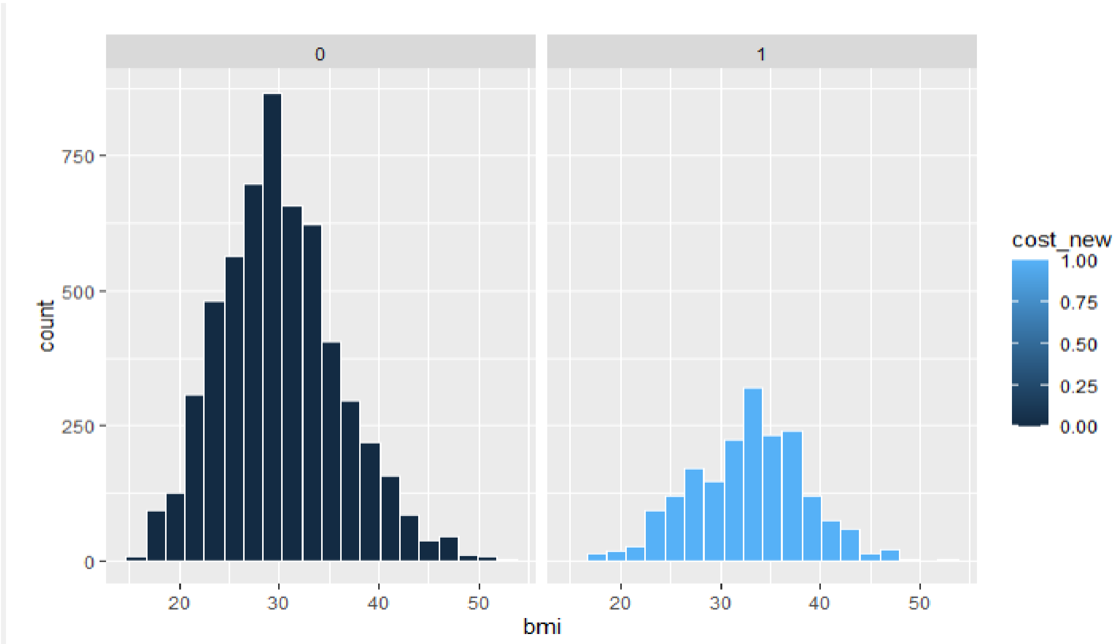


We plotted the marriage variable to the cost variable and found that people who are married or not married do not affect the cost. The cost variable is independent of the married variable

