IST 687

Healthcare Cost Prediction

Produced By:-

Yashanjali Chavan

Matthew Penn

Immanuel Odarteifio

Sahil Wani

Abirami Rajalingam

TABLE OF CONTENTS

Introduction

One of the most typical recurring expenses in a person's life is medical bills. BMI, aging, smoking, and other conditions have all been linked to higher expenses for personal medical care in various study studies. To help develop measures for preventing obesity that are both affordable and successful, estimates of the costs of healthcare connected to obesity and other such important conditions are required. Preventing young age obesity, and encouraging people to adopt an active lifestyle at a young age is a key priority in public health, clinical practice, and global health.

Dataset contains information about various factors like age, BMI, smoking, exercise, marriage status, number of children of patients in 7 states of America. We can predict the healthcare cost for people using certain traits from our dataset which can help us analyze who is likely to spend more money on their healthcare for next year.

Our goal is to give the HMO specific advice on how to reduce health care expenditures in order to give them concrete information into how to minimize their overall health care costs.

Business Understanding

For this project, we wanted to look at the variables that correlate with customers being expensive for a Health Management Organization (HMO). To do this, we utilized a dataset containing information on customers, including how much their health insurance cost. Using this data, we build various models to understand what factor seemed to lead to higher costs in order to identify ways to lower costs in those areas.

Business Questions

- ➢ Is there a disparity between gender and the cost of healthcare?
- ➢ Do older people, who are more prone to hospitalization, have a large healthcare cost?
- ➢ What's the mean cost per BMI?
- ➢ Does a combination of high BMI & hypertension contribute to expensive healthcare?
- ➢ Do regular exercise contribute to a healthier lifsestyle which in turn reduce the health cost?
- ➢ Will the location type have a direct impact on health?
- ➢ Do Active Smoker spend more money on health care than the rest?
- ➢ Is pediatrics expensive than the general health expense?

Technical Details

There are 7582 rows and 13 columns in our dataset.

Here are the variables that we found in our data file:

- 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.

Based on the cost we created a column named expensive which will determine whether the particular person has a expensive health care or not. This was achieved by taking the third quartile i.e the 75th percentile, if the cost is greater than the 75th percentile ($ 4775) then that person has expensive health care or spend more money on the health care

Goal of our analysis

Using prediction algorithms like linear model, random forest regression, generalized linear model and support vector machines, we aim to provide best actionable insights to Health Management Organization in terms of how to lower a customer’s healthcare costs.

Data Acquisition

```
> glimpse(RawhosData)
Rows: 7,582
Columns: 15
$ X               <int> 1, 2, 3, 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 32, 33, 34, 35, 37, 39, 40, 41, 42,…
$ age             <int> 18, 19, 27, 34, 32, 47, 36, 59, 24, 61, 22, 57, 26, 18, 23, 57, 31, 60, 30, 19, 32, 37, 58, 62, 56, 31, 19, 20, 63, 28, 63, 35, 58, …
$ bmi             <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 33.440, 29.830, 25.840, 26.220, 26.290, 34.400, 39.820, 42.130, 24.600, 23.845, 40.300, 35.3…
$ children        <int> 0, 1, 3, 0, 0, 1, 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 2, 3, 0, 2, 2, 0, 5, 0, 1, 3, 1, 0, 0, 2, 1, 2, 1, 0, 2, 0, 0, 1, 2, …
$ smoker          <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no", "yes", "no", "no", "yes", "no", "no", "no", "yes", "no", "no", "no", "yes", "…
$ location        <chr> "CONNECTICUT", "RHODE ISLAND", "MASSACHUSETTS", "PENNSYLVANIA", "PENNSYLVANIA", "PENNSYLVANIA", "PENNSYLVANIA", "PENNSYLVANIA", "PEN…
$ location_type   <chr> "Urban", "Urban", "Urban", "Country", "Country", "Urban", "Urban", "Country", "Urban", "Urban", "Urban", "Urban", "Urban", "Country"…
$ education_level <chr> "Bachelor", "Bachelor", "Master", "Master", "PhD", "Bachelor", "Bachelor", "Bachelor", "Bachelor", "No College Degree", "Bachelor", …
$ yearly_physical <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "No", "No", "Yes", "N…
$ exercise        <chr> "Active", "Not-Active", "Active", "Not-Active", "Not-Active", "Not-Active", "Active", "Not-Active", "Active", "Active", "Not-Active"…
$ married         <chr> "Married", "Married", "Married", "Married", "Married", "Married", "Married", "Married", "Married", "Married", "Married", "Married", …
$ hypertension    <int> 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ gender          <chr> "female", "male", "male", "male", "male", "female", "male", "female", "male", "female", "male", "female", "male", "male", "male", "m…
$ cost            <int> 1746, 602, 576, 5562, 836, 3842, 1304, 9724, 201, 4492, 717, 4153, 5336, 382, 294, 1382, 15058, 3384, 761, 146, 18100, 1496, 2876, 3…
$ expensive       <dbl> 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, …
```

Packages used

The following packages were used:

- Tidyverse – collection of R packages
- caret- build machine learning models
- ggplot2- primarily used for data visualization
- party- create decision trees
- ggpubr- produce production-quality visualizations
- kernlab- Used for kernel-based machine Learning
- arules- represent, manipulate, and analyze transaction data and patterns.
- arulesViz- handling and mining association rules.
- Maps- used to make map outlines and points
- ggmap- functions to visualize spatial data and models
- mapproj- convert latitude/longitude to coordinates
- rworldmap- maps global data
- rpart - to create regression decision tree
- Zoo - used for handling irregular or NA values in the dataset

