Syracuse University
School of Information Studies

HEALTHCARE EXPENSE ANALYSIS

Shubh Mody
Kunjan Ashish Chauhan
Yixing (Leo) Zhu
Nicholas Lukowsky

Introduction

Healthcare is an important part of everyone’s life. However, healthcare often costs a lot and even become affordable to some people. Thus, it is important to help people reduce healthcare cost. The dataset includes healthcare customer cost information from an HMO (Health Management Organization). In order to reduce healthcare cost, it is necessary to first predict and understand why and what groups of people are more expensive in terms of health care cost.

Project Goal

Our goal is to provide insights from business mindset. We used different models and data analysis techniques, looking to find common trends between the individual customer characteristics and healthcare cost.

Business Questions:

- Why are some customer's healthcare costs more expensive than others?
- What are the factors that cause a customer's healthcare expenses to increase or decrease?
- What can HMO do to reduce their expenditure on customers?

Data Characteristics

The raw dataset contains healthcare cost information for an HMO (Health Management Organization). It has 7,582 rows (observations), 14 columns (variables) for healthcare customers ranging from age 25 to 66.

	X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
1	1	18	27.900	0	yes	CONNECTICUT	Urban	Bachelor	No	Active	Married	0	female	1746
2	2	19	33.770	1	no	RHODE ISLAND	Urban	Bachelor	No	Not-Active	Married	0	male	602
3	3	27	33.000	3	no	MASSACHUSETTS	Urban	Master	No	Active	Married	0	male	576
4	4	34	22.705	0	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	1	male	5562
5	5	32	28.880	0	no	PENNSYLVANIA	Country	PhD	No	Not-Active	Married	0	male	836
6	7	47	33.440	1	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0	female	3842
7	9	36	29.830	2	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	1304
8	10	59	25.840	0	no	PENNSYLVANIA	Country	Bachelor	No	Not-Active	Married	1	female	9724
9	11	24	26.220	0	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	201
10	12	61	26.290	0	yes	CONNECTICUT	Urban	No College Degree	No	Active	Married	0	female	4492

Before implementing the data analysis, we found some variables that might be closely related to the healthcare cost and raised some questions:

Age:

- How one's health care cost is affected by his/her age?

BMI:

- Whether BMI (body mass index) plays an important role in determining one's health care cost?

Smoker:

- How one's lifestyle habits affect how much they spend on healthcare?

Exercise:

- Whether the activity level of a customer influences their healthcare expenses?

Location:

- Whether the healthcare expenditure relates to the state where someone lives?

Data Cleaning and Processing

Data Cleaning:

We first checked for missing values, and we did found 78 missing values in "bmi " and 80 missing values in "hypertension". Then we used the "imputeTS" R package to implement missing value imputation on "bmi" but replaced missing values on "hypertension" with 0 directly so that there is no missing value in dataset anymore.

```
     age   bmi children smoker location     locat…¹ educa…² yearl…³ exerc…⁴ married hyper…⁵
   <dbl> <dbl>    <dbl> <chr>  <chr>        <chr>   <chr>   <chr>   <chr>   <chr>     <dbl>
 1    19    NA        0 no     PENNSYLVA…   Urban   No Col… No      Active  Not_Ma…       0
 2    35    NA        1 yes    RHODE ISL…   Urban   Bachel… Yes     Not-Ac… Married       0
 3    19    NA        0 no     MARYLAND     Urban   PhD     Yes     Not-Ac… Not_Ma…       0
 4    19    NA        0 no     PENNSYLVA…   Urban   Master  No      Not-Ac… Married       0
 5    59    NA        0 no     PENNSYLVA…   Country Master  No      Active  Not_Ma…       0
 6    26    NA        0 no     MARYLAND     Country Bachel… No      Active  Married       0
 7    19    NA        0 no     PENNSYLVA…   Urban   Master  No      Not-Ac… Married       0
 8    19    NA        0 no     MARYLAND     Urban   Master  No      Not-Ac… Married       1
 9    37    NA        0 no     PENNSYLVA…   Urban   Bachel… No      Not-Ac… Married       0
10    57    NA        0 no     RHODE ISL…   Country Bachel… Yes     Not-Ac… Married       0
# … with 68 more rows, 2 more variables: gender <chr>, cost <dbl>, and abbreviated
```

```
     age   bmi children smoker location     locat…¹ educa…² yearl…³ exerc…⁴ married hyper…⁵
   <dbl> <dbl>    <dbl> <chr>  <chr>        <chr>   <chr>   <chr>   <chr>   <chr>     <dbl>
 1    42  39.5        0 no     MASSACHUS…   Urban   Bachel… Yes     Not-Ac… Married      NA
 2    34  33.7        1 no     PENNSYLVA…   Country Master  Yes     Not-Ac… Not_Ma…      NA
 3    48  30.2        2 no     NEW YORK     Urban   Bachel… No      Not-Ac… Married      NA
 4    20  20.7        0 no     NEW YORK     Country Bachel… No      Not-Ac… Married      NA
 5    35  33.2        1 no     CONNECTIC…   Country Bachel… Yes     Active  Not_Ma…      NA
 6    34  27          2 no     NEW YORK     Country Master  No      Not-Ac… Married      NA
 7    60  36.1        3 no     MARYLAND     Country Bachel… No      Not-Ac… Married      NA
 8    36  29.0        4 no     NEW JERSEY   Urban   No Col… No      Not-Ac… Married      NA
 9    18  27.3        3 yes    NEW JERSEY   Country Bachel… Yes     Not-Ac… Not_Ma…      NA
10    29  37.3        2 no     PENNSYLVA…   Country No Col… No      Not-Ac… Not_Ma…      NA
# … with 70 more rows, 2 more variables: gender <chr>, cost <dbl>, and abbreviated
```