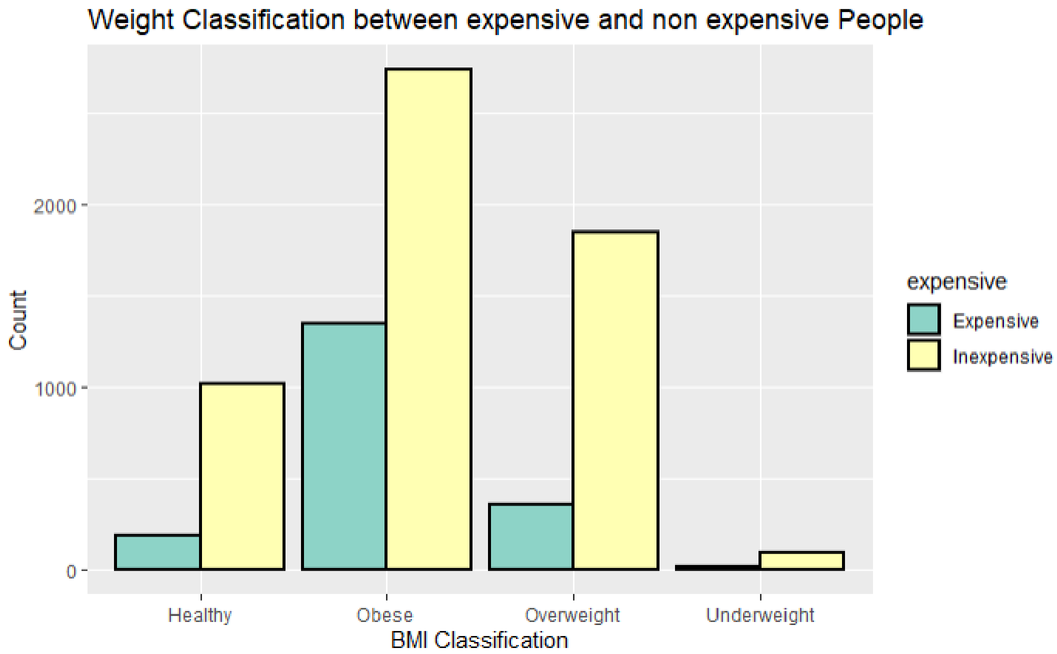
Histogram



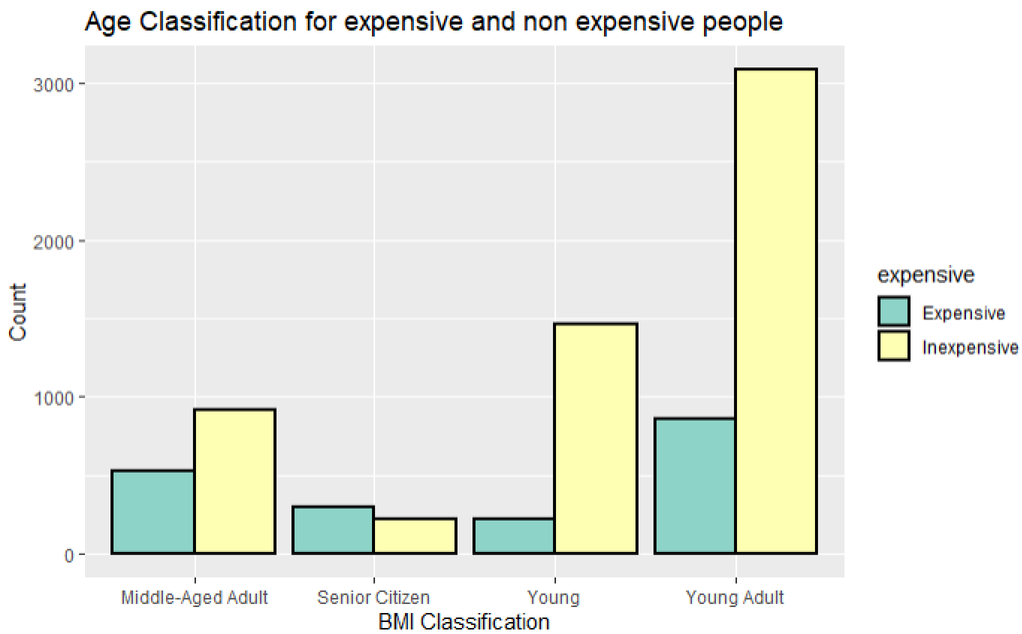
We plot a histogram of age with the cost and find that people who are the most expensive fall in the age range of 40 - 60 years. People who are in the age range of 20-30 have a low cost and therefore age reflects the cost factor.



We plot histogram for bmi and find that for both expensive and non expensive the bmi is normally distributed. We find that people who have the bmi between 25-30 are high in count and therefore they are not expensive since a large number of people who are not expensive have a bmi which is in the accepted age.



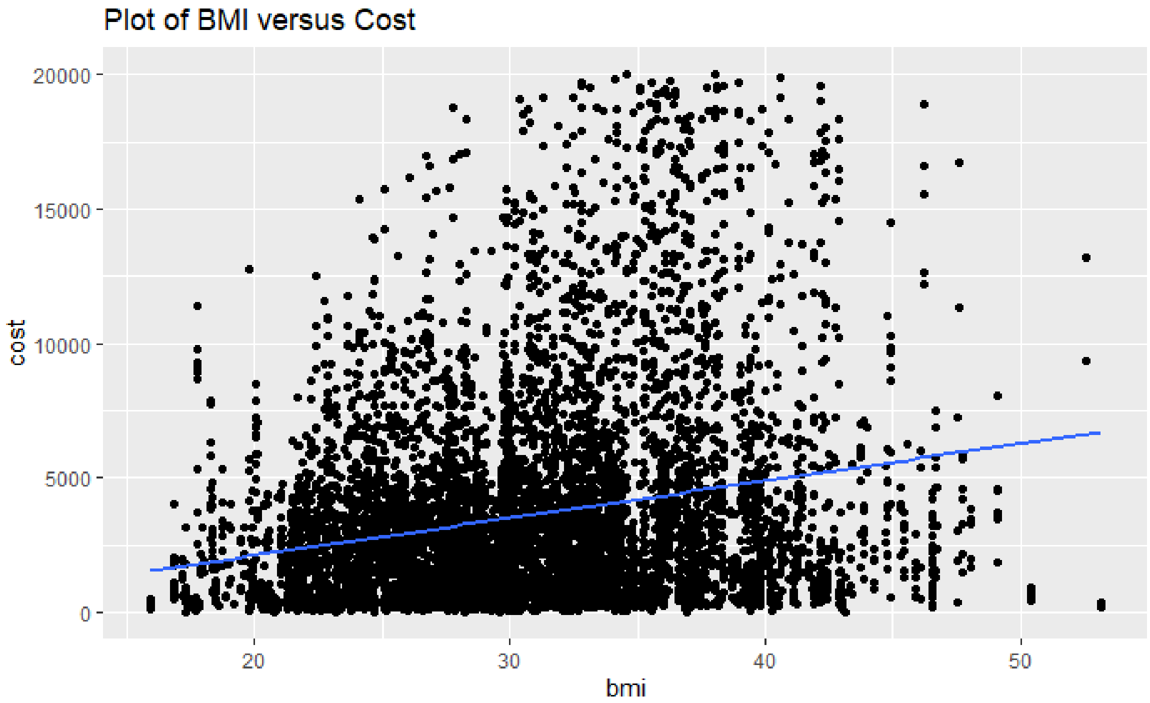
We classify the BMI with the cost variable, “yellow” color indicates inexpensive and green color indicates people are “expensive”. We see that the count of people who are healthy is higher when the person is inexpensive. Therefore bmi directly affects the cost.



We classify age with the cost variable, “yellow” color indicates inexpensive and green color indicates people are “expensive”. Young adults are high in count and are inexpensive whereas, senior citizens are more expensive.