Data Cleaning

```
RawhosData <-  read.csv(datafile)
view(RawhosData)
```

	X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
1	1	18	27.9000	0	yes	CONNECTICUT	Urban	Bachelor	No	Active	Married	0	female	1746
2	2	19	33.7700	1	no	RHODE ISLAND	Urban	Bachelor	No	Not-Active	Married	0	male	602
3	3	27	33.0000	3	no	MASSACHUSETTS	Urban	Master	No	Active	Married	0	male	576
4	4	34	22.7050	0	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	1	male	5562
5	5	32	28.8800	0	no	PENNSYLVANIA	Country	PhD	No	Not-Active	Married	0	male	836
6	7	47	33.4400	1	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0	female	3842
7	9	36	29.8300	2	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	1304
8	10	59	25.8400	0	no	PENNSYLVANIA	Country	Bachelor	No	Not-Active	Married	1	female	9724
9	11	24	26.2200	0	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	201
10	12	61	26.2900	0	yes	CONNECTICUT	Urban	No College Degree	No	Active	Married	0	female	4492
11	13	22	34.4000	0	no	MARYLAND	Urban	Bachelor	No	Not-Active	Married	0	male	717
12	14	57	39.8200	0	no	MARYLAND	Urban	Bachelor	Yes	Not-Active	Married	0	female	4153
13	15	26	42.1300	0	yes	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	5336
14	16	18	24.6000	1	no	PENNSYLVANIA	Country	No College Degree	Yes	Not-Active	Not_Married	0	male	382
15	18	23	23.8450	0	no	MASSACHUSETTS	Urban	No College Degree	No	Active	Married	0	male	294
16	19	57	40.3000	0	no	PENNSYLVANIA	Urban	Bachelor	Yes	Active	Not_Married	0	male	1382
17	20	31	35.3000	0	yes	PENNSYLVANIA	Urban	PhD	No	Not-Active	Married	0	male	15058
18	21	60	36.0050	0	no	PENNSYLVANIA	Urban	PhD	No	Active	Married	0	female	3384
19	22	30	32.4000	1	no	PENNSYLVANIA	Urban	Master	No	Active	Married	0	female	761
20	23	19	32.1600	0	no	PENNSYLVANIA	Urban	No College Degree	No	Active	Not_Married	0	male	146
21	24	32	31.9200	1	yes	NEW JERSEY	Urban	No College Degree	Yes	Not-Active	Not_Married	0	female	18100
22	25	37	28.0250	2	no	PENNSYLVANIA	Urban	PhD	No	Active	Married	0	male	1496
23	26	58	27.7200	3	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Not_Married	0	female	2876
24	27	62	23.0850	0	no	PENNSYLVANIA	Country	No College Degree	No	Active	Not_Married	1	female	3541
25	28	56	32.7750	2	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	0	female	3962
26	30	31	36.3000	2	yes	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0	male	11302
27	32	19	26.3150	0	no	PENNSYLVANIA	Urban	Master	No	Not-Active	Not_Married	0	female	484
28	33	20	28.6000	5	no	MARYLAND	Urban	Bachelor	No	Active	Married	0	female	466
29	34	63	28.3100	0	no	MASSACHUSETTS	Country	Master	No	Not-Active	Not_Married	0	male	3660

```
> dim(RawhosData)
[1] 7582   15
```

```
> str(RawhosData)
'data.frame':   7582 obs. of  15 variables:
 $ X              : int  1 2 3 4 5 7 9 10 11 12 ...
 $ age            : int  18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
 $ children       : int  0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : chr  "yes" "no" "no" "no" ...
 $ location       : chr  "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...
 $ location_type  : chr  "Urban" "Urban" "Urban" "Country" ...
 $ education_level: chr  "Bachelor" "Bachelor" "Master" "Master" ...
 $ yearly_physical: chr  "No" "No" "No" "No" ...
 $ exercise       : chr  "Active" "Not-Active" "Active" "Not-Active" ...
 $ married        : chr  "Married" "Married" "Married" "Married" ...
 $ hypertension   : int  0 0 0 1 0 0 0 1 0 0 ...
 $ gender         : chr  "female" "male" "male" "male" ...
 $ cost           : int  1746 602 576 5562 836 3842 1304 9724 201 4492 ...
 $ expensive      : num  1 1 1 0 1 1 1 0 1 1 ...
```

Checking for the NA and NULL values

```
> sapply(RawhosData,function(x) sum(is.null(x)))
            X           age           bmi      children        smoker      location  location_type education_level yearly_physical
            0             0             0             0             0             0             0             0             0
     exercise       married  hypertension        gender          cost
            0             0             0             0             0
```

```
> sapply(RawhosData,function(x) sum(is.na(x)))
            X           age           bmi      children        smoker      location  location_type education_level yearly_physical
            0             0            78             0             0             0             0             0             0
     exercise       married  hypertension        gender          cost
            0             0            80             0             0
```

We checked the data for missing or null value and found that there are few NA values in bmi and hypertension in which for the bmi we used na.approx function which replaces the NA values with the mean approximate values and for the hypertension since it is a numeric values with either 0 or 1 we assumed that it will no affect data so removed the 80 rows of NA values based on the hypertension.