Data Processing:

We created age groups based on customers' ages. We defined customers younger than age 25 as “young adults”, customers from age 25 to 40 as “adults”, customers from age 40 to 55 as “older adults”, customers older than age 55 as “senior citizens”.

	age	ageGroup
1	18	young adults
2	19	young adults
3	27	adults
4	34	adults
5	32	adults
6	47	older adults
7	36	adults
8	59	senior citizens
9	24	young adults
10	61	senior citizens

We also defined the 25% most expensive customers as “expensive” and the remaining 75% as “inexpensive”. The threshold for an “expensive” customer is a cost of $4,775.

	cost	category
1	1746	inexpensive
2	602	inexpensive
3	576	inexpensive
4	5562	expensive
5	836	inexpensive
6	3842	inexpensive
7	1304	inexpensive
8	9724	expensive
9	201	inexpensive
10	4492	inexpensive

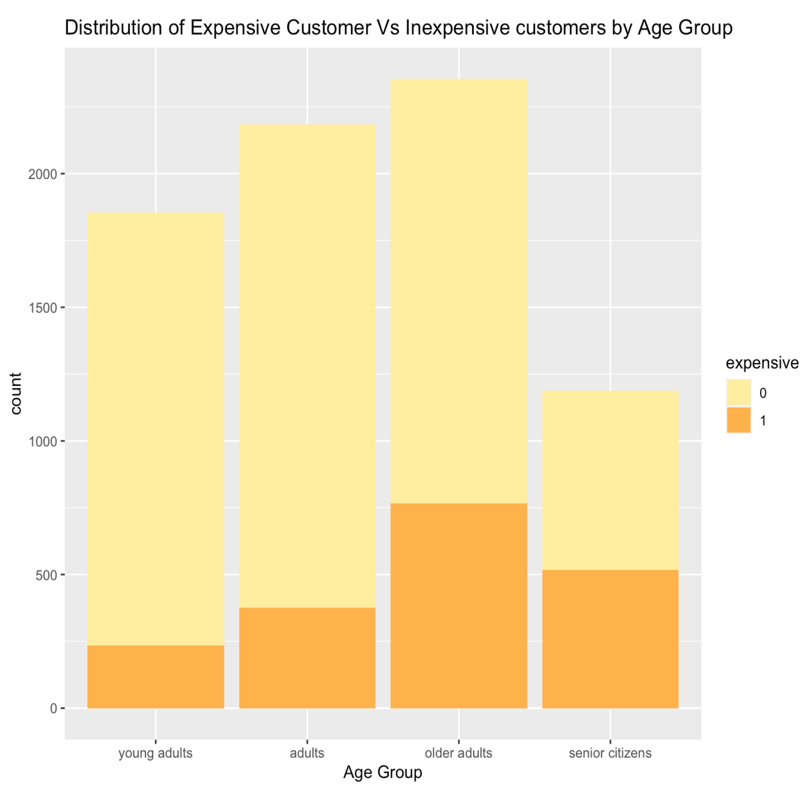
We then converted categorical variables, including “smoker”, “location”, “location_type”, “education_level”, “yearly_physical”, “exercise”, “married”, “gender”, into factors (digital values).