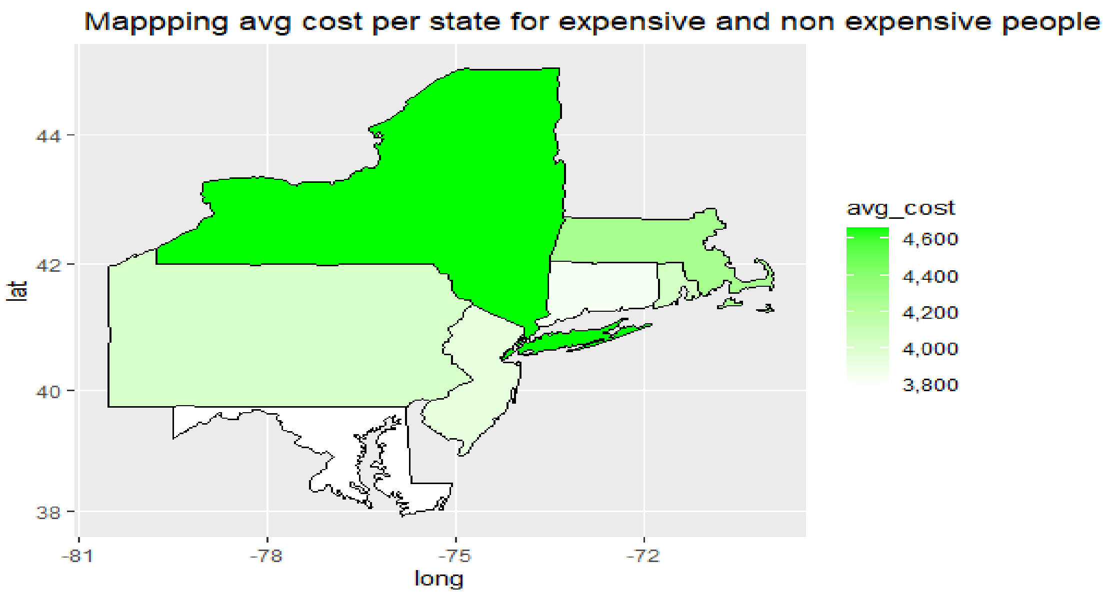
Bivariate Analysis

Scatterplot

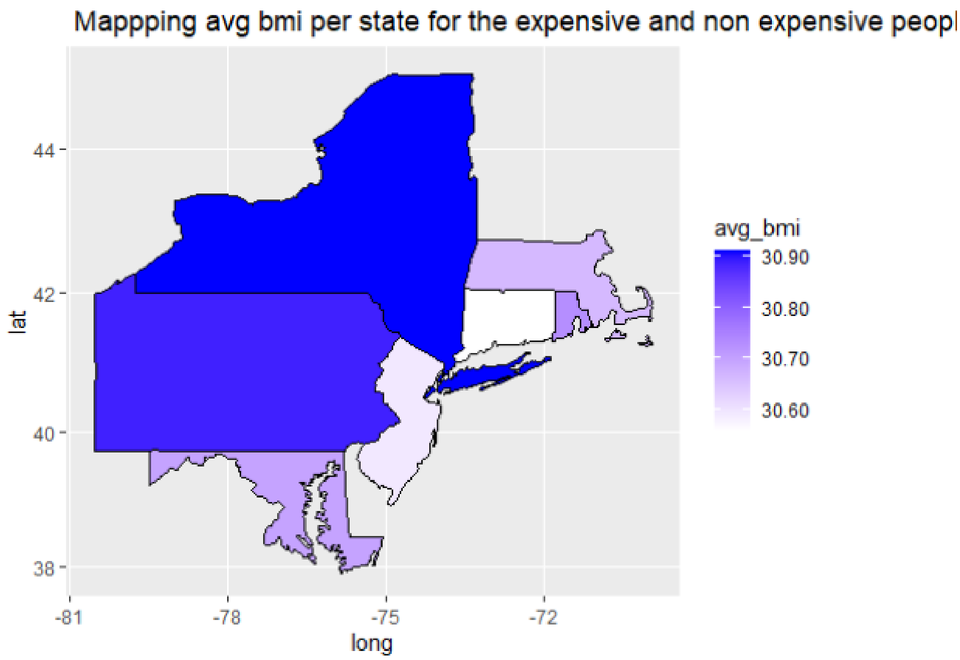


We plot bmi with the cost variable and see a direct linear relationship between bmi and cost. As the bmi increases the cost increases and therefore we clearly can state that bmi affects the health care cost of an individual