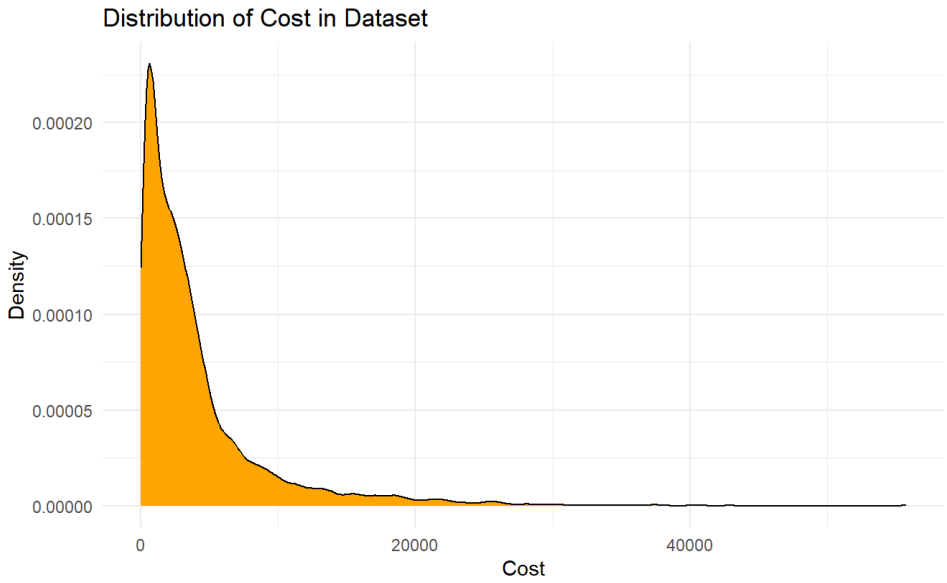
```
> RawhosData$bmi <- na.approx(RawhosData$bmi)
> table(is.na(RawhosData$bmi))

FALSE
 7582
```

```
#removing the na values from hypertension
RawhosData <- RawhosData[complete.cases(RawhosData), ]
```

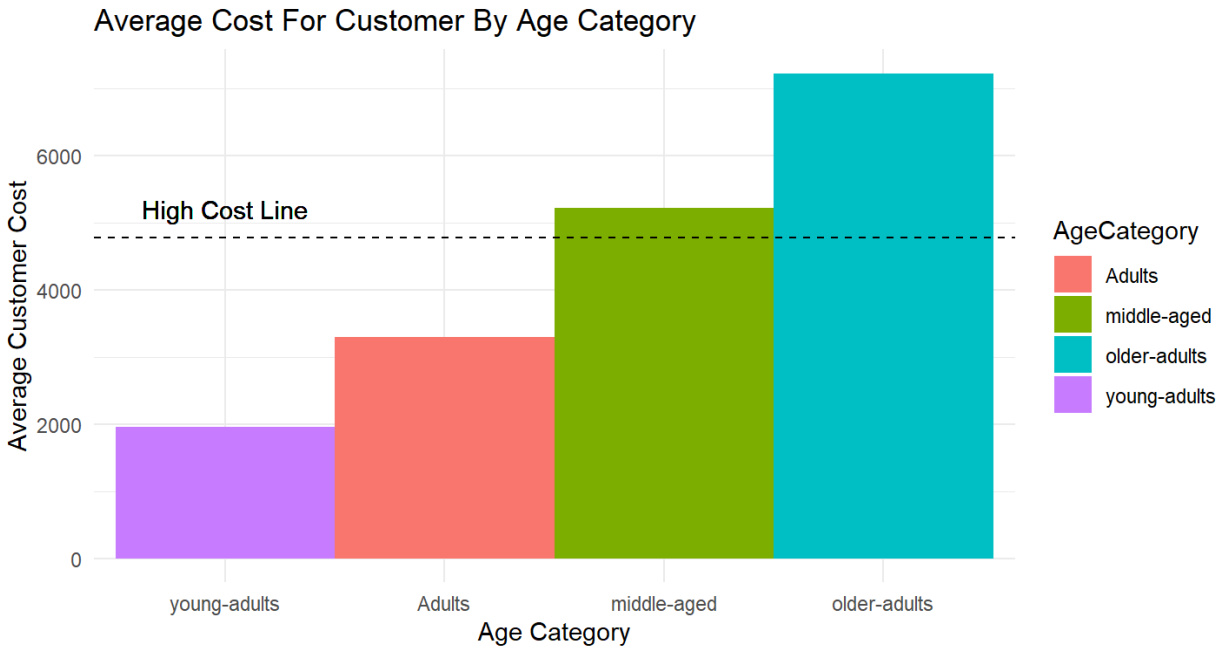
Variables Analysis

Number of health care cost Vs Number of expensive		
Row Labels	**Count of Expensive**	**Percentage**
0	1878	25.03%
1	5624	74.97%
Total	7502	