	smoker	location	location_type	education_level	yearly_physical	exercise	married	gender
1	yes	CONNECTICUT	Urban	Bachelor	No	Active	Married	female
2	no	RHODE ISLAND	Urban	Bachelor	No	Not-Active	Married	male
3	no	MASSACHUSETTS	Urban	Master	No	Active	Married	male
4	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	male
5	no	PENNSYLVANIA	Country	PhD	No	Not-Active	Married	male

	smoker	location	location_type	education_level	yearly_physical	exercise	married	gender
1	1	0	1	0	0	1	1	0
2	0	1	1	0	0	0	1	1
3	0	2	1	1	0	1	1	1
4	0	3	0	1	0	0	1	1
5	0	3	0	2	0	0	1	1

Data Visualization

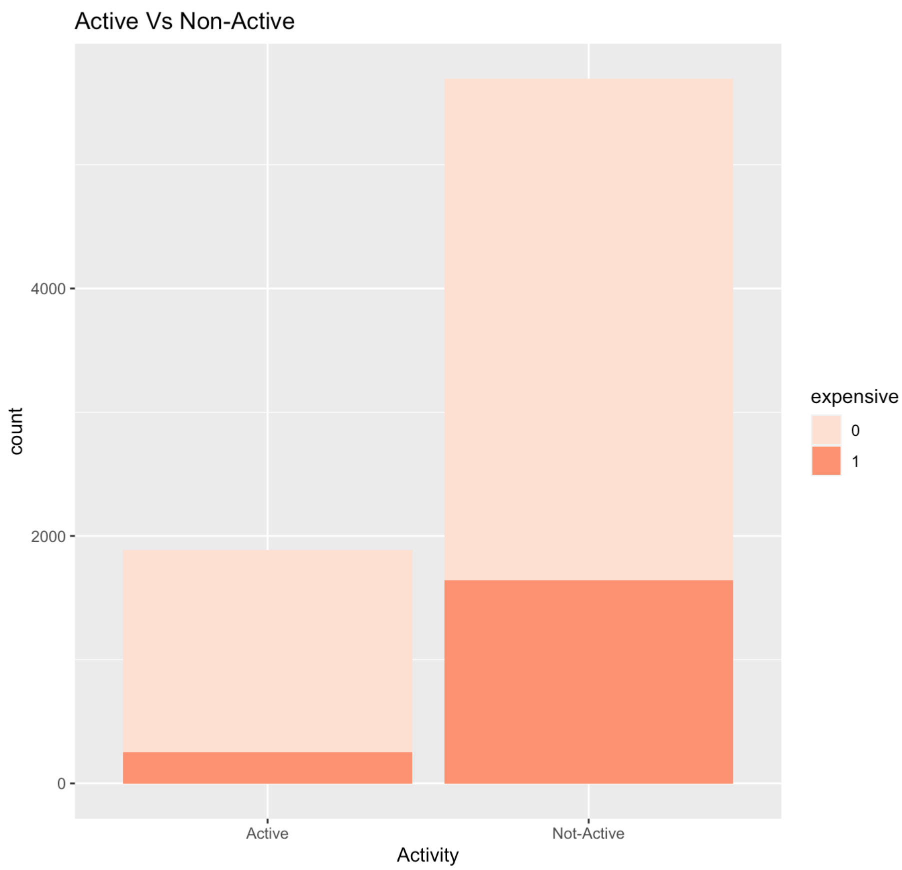
Healthcare Cost of Different Age Group:



- Threshold to consider a customer expensive set at $4,775 (Q3 value).