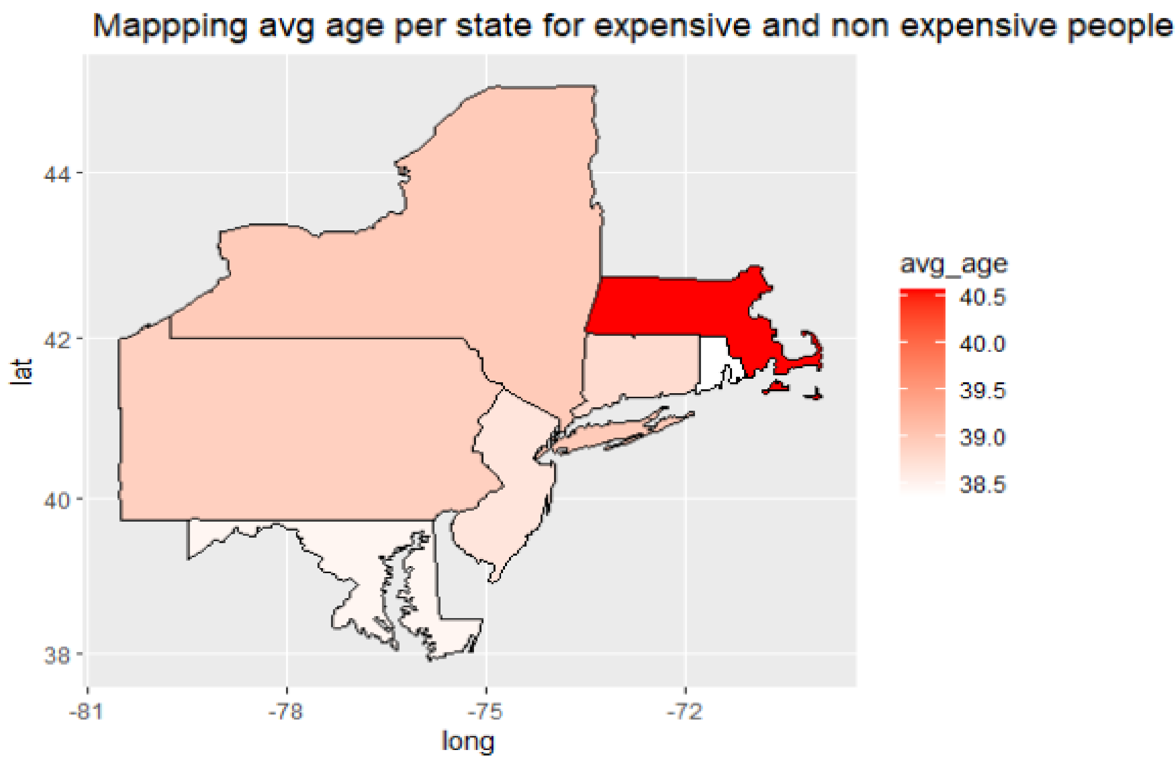
Multivariate Analysis



We map our health care data in the USA map and find key insights. The map shows us that the average cost per state is high for New York and second highest in Massachusetts. The average cost is lowest in Maryland and Connecticut. Therefore we can clearly see that New York has high cost for health care and avg cost is the highest in New York. The city of Boston also has the second highest average cost.



The map shows us that the average bmi per state is high for New York and Pennsylvania and lowest for Connecticut. This clearly indicates that people in New York are not healthy and fall in the out of range of bmi values. People in Connecticut have a low bmi and therefore they are healthy and have low health care cost.



Avg age is highest in the state of Massachusetts and second highest in New York and lowest in Rhode island. People in Massachusetts are older as compared to people in other states and they spend more on healthcare as indicated by maps earlier.

Model Building

Several models were used to perform predictive analysis on the data in order to see what were the significant attributes that lead to a person being classified as expensive. Each model uses a different method to identify key factors in their order of importance and a confusion matrix is also formed on the basis of the predictions.

To start the modeling process the data was first cleaned and processed. All categorical variables along with cost were converted to numerical variables to make it easier for the models to run. Splitting the cost variable on the 75th percentile converts the task of predicting people into a binary classification problem. All variables were stored into a new data frame df which was then used to run all the models. The data was then partitioned into a test set and a train set to train and test the models. The train-test split ratio is 75:25.

Note : The original cost variable must be removed from df, otherwise the data is overfitted