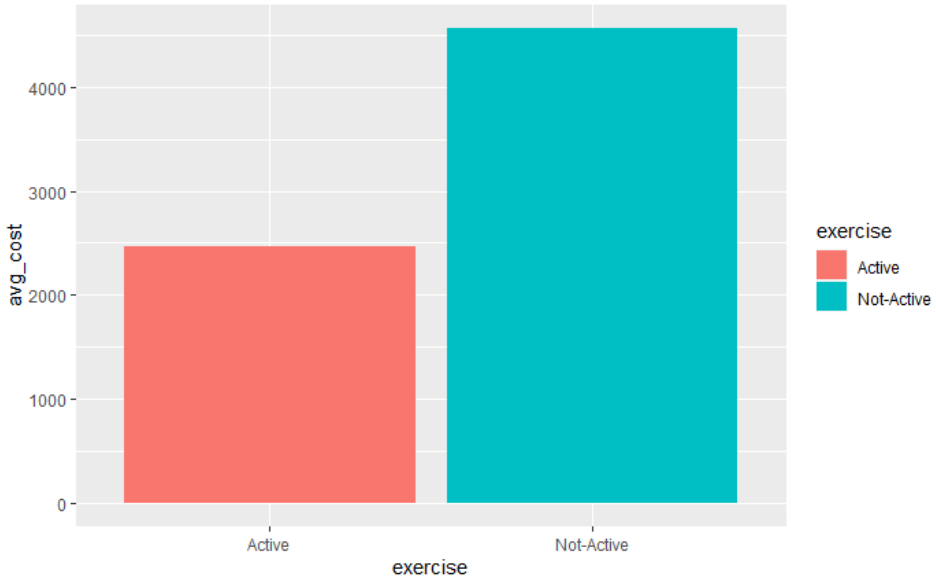
Analyzing our data variables



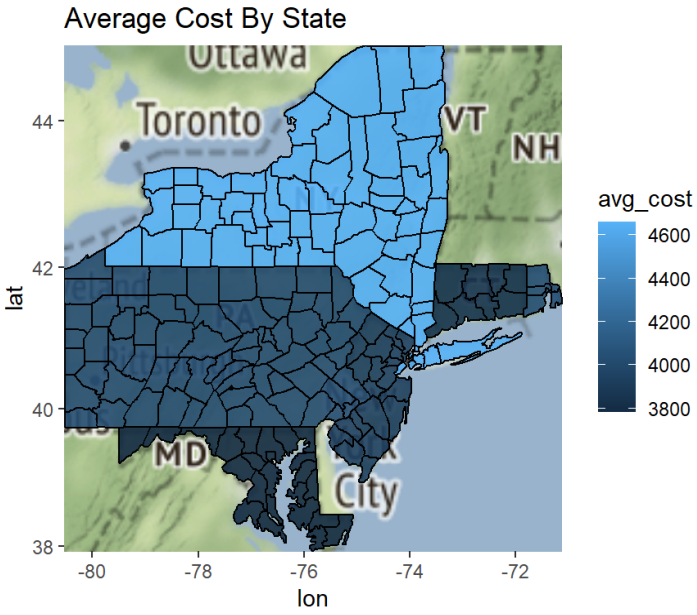
The above graph represents the distribution of cost in our dataset. As we can observe that it is left skewed indicating that the cost of healthcare for individuals generally falls between the range of $0-$20000 while some of it extending up to even $40000.

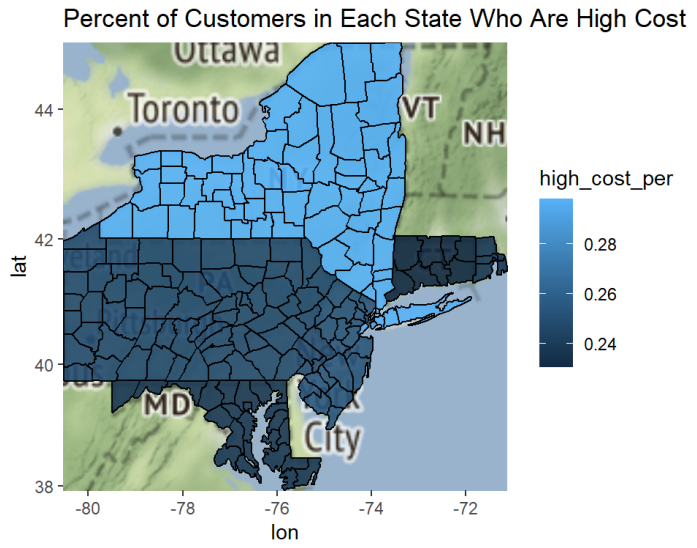


The bar plot above shows the average cost for each of four designated categories.. Initially the age of the individuals was divided into four categories as young-adults ranging from the age 18-25, adults between 25-40, middle-aged adults between 40-59 and older adults above the age of 59. As the groups age, they tend to pay more in average costs.

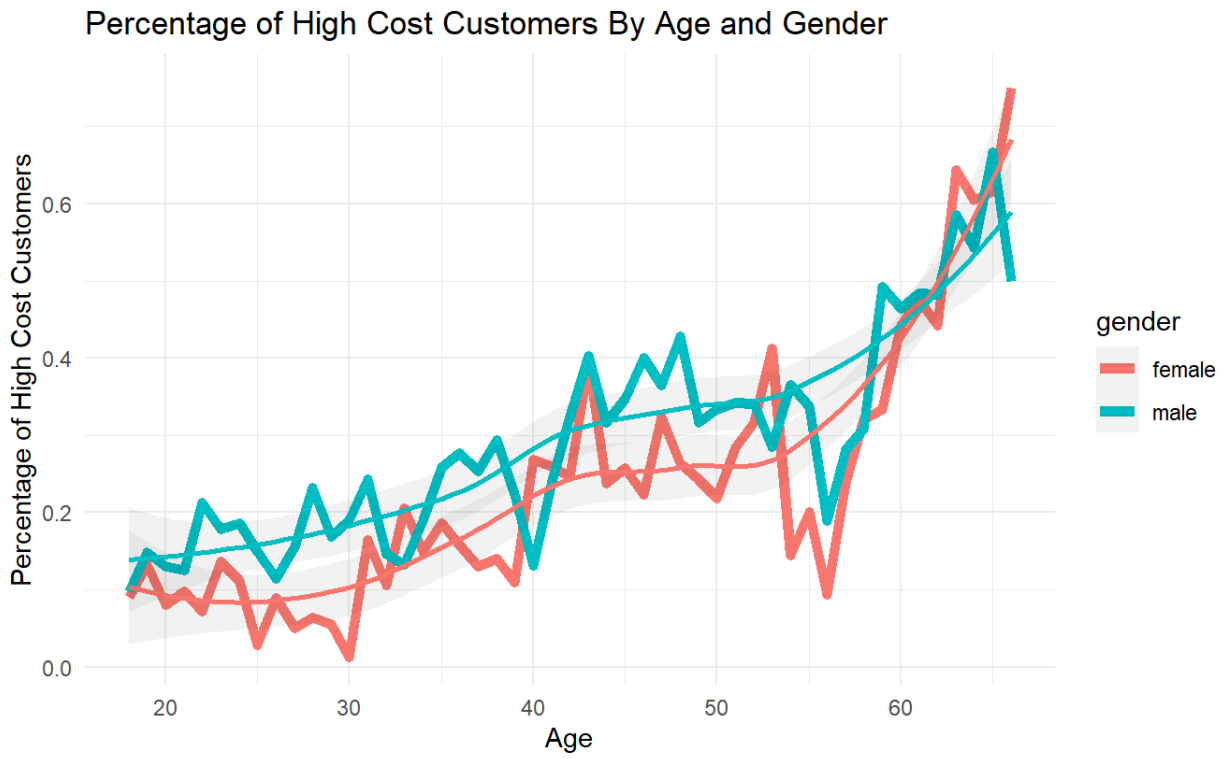


A bar plot was used to distribute the cost between two categories of individual based on their exercising activity. As it is observed that individuals who do not have an active exercising routine tend to spend more on health than the one's who exercise regularly as it is a great way to keep healthy.