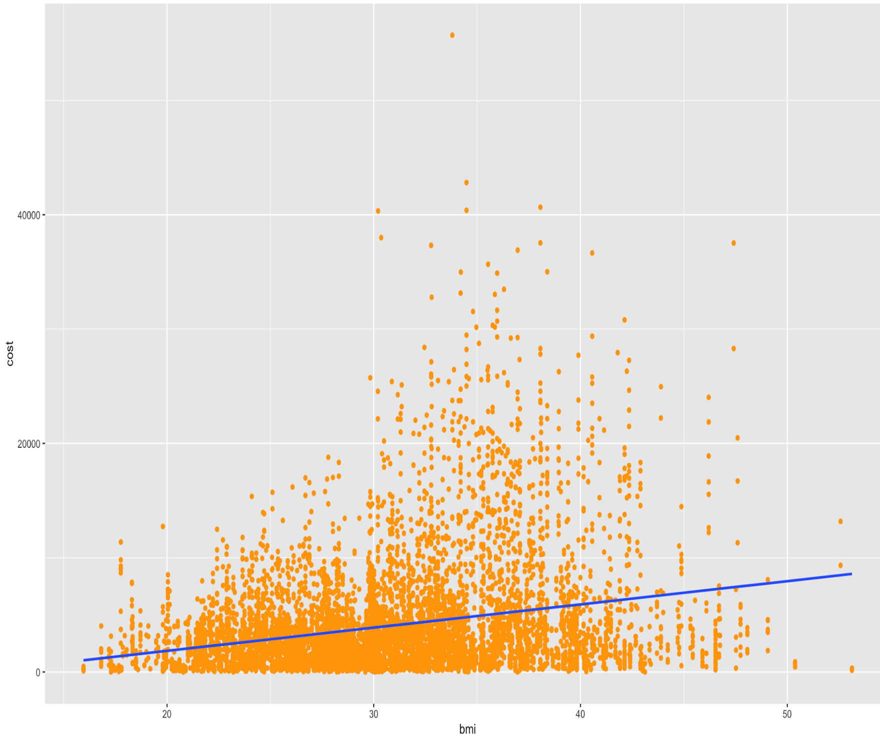
- The percentage of expensive customers increases with increase in the age group of the customers.
- Even though there are fewer data points for customers in the senior citizens category, we still see that their healthcare cost is far more than young adults and adults.

Healthcare Cost of Activity Levels:



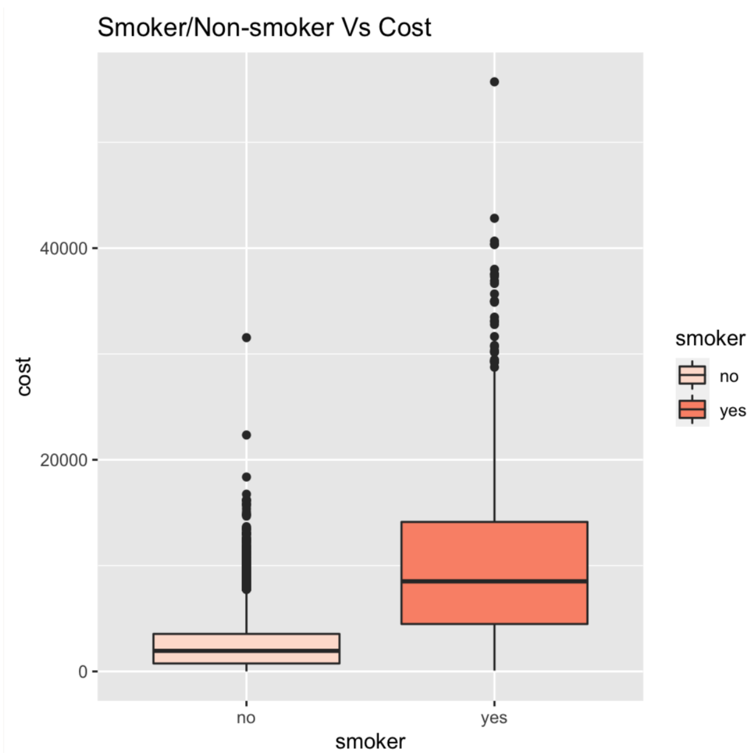
- Customers who are less active are more likely to have higher healthcare cost expenses.
- Fewer customers identify as active altogether, but they have a smaller percentage of health care costs that are classified as expensive.