Linear Multiple Regression Model

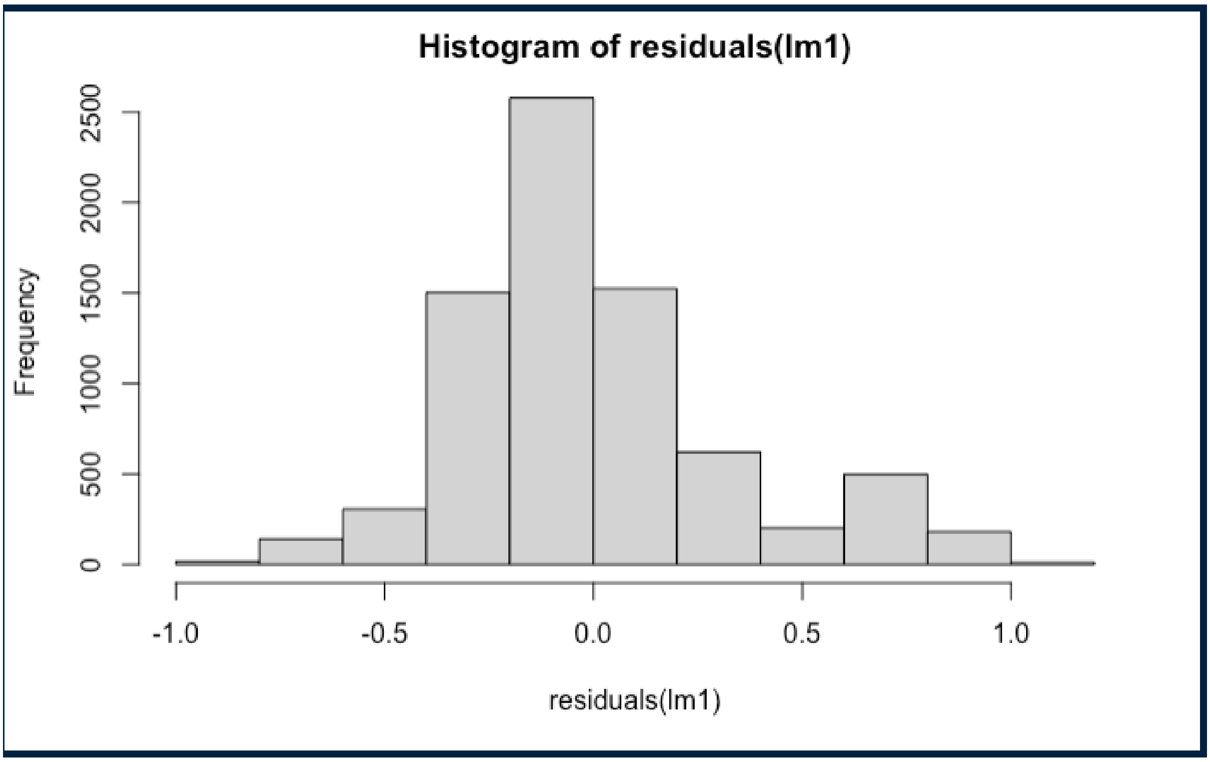
The first model we ran was the Linear Multiple regression model. The multiple regression model allows us to find what predictors are statistically significant from a list of independent variables given in the data set.

```
Call:
lm(formula = cost ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.94586 -0.20516 -0.05825  0.12797  1.14842 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.4984674  0.0395134 -37.923  < 2e-16 ***
age              0.0073962  0.0002679  27.610  < 2e-16 ***
bmi              0.0126012  0.0006349  19.849  < 2e-16 ***
children         0.0115141  0.0031046   3.709 0.000210 ***
smoker           0.5946685  0.0095505  62.266  < 2e-16 ***
location         0.0006767  0.0020519   0.330 0.741566    
location_type   -0.0098348  0.0087017  -1.130 0.258420    
education_level -0.0001466  0.0038113  -0.038 0.969324    
yearly_physical  0.0216209  0.0087263   2.478 0.013246 *  
exercise         0.1688931  0.0087212  19.366  < 2e-16 ***
married          0.0083760  0.0080070   1.046 0.295555    
hypertension     0.0349432  0.0094486   3.698 0.000219 ***
gender           0.0148411  0.0075911   1.955 0.050612 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3282 on 7569 degrees of freedom
Multiple R-squared:  0.4265,	Adjusted R-squared:  0.4256 
F-statistic: 469.2 on 12 and 7569 DF,  p-value: < 2.2e-16
```

As the name suggests, the linear multiple regression model helps analyze linear relationships between dependent and independent variables.

According to our model, age, body mass index, number of children, smoking status, exercise routine and hypertension are the statistically significant variables which can be used to predict the health expense of an individual based on the dataset provided by the HMO. The alpha level considered for our study was 0.001 and p-value less than the alpha level lets us reject the null hypothesis. The null hypothesis in this case is that there is no linear relationship between the dependent and independent variables. The adjusted R-squared value shows us that the independent variables

accounted for 42.67% variability in the dependent variable. The p-value at the very bottom of the model shows that this model is statistically significant; rejecting the null hypothesis that R-squared is 0. The histogram of residuals of the model appears



to be normal and centered at 0 which infers that there are no underlying non linear relationships in our data.

Now that we know the significant predictors, we created a simple linear regression model with just those predictors to see if we get a better model with a better R-squared value. However this model gave us a lower R-squared value (0.4232) so we rejected it.

We ran several simple linear regression models to see which independent variables account for the most variability in the dependent variable. The R-squared values are reported in the following table.

Model	Adjusted R-squared
cost ~ smoker	0.2981
cost ~ age	0.06991
cost ~ bmi	0.0386

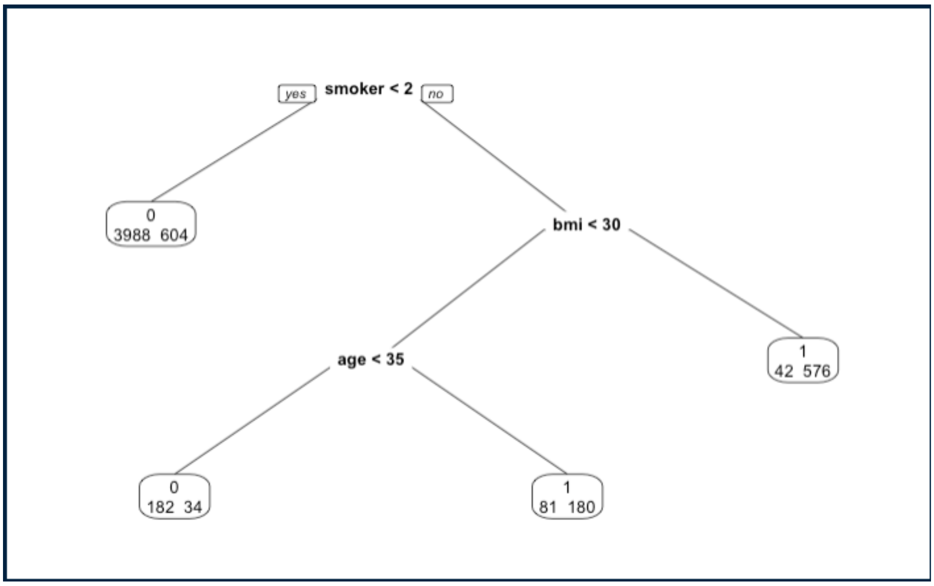
cost ~ exercise	0.02408

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1322  211
         1   72  290