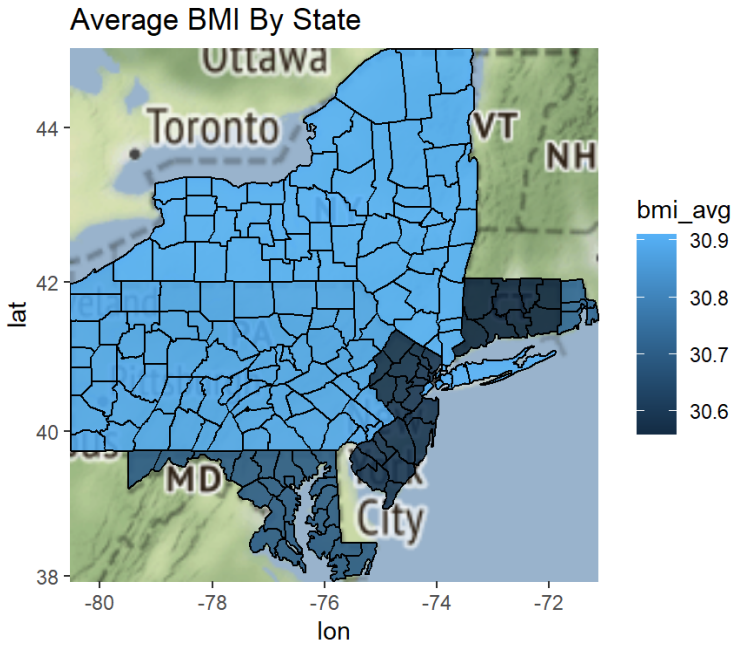




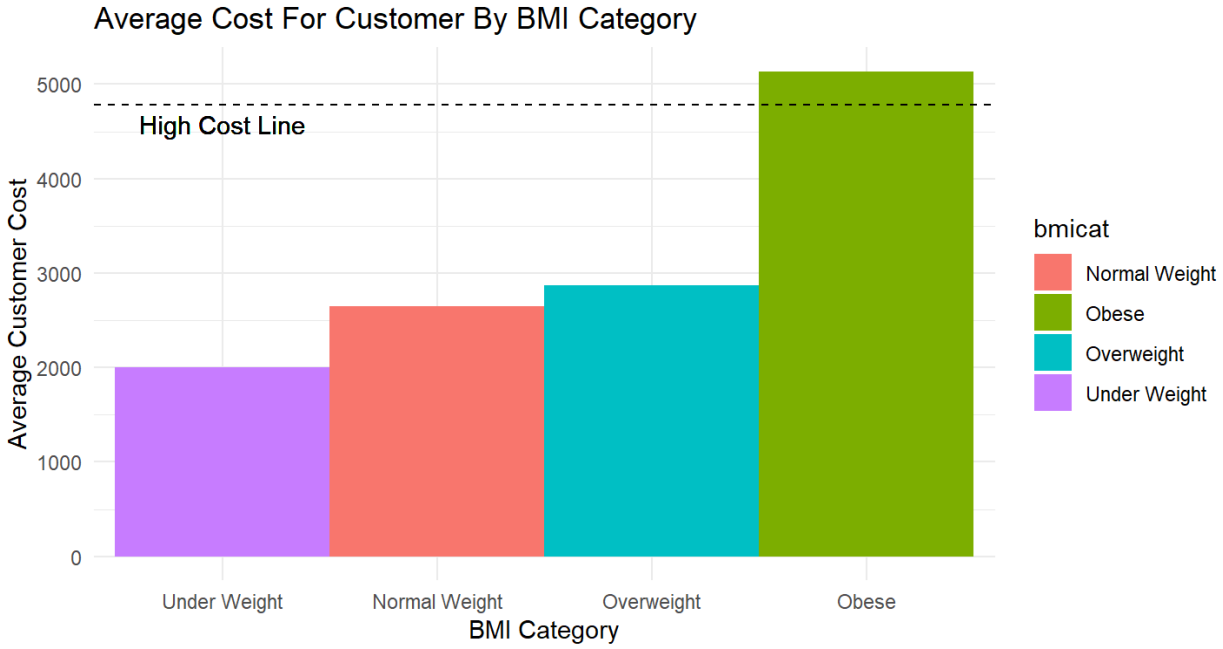
We used maps to visualize the cost and high cost distribution by state. New York had both the highest average cost per customer and the highest percentage of customers who were high cost. Maryland had the lowest average cost, while Connecticut had the lowest percentage of customers who were high cost, which seems to indicate that there were many outliers with a high cost in Connecticut that brought up the average in the state.



The line chart shows the percentage of customers who are high cost, broken down by age and gender. The trend indicates that customers are more likely to be high cost as they get older, with males having a slightly higher chance to be high cost until about age 55 when they become very similar.