Healthcare Cost Relation with BMI:



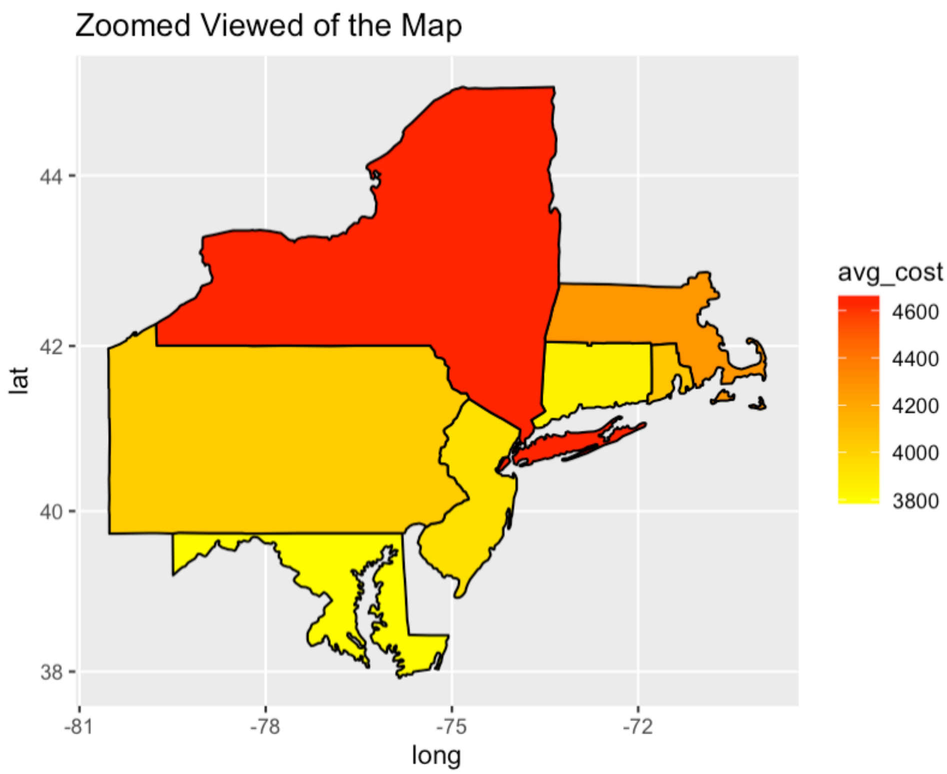
- As the BMI increases, customers tend to have more healthcare issues.
- High BMI usually means obesity and unhealthy lifestyle.

Healthcare Cost Relation with Smoking:



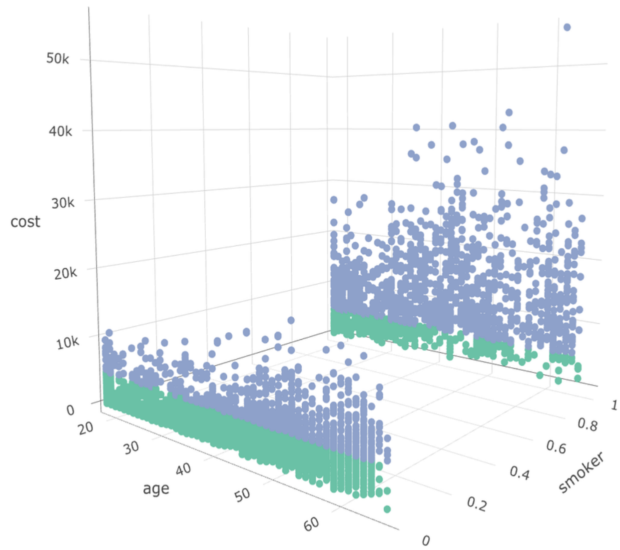
- If a customer is a smoker, it has a drastic impact on the cost of his/her healthcare.
- The median cost for smokers is 4 times that of non-smokers, as well as the group containing larger outliers.