               Accuracy : 0.8507
                 95% CI : (0.8338, 0.8664)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5786

 Mcnemar's Test P-Value : 2.34e-16

            Sensitivity : 0.9484
            Specificity : 0.5788
         Pos Pred Value : 0.8624
         Neg Pred Value : 0.8011
             Prevalence : 0.7356
         Detection Rate : 0.6976
   Detection Prevalence : 0.8090
      Balanced Accuracy : 0.7636

       'Positive' Class : 0
```

We ran the simple regression model with the training data and used the test set to get a confusion matrix

The measurements of significance for us are the accuracy and sensitivity. The confusion matrix reported an accuracy of 85.07% and sensitivity 94.84%.

Tree Bag Model



The next model we used was a Tree Bag model. Bagging models are generally used to achieve a higher accuracy in classification problems. Tree bag algorithm uses multiple subsets of training data to construct a final aggregated model with the best accuracy.

Each node shows the predicted class. If a person is a smoker the cost would be 3988,604. For a non-smoker, the model first checks if the body mass index is greater than 30. If it is then the cost is 42,576, otherwise the model checks the age. If age is less than 35 the cost is 182,34 and if age is over 35 the cost is 81,180

Running the varImp function gives us the most important variables considered by the model

	Overall <dbl>
bmi	100.0000000
age	93.4208907
smoker	54.9893379
exercise	25.2835446
location	22.8062861
children	20.6775534
education_level	13.2378651
gender	2.3296319
married	1.1912997
yearly_physical	0.4075212

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1324  142
         1   70  359

               Accuracy : 0.8881
                 95% CI : (0.8731, 0.902)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6985

 Mcnemar's Test P-Value : 1.081e-06

            Sensitivity : 0.9498
            Specificity : 0.7166
         Pos Pred Value : 0.9031
         Neg Pred Value : 0.8368
             Prevalence : 0.7356
         Detection Rate : 0.6987
   Detection Prevalence : 0.7736
      Balanced Accuracy : 0.8332

       'Positive' Class : 0
```

We used this model with the test set to predict the cost and then form a confusion matrix

The treebag model achieved a better accuracy than the linear model with the given training data and a slightly higher sensitivity.

Support Vector Machine Model

Support vector machines is a supervised learning technique used especially for classification and regression problems. The model works by creating two support vectors above and below the main vector which divides the data into classes. The algorithm then minimizes the error to best fit the data. The SVM model ran with all the independent variables in the training dataset

	Overall <dbl>
smoker	100.00000000
age	23.57877920
bmi	14.56006341
exercise	7.80266838
gender	1.78182397
children	0.85696299
hypertension	0.38021220
location_type	0.11558424
education_level	0.08910050
yearly_physical	0.07339697

The varImp function gave us all the independent variables that the SVM model considered to be the most significant. We used the test set to get the confusion

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1362  232
         1   32  269

               Accuracy : 0.8607
                 95% CI : (0.8443, 0.876)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5893

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

            Sensitivity : 0.9770
            Specificity : 0.5369
         Pos Pred Value : 0.8545
         Neg Pred Value : 0.8937
             Prevalence : 0.7356
         Detection Rate : 0.7187
   Detection Prevalence : 0.8412
      Balanced Accuracy : 0.7570