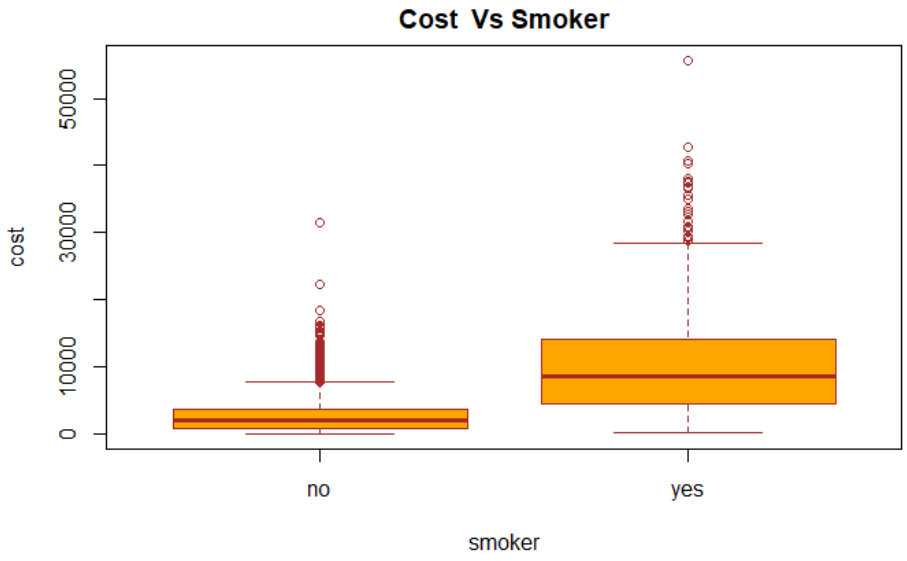


This map is a distribution of bmi over different states. New York had the highest average BMI, while Connecticut had the lowest.

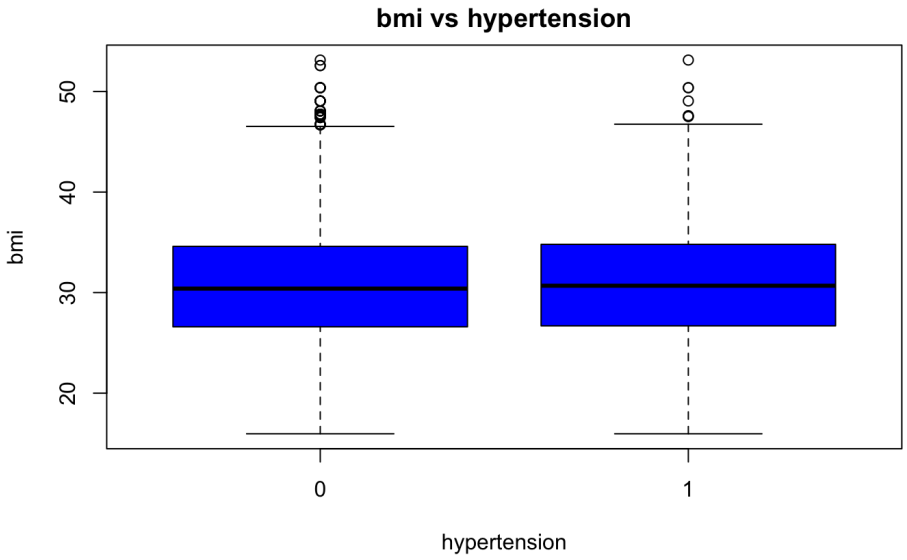


Another bar chart was used to show the average cost for each BMI weight class. Bmi was distributed in different categories, namely under weight falling under the index 18.5, normal weight falling between 18.5-25, overweight between 25-30 and obese to

be over 30. The obese group has the highest average cost by far and are the only group of the four to have an average cost that is above the high cost qualification. Under weight has the lowest average cost, followed by normal weight, but the difference between those 3 groups is much smaller than th gap between them and the obese group.



A box plot was used to illustrate the cost distribution between smokers and non smokers. As it is observed the median cost for an active smoker is much higher than that of a non smoking individual. However the outliers indicate that there might also be other health concerns taken into consideration apart from smoking activity.



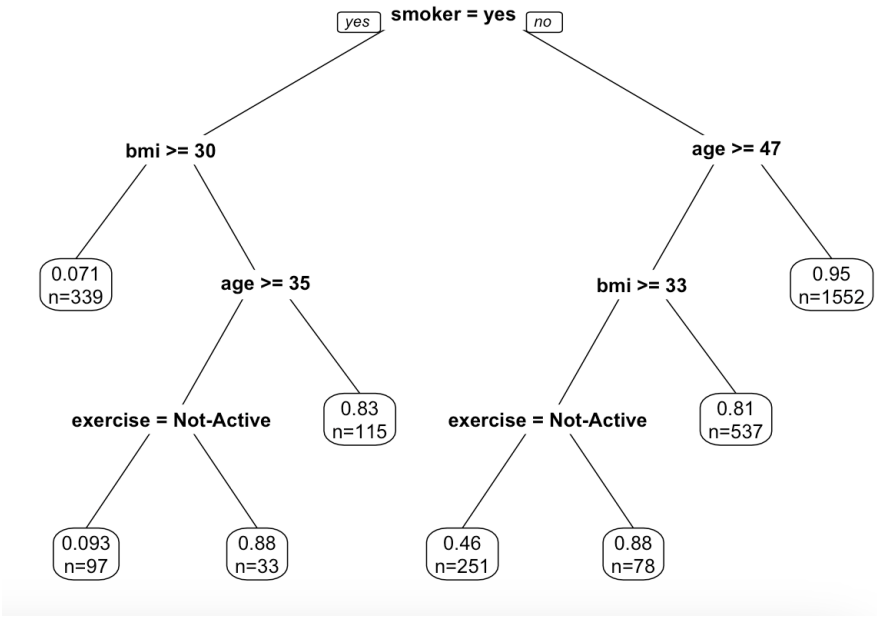
The box plot below shows the BMI distribution between hypertensive and nonhypertensive individuals.. However there is no significant difference in the plots.

Model Prediction

After analyzing the data and making sure that we do not have any NA values. We produced a set of 4 features which we thought would be the best for running the model. First, we ran varIMP function on our dataset to see which features have the highest importance leading to accurate prediction of expensive health care as most of the data have categorical variables and this function helped us to classify based on the results of the variables. The higher the value, the more its significance is of that feature. We realized Somker, exercise, age, bmi is a few of the most significant variables and used those features in our Random Forest, SVM Model and glm model.
Note: The train() command that computed the Regression Trees and SVMs were not included in these results because their ksvm() and rpart() counterparts had better performances (like accuracy).