Northeastern States Map of Average Cost:



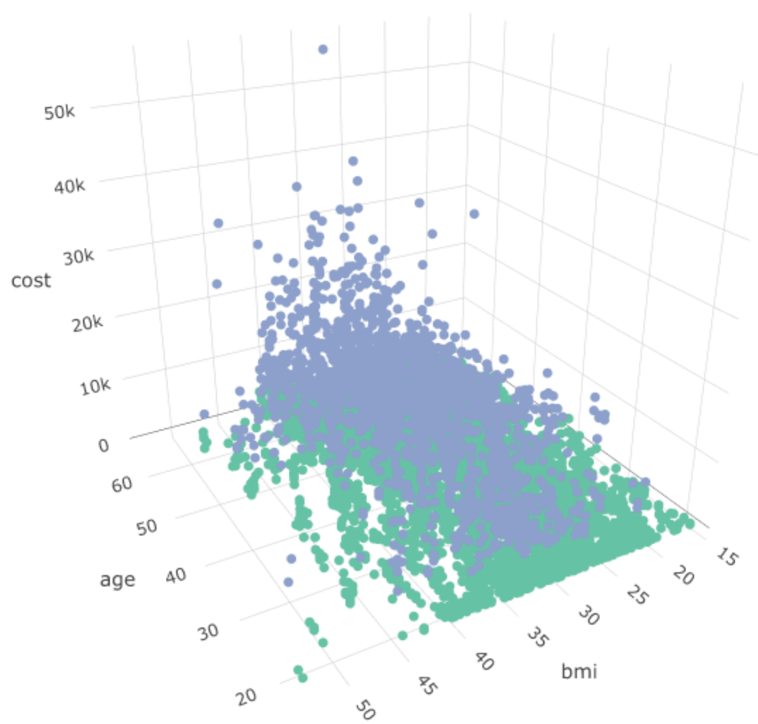
- For the states included in the dataset the average cost of healthcare is shown.
- Massachusetts and especially New York have higher cost while Connecticut and Maryland have lower cost.

Healthcare Cost Relation with Age & Smoking:



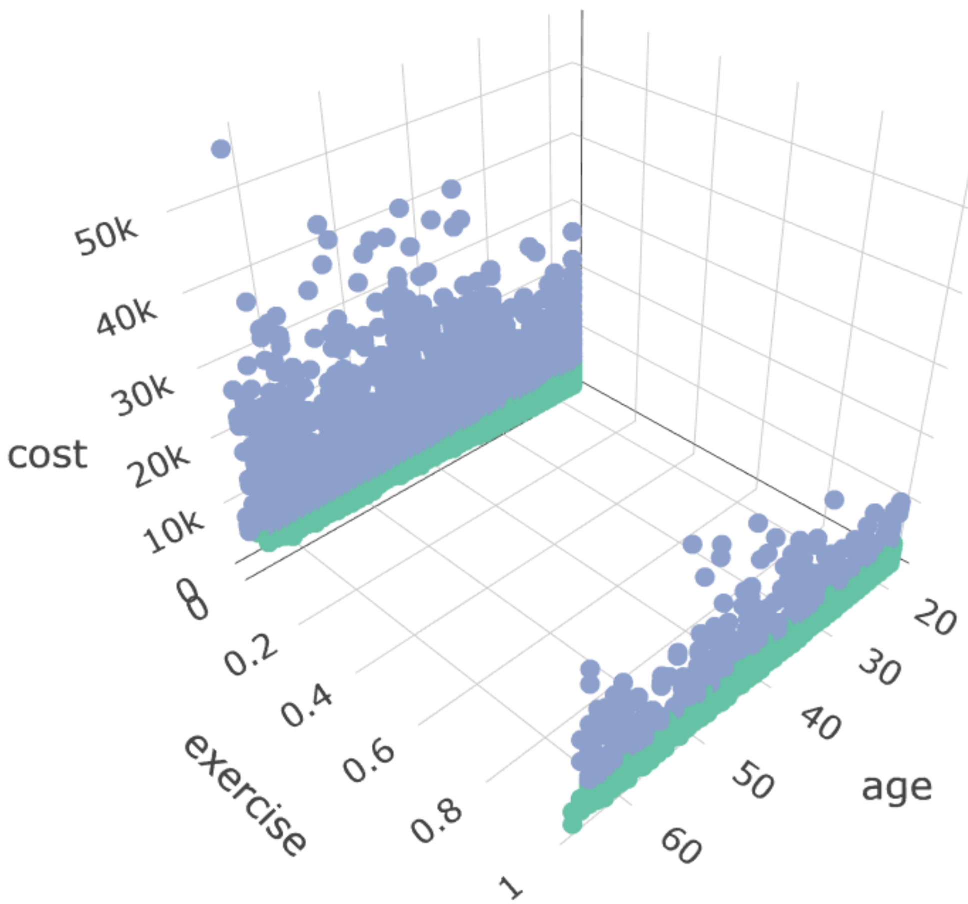
- Using a three-dimensional plot of age and smoker is shown against the healthcare cost.
- For non-smokers the cost is relatively low & constant as age increases, however for smokers the cost is generally larger and increases with age.