       'Positive' Class : 0
```

matrix.

As we can see, the SVM model gave us the highest accuracy (86.07%) and sensitivity (97.70%) among all the models we considered. Thus we used the SVM as our final model.

Significant predictors by model

	treebag Model	Linear model (cost~.)	SVM Model
1	**bmi**	**smoker**	**smoker**
2	**age**	**age**	**age**
3	**smoker**	**bmi**	**bmi**
4	**exercise**	**exercise**	**exercise**

Model Confusion

	treebag	Linear model (cost~.)	Linear model (cost~significant)	SVM
Accuracy	88.81%	85.07%	85.07%	86.07%
Sensitivity	94.98%	94.84%	94.84%	97.70%

Numbers and Insights

Insights from the data

- 27% men were expensive while only 21% women were expensive
- Average cost per patient was about $4043
- Average BMI for expensive patients was 32.83
- Average BMI for inexpensive patients was 30.11
- Average age for expensive patients was between 36 and 37
- Average age for non expensive patients was 45

NOTE : the average BMI in the dataset was 30.79. According to the standard BMI scale this lies in the first degree of obesity. Thus a lot of patients were overweight according to the dataset.

Insights from the models

- The most significant predictor of cost was the smoking habits of the patient followed by their age, BMI and their exercise regiment

Analysis and Recommendations

For Smokers

- Smoking is the most significant predictor of high healthcare costs for patients. As smoking affects the health and the health care cost the HMO can run quit smoking campaigns along with providing nicotine replacement therapy and counseling sessions to patients

For patients with high BMI

- When BMI is high, healthcare cost increases, we therefore recommend the HMO to counter this problem by offering sessions with dietitians and having a tie up with fitness clubs which specialize in overweight patients.

For people who don't exercise

- We found non active people in the data have more health care costs, and therefore we recommend the HMO start daily exercise programs for patients such as walking, jogging, yoga would significantly reduce the health care cost.

For older people

- According to the data, as people grow older their healthcare costs increase significantly. Some of these costs are unavoidable but following a good exercise routine and eating right can help overcome a lot of health problems. So we would recommend the HMO to create some self-care programs for older adults.

Shiny App