Cart model decision tree

The working of decision tree models is based on repeated partitioning the data into multiple sub-spaces, so that the outcome in each final sub-space is as homogeneous as possible. This approach is technically called recursive partitioning.



Linear Model

We used linear model for all the variables in our dataset

```
Call:
lm(formula = expensive ~ smoker + exr + age + bmi + chld + hyper,
    data = tdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13807 -0.12573  0.05785  0.20385  0.92957

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.6884403  0.0229870  73.452  < 2e-16 ***
smokeryes     -0.5978698  0.0095592 -62.544  < 2e-16 ***
exrNot-Active -0.1686117  0.0087464 -19.278  < 2e-16 ***
age           -0.0073365  0.0002688 -27.295  < 2e-16 ***
bmi           -0.0126862  0.0006361 -19.944  < 2e-16 ***
chld          -0.0113903  0.0031152  -3.656 0.000258 ***
hyper         -0.0335965  0.0094557  -3.553 0.000383 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3278 on 7495 degrees of freedom
Multiple R-squared:  0.4276,    Adjusted R-squared:  0.4271
F-statistic: 933.2 on 6 and 7495 DF,  p-value: < 2.2e-16
```

The adjusted R-squared from this model is 0.4271. The adjusted R-squared of 0.4271 means that the "Smoker", "exercise", "age", "age", "bmi", “Children” and "hypertension" variables explain 74.97% of the "expensive" (whether or not the person has expensive health care) variable's variation. The p-value of the model is 2.2e-16, so it is highly likely that the changes caused by these variables do not occur by chance.
The "age", "bmi", "children", and "hypertension" are statistically significant without any conditions because they are metric variables. The "smoker" variable is statistically significant when its value is equal to "yes". Similarly, the "exercise" variable is statistically significant when its value is equal to “Not Active”.

Random Forest Regression

Random Forest is a collection of decision trees. To obtain more precise predictions, it constructs and merges several decision trees. Based on the significant variables i.e age, bmi, exercise, children, hypertension and smoker we used Random forest algorithm to train the model.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  691   81
         1  434 3294

               Accuracy : 0.8856
                 95% CI : (0.8759, 0.8947)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6592

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

            Sensitivity : 0.6142
            Specificity : 0.9760
         Pos Pred Value : 0.8951
         Neg Pred Value : 0.8836
             Prevalence : 0.2500
         Detection Rate : 0.1536
   Detection Prevalence : 0.1716
      Balanced Accuracy : 0.7951

       'Positive' Class : 0
```

From Random Forest model , the accuracy is 88.56 % along with the INR value 0.75 and p - value is < 2.2e-16. Based on these values we got a pretty good model.

Generalized linear model

The glm() function in R can be used to implement, a family of regression models that accommodates non-normal distributions and allows us to apply a variety of regression technique. Since our dependant variable is a categorical variable we used binomial family in the glm function.

```
Call:
glm(formula = expensive ~ ., family = binomial(), data = trainSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2788  -0.0744   0.2511   0.5258   2.6309

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   11.314860   0.511408  22.125  < 2e-16 ***
age           -0.075892   0.004916 -15.438  < 2e-16 ***
smokeryes     -4.035003   0.165004 -24.454  < 2e-16 ***
hyper         -0.311459   0.137353  -2.268   0.0234 *
exrNot-Active -1.767758   0.165930 -10.654  < 2e-16 ***
chld          -0.208057   0.045199  -4.603 4.16e-06 ***
bmi           -0.133946   0.010547 -12.700  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3377.4  on 3001  degrees of freedom
Residual deviance: 1946.5  on 2995  degrees of freedom
AIC: 1960.5

Number of Fisher Scoring iterations: 6
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  747  202
         1  378 3173

               Accuracy : 0.8711
                 95% CI : (0.861, 0.8808)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6374

 Mcnemar's Test P-Value : 3.69e-13

            Sensitivity : 0.6640
            Specificity : 0.9401
         Pos Pred Value : 0.7871
         Neg Pred Value : 0.8936
             Prevalence : 0.2500
         Detection Rate : 0.1660
   Detection Prevalence : 0.2109
      Balanced Accuracy : 0.8021

       'Positive' Class : 0
```

As we can see, the accuracy is 87.11 % along with the INR value 0.75 and p - value is < 2.2e-16. Based on these values we got a pretty good model.

Svm model

Support vector machines, or SVMs, are supervised machine learning algorithms that are mostly used to categorize data into groups.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  680  100
         1  445 3275

               Accuracy : 0.8789
                 95% CI : (0.869, 0.8883)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6403

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

            Sensitivity : 0.6044
            Specificity : 0.9704
         Pos Pred Value : 0.8718
         Neg Pred Value : 0.8804
             Prevalence : 0.2500
         Detection Rate : 0.1511
   Detection Prevalence : 0.1733
      Balanced Accuracy : 0.7874

       'Positive' Class : 0
```

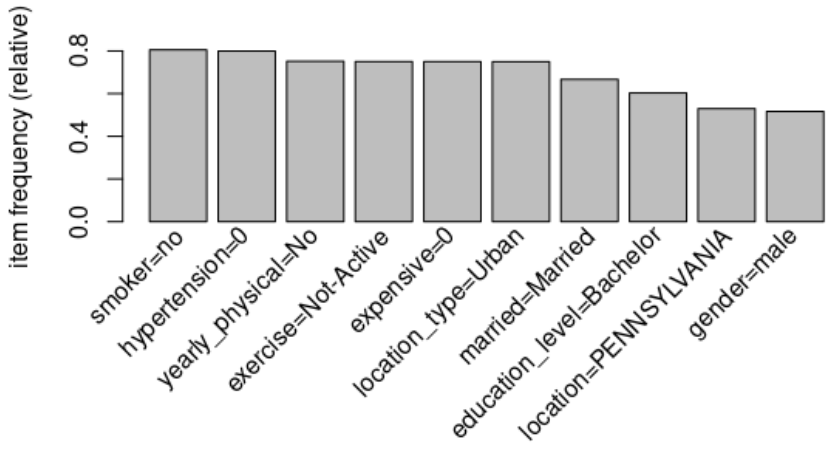
By using SVM model, the accuracy is 87.89 % along with the INR value 0.75 and p - value is < 2.2e-16. Based on these values we got a pretty good model.