Healthcare Cost Relation with Age & BMI:



- Increasing BMI with increasing age has higher number of customers which are expensive.
- BMI alone with cost was quite scattered, but this graph shows BMI of 30 for age 20-30 doesn't affect cost by much. But the same BMI for age above 40 have higher cost.

Healthcare Cost Relation with Age & Exercise:



- Using a three-dimensional plot age and the categorical variable of exercise is shown against the healthcare cost.
- For active people the cost is relatively low & constant as age increases, however for non-active people the cost is generally larger.

Cost Prediction Using Linear Modeling

We first used all variables as predictors of the linear model to predict the cost, and got variable coefficients and other statistical values.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -6749.099    264.288 -25.537  < 2e-16 ***
age                102.453      2.630  38.954  < 2e-16 ***
bmi                181.388      6.232  29.105  < 2e-16 ***
children           232.944     30.478   7.643 2.38e-14 ***
smokeryes         7664.227     93.755  81.747  < 2e-16 ***
location1          114.829    178.206   0.644 0.519360    
location2            9.494    198.377   0.048 0.961830    
location3           17.048    139.987   0.122 0.903072    
location4         -130.369    175.782  -0.742 0.458322    
location5          113.071    194.552   0.581 0.561131    
location6          467.802    189.782   2.465 0.013726 *  
location_type1     -10.484     85.432  -0.123 0.902334    
education_level1   -97.075     95.118  -1.021 0.307488    
education_level2  -233.763    129.884  -1.800 0.071935 .  
education_level3    41.581    126.295   0.329 0.741986    
yearly_physical1   139.284     85.695   1.625 0.104130    
exercise1        -2263.242     85.628 -26.431  < 2e-16 ***
married1          -134.309     78.594  -1.709 0.087510 .  
hypertension       341.385     92.750   3.681 0.000234 ***
gender1             29.613     74.532   0.397 0.691141    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3220 on 7562 degrees of freedom
Multiple R-squared:  0.5742,    Adjusted R-squared:  0.5731 
F-statistic: 536.6 on 19 and 7562 DF,  p-value: < 2.2e-16
```

We then dropped these variables with p-values larger than 0.05 and used the remaining variables as predictors to implement the second linear modeling. However, there is no improvement for the second model in terms of model accuracy as the “Adjusted R-squared” values for both models are almost identical.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6779.989    215.864 -31.409  < 2e-16 ***
age           102.300      2.629  38.917  < 2e-16 ***
bmi           181.333      6.222  29.143  < 2e-16 ***
children      235.109     30.442   7.723 1.28e-14 ***
smokeryes    7666.709     93.442  82.048  < 2e-16 ***
exercise1   -2259.719     85.605 -26.397  < 2e-16 ***
hypertension  333.293     92.752   3.593 0.000328 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3222 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
```

Classification Using Machine Learning

We need to analyze the data and find answers to the below questions:

Machine learning is the science of getting computers to act by feeding them data and letting them learn a few tricks on their own, without being explicitly programmed to do so

The key to machine learning is the data. Machines learn just like us humans. We humans need to collect information and data to learn, similarly, machines must also be fed data in order to learn and make decisions.

Types of machine learning:

1. Supervised learning

Supervised means to oversee or direct a certain activity and make sure it's done correctly. In this type of learning the machine learns under guidance.

At school, our teachers guided us and taught us, similarly in supervised learning, you feed the model a set of data called training data, which contains both input data and the corresponding expected output. The training data acts as a teacher and teaches the model the correct output for a particular input so that it can make accurate decisions when later presented with new data.

2. Unsupervised learning

Unsupervised means to act without anyone's supervision or direction.

In unsupervised learning, the model is given a data set which is neither labeled nor classified. The model explores the data and draws inferences from data sets to define hidden structures from unlabeled data

An example of unsupervised learning is an adult like you and me. We don't need a guide to help us with our daily activities, we figure things out on our own without any supervision.