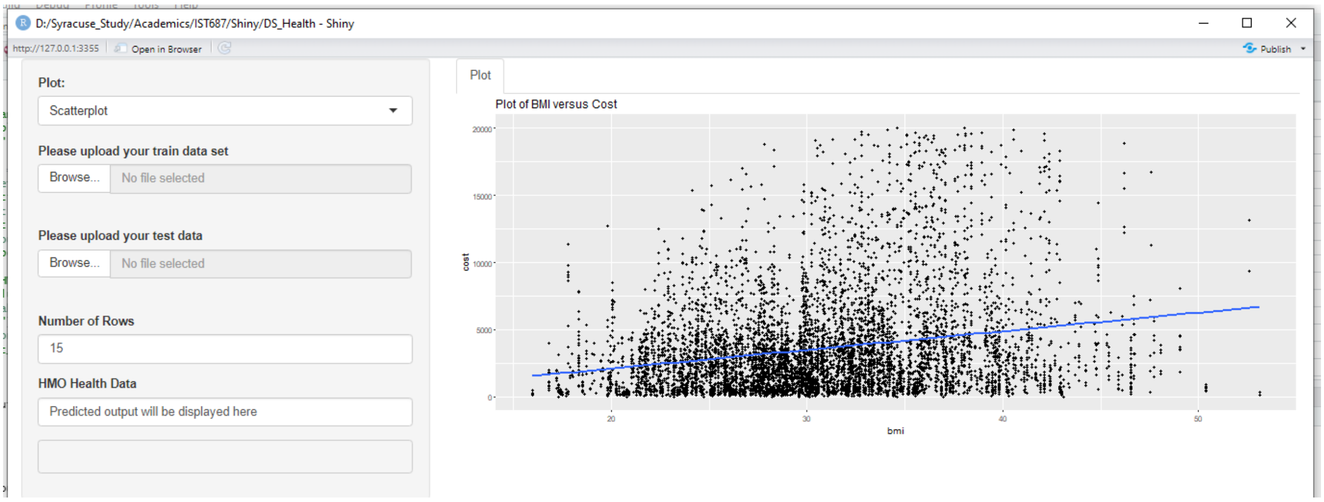
Shiny is an R package that makes it easy to build interactive web apps straight from R. You can host standalone apps on a webpage or embed them in R Markdown documents or build dashboards.

We have developed a Dashboard on Shiny App which gives the following two outputs:

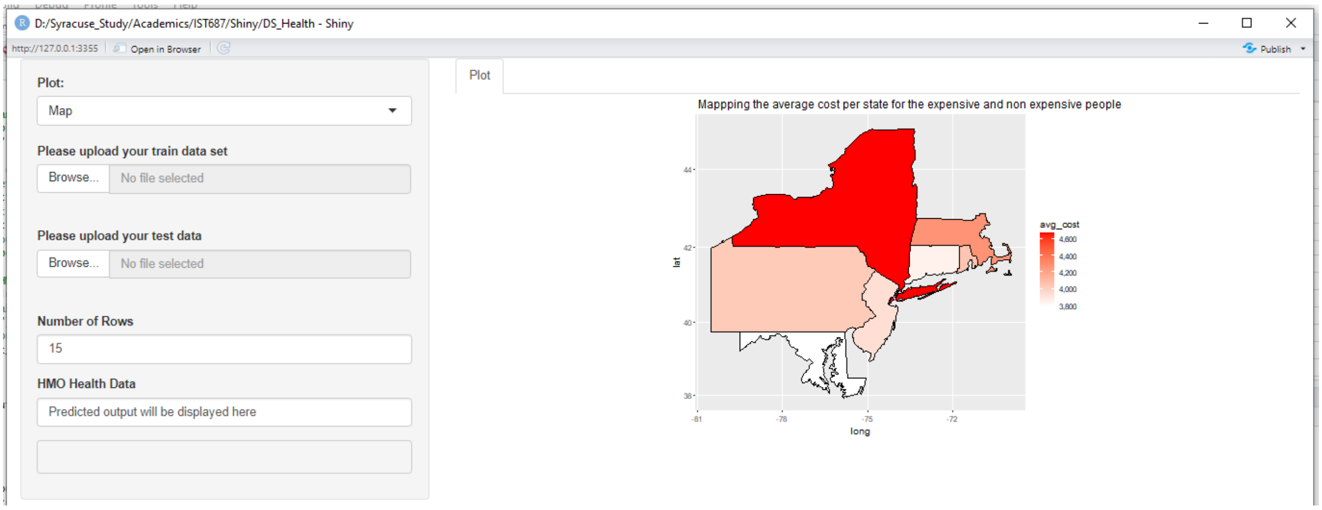
i) Presents the HMO_data in visual form (Histogram, Scatterplot, Boxplot and Map).

ii) Takes trainset and testset from the user and displays the output of the test data.

- Data Visualization

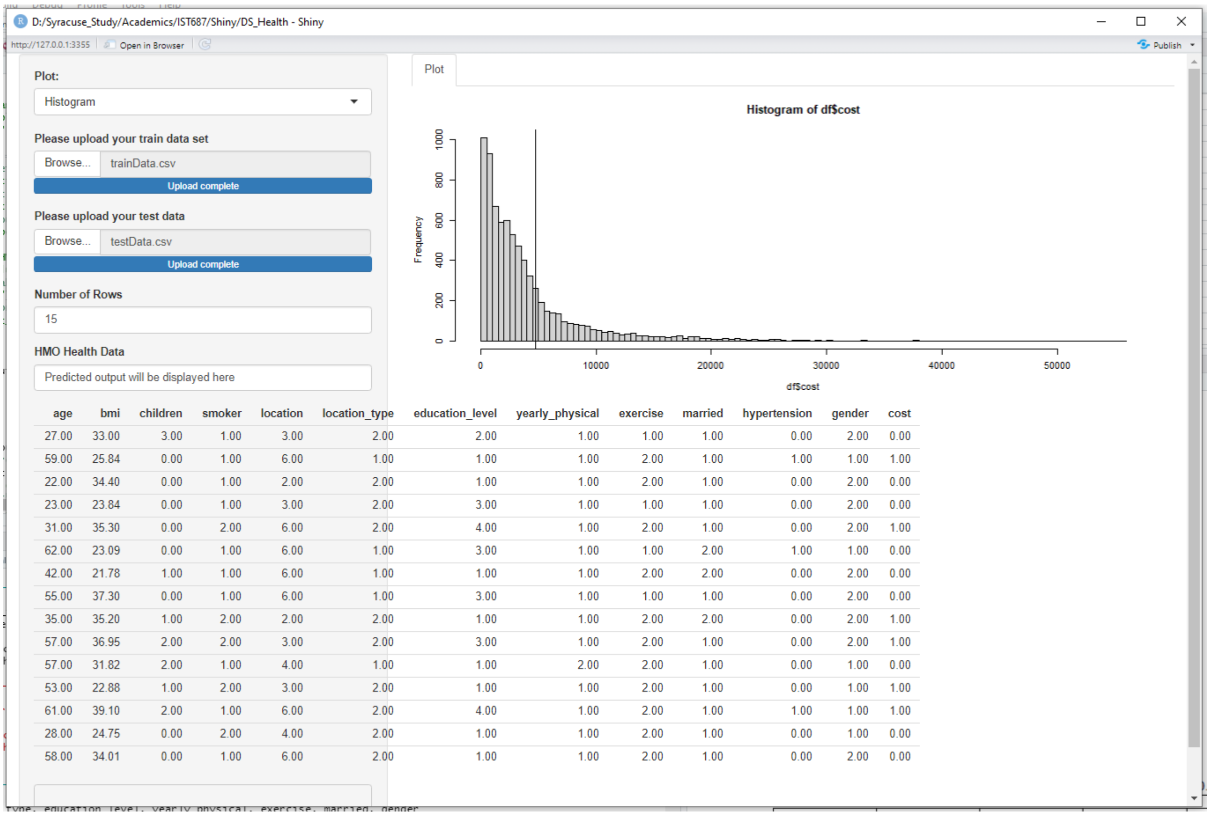


Scatterplot



Map

Taking trainset and testset from user and displaying the output



age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
27.00	33.00	3.00	1.00	3.00	2.00	2.00	1.00	1.00	1.00	0.00	2.00	0.00
59.00	25.84	0.00	1.00	6.00	1.00	1.00	1.00	2.00	1.00	1.00	1.00	1.00
22.00	34.40	0.00	1.00	2.00	2.00	1.00	1.00	2.00	1.00	0.00	2.00	0.00
23.00	23.84	0.00	1.00	3.00	2.00	3.00	1.00	1.00	1.00	0.00	2.00	0.00
31.00	35.30	0.00	2.00	6.00	2.00	4.00	1.00	2.00	1.00	0.00	2.00	1.00
62.00	23.09	0.00	1.00	6.00	1.00	3.00	1.00	1.00	2.00	1.00	1.00	0.00
42.00	21.78	1.00	1.00	6.00	1.00	1.00	1.00	2.00	2.00	0.00	2.00	0.00
55.00	37.30	0.00	1.00	6.00	1.00	3.00	1.00	1.00	1.00	0.00	2.00	0.00
35.00	35.20	1.00	2.00	2.00	2.00	1.00	1.00	2.00	2.00	0.00	2.00	1.00
57.00	36.95	2.00	2.00	3.00	2.00	3.00	1.00	2.00	1.00	0.00	2.00	1.00
57.00	31.82	2.00	1.00	4.00	1.00	1.00	2.00	2.00	1.00	0.00	1.00	0.00
53.00	22.88	1.00	2.00	3.00	2.00	1.00	1.00	2.00	1.00	0.00	1.00	1.00
61.00	39.10	2.00	1.00	6.00	2.00	4.00	1.00	2.00	1.00	1.00	1.00	1.00
28.00	24.75	0.00	2.00	4.00	2.00	1.00	1.00	2.00	1.00	0.00	1.00	0.00
58.00	34.01	0.00	1.00	6.00	2.00	1.00	1.00	2.00	1.00	0.00	2.00	0.00

Conclusion

The analysis of data gathered by the Health Management Organisation about their patients revealed some key information about why some people are more expensive than others. The most significant factors that lead to high healthcare costs included a patient's age, smoking habits, body mass index and their exercising habits. The data also revealed that the average body mass index of the patients was quite high. Using the models we created, the HMO can predict whether a client is going to be expensive or not based on their habits and some information about their lifestyle. To reduce cost of existing clients, we recommend the HMO start various health, fitness and quit smoking programs.