Although, all three models gave a pretty good accuracy we went with Random Forest modelling for our data set because of its ability to prevent overfitting of the model is also among its many advantages.The random forest can handle a lot of features and aids in determining the crucial characteristics.

Association Rules

Machine learning associations between variables in large databases are discovered using association rule learning.Associative rules are designed to discover the rules that determine how or why certain items are related in any situation where there are many items. They are developed on the basis of some measures of interestingness which are used to identify strong rules discovered in databases.Using Association rule learning we found some interesting relations between our variables.

- There is a 91.5 % chance for an individual who is a married male who does not actively exercise , has no hypertension and does get a yearly physical check up to spend more on healthcare.
- There is a 51.2 % chance for an individual having 3 children and between the age group 47 - 66 to spend more on healthcare.
- There is a 79% chance that an individual living in the state of New York and 76.4 % chance that an individual living in the state of New Jersey is likely to spend more than the other state residents.
- There is an 80% chance for an individual having 2 children and being an active smoker to spend more on healthcare.



Shiny App

In addition to the analysis detailed above, we also built a Shiny web app within R. The app has 4 key features. First, it allows the user to upload two files. The first file should contain the data about the customers, without the cost the paid, while the second contains the cost of each customer in that first data set. The app will then display the first 5 rows of data in the customer file and allows for the user to adjust how many rows are shown. Two graphs are shown with some exploratory information from the training set we were given. The first looks at the percentage of smokers who were deemed high cost compared to the percentage of non-somers who received that designation. Underneath that is a density plot showing the distribution of cost throughout the entire dataset. Two check boxes are provided to the user allowing them to toggle between including each of the smoker and non-smoker categories. If one of those boxes are unchecked, the distribution changes itself to reflect that on the graph.

Once the data files are uploaded, the analysis from the preloaded model will be conducted and the results will be shown in a text box and table at the bottom of the page. In the textbox, the sensitivity and accuracy are taken from the confusion matrix, which is calculated from comparing the model's predictions to the solutions provided. That sensitivity is then compared to the one we got with our training data, which was approximately 0.88. Underneath that, the full confusion matrix is shown in a table. The first column shows whether the customer was predicted to be expensive (1) or not (0), the second column shows whether they actually were expensive (1) or not (0), and the final column shows how many times that combination was observed in the dataset.

When the app was completed, we published it to Shinyapps.io and it can be viewed here: https://matthewpenn.shinyapps.io/Group1_Project/.

Conclusion

We initially started with all 15 variables . After running through CART, we realized the 6 most important variables and went through with them in our completed model.

In our introduction, we stated the goal of our analysis is to answer two key questions

- Do older people, who are more prone to hospitalization, have a large healthcare cost?
- Does a combination of high BMI & hypertension contribute to expensive healthcare?
- Do regular exercise contribute to a healthier lifsestyle which in turn reduce the health cost?

Based on our analysis these are the significant and potential casual relationships that can contribute to expensive health care.

a) Age

Based on our analysis Age is a major role in the health care expensive. Older Adults people i.e people who are above 59 years of age are more likely to spend on health care than the other gae group.

b) Exercise

One factor we found is that being in-active in regular exercise can cause health issues which can impact on the health care cause. Whereas, people who regurarly exercise lead a healthier lifestyle.

c) Smoker

Through our analysis, one of the major factor that contributed to the more health care expense is the habit of smoking. People who are active smokers are likely to spend more money on health care than the other people irrespective of the age.

d) Hypertension

From the analysis we inferred that hypertension and stress can also cause serious health problems and lead to a expensive health care.

e) bmi

Bmi is body mass index which is a value derived from height and weight of the person. People who do not maintain the ideal bmi value ie. value between 18.5 to 25

are more likely to spend on health care. Among those who are overweight (26-30) and obese (above 30) are the one affected the most with the health cost.

f) Children

We see that family with two or more children tend to spend more on health care, from which we can infer that pediatrics contributes a larger portion in the health care expenditure.

Recommendations

- Help young adults cultivate an active lifestyle
    - This helps them to stay active when they get older to reduce the number of complicated health issues they may face
- Invest in campaigns that fight against smoking
    - Smoking leads to a lot of breathing problems in addition to other organ failure related issues. Providing resources to educate people about smoking and also making rehabilitation easily accessible.

Limitations

➔ Lack of customer behaviour data

➔ From our analysis, we discovered that variables relating to the customer's behavior, such as past medical history, show significance in leading to expensive health care. Unfortunately, our dataset is limited in providing variables that relate to customer medical records. However, we understand such data is not readily available and might be difficult to monitor.

Future Studies

★ Include more customer data

One recommendation that can improve our study is to source customer behavior data. Examples include data which provides information about customers' medical records, type of health care benefits taken and if any health issues were resolved or

not. With this additional data, we might understand the customer's health journey in greater details and identify alternate good predictors for health care costs.

- ★ Cross reference results and findings from different hospitals and locations

To improve our model and increase its relevance to different hospitals, one recommendation is to perform a similar analysis of data from various hospitals from other parts of US state and cross-reference the results. Doing this allows us to verify our insights.