What is SVM?

SVM (Support Vector Machine) is a supervised machine learning algorithm which is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyperplane which acts like a decision boundary between the various classes.

SVM can be used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data.

Before moving further, let’s discuss the features of SVM:

SVM is a supervised learning algorithm. This means that SVM trains on a set of labeled data. SVM studies the labeled training data and then classifies any new input data depending on what it learned in the training phase.

A main advantage of SVM is that it can be used for both classification and regression problems. Though SVM is mainly known for classification, the SVR (Support Vector Regressor) is used for regression problems.

SVM can be used for classifying non-linear data by using the kernel trick. The kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data. After which you can easily draw a hyperplane between the various classes of data.

Our model performance:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1103  148
         1   34  231

               Accuracy : 0.8799
                 95% CI : (0.8625, 0.8959)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6442

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

            Sensitivity : 0.9701
            Specificity : 0.6095
         Pos Pred Value : 0.8817
         Neg Pred Value : 0.8717
             Prevalence : 0.7500
         Detection Rate : 0.7276
   Detection Prevalence : 0.8252
      Balanced Accuracy : 0.7898

       'Positive' Class : 0
```

Our SVM model’s Accuracy was 87.99% with a sensitivity of 97.01%

Decision trees in R

Decision Trees is useful supervised Machine learning algorithms that have the ability to perform both regression and classification tasks. It is characterized by nodes and branches, where the tests on each attribute are represented at the nodes, the outcome of this procedure is represented at the branches and the class labels are represented at the leaf nodes. Hence it uses a tree-like model based on various decisions that are used to compute their probable outcomes. These types of tree-based algorithms are one of the most widely used algorithms due to the fact that these algorithms are easy to interpret and use. Apart from this, the predictive models developed by this algorithm are found to have good stability and decent accuracy due to which they are very popular.

As it can be seen that there are many types of decision trees but they fall under two main categories based on the kind of target variable, they are

Categorical Variable Decision Tree: This refers to the decision trees whose target variables have limited value and belongs to a particular group.

Continuous Variable Decision Tree: This refers to the decision trees whose target variables can take values from a wide range of data types.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1113  164
         1   24  215

               Accuracy : 0.876
                 95% CI : (0.8583, 0.8922)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6229

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

            Sensitivity : 0.9789
            Specificity : 0.5673
         Pos Pred Value : 0.8716
         Neg Pred Value : 0.8996
             Prevalence : 0.7500
         Detection Rate : 0.7342
   Detection Prevalence : 0.8423
      Balanced Accuracy : 0.7731

       'Positive' Class : 0
```

Tree Model's performance accuracy was 87.6% with 97.89% sensitivity.

Advantages of the Decision tree:

Easy Interpretation

Making predictions is fast

Easy to identify important variables

Handless missing data

One of the drawbacks is to can have high variability in performance.

Recursive portioning- basis can achieve maximum homogeneity within the new partition.

Conclusion and Recommendations

We need to analyze the data and find answers to the below questions:

Why are some customers' healthcare costs more expensive than others?

What are the factors that cause a customer's healthcare expenses to increase or decrease?

What can HMOs do to reduce their expenditure on customers?

Having associated the attributes for expensive customers, we can be selective in attracting our future customers.

As a second option, we can increase monthly premiums for customers who are potentially expensive.

Younger people tend to have lower healthcare costs while older people tend to have higher costs. Based on this we can group customer age brackets and determine a higher monthly premium based on the age bracket.

Three factors affect the person's health (BMI, Smoking, Exercise).

A balanced, calorie-controlled diet is the ticket to a healthy BMI – the safe way.

Woww! Look at what you all can buy if you "Quit" smoking

STOP SMOKING, START LIVING

Smoking is an expensive habit. In Louisiana, the average cost of a pack of cigarettes is $5.85.
You may be surprised by how much you'll save in the weeks, months and years ahead.

26 packs = $150
Dinner out for two

43 packs = $250
Concert tickets for two

171 packs = $1,000
1-Year gym membership

85 packs = $500
Designer handbag

350 packs/
1-year span = $2,135
7-Day Caribbean
cruise for two

855 packs = $5,000
New motorcycle

2,051 packs = $12,000
New car

3,650 packs/
10-year span = $21,000
Home renovation