IST 687 – Introduction to Data Science

Final Project

Health Management Organization

Prachi Sadarangani
Spruha Band
Yashaswini Kulkarni
Akshitha Reddy Kalvakuntla

Table of Contents

Overview of the Project:

IST 687: Introduction to Data Science.

As a group, we are given a data set and the overall goal of the case is to provide actionable insights, based on the data available, as well as accurately predict which people are expensive. The data set that we got contains healthcare cost information from a Health Management Organization from which each row in the dataset represents a person. Our goal is to understand the key drivers of why some people are expensive and also predict which people will be expensive in the future.

Therefore, on the most precise level, we have two major goals to meet through the course of this project:

1. Predict people who will spend a lot of money on health care the coming year
2. Provide actionable insights to the HMO, in terms of how to lower their health care costs, by providing a specific recommendation on how to lower the health care costs.

To begin with, we as a team started by exploring and cleaning the data set. This helped us in better understanding of the structure of the dataset. Then we tried to summarize variables within dataset to figure out what attributes are contributing to the health costs. Next, we tried to come up with visualizations of various factors. Finally, we created a Shiny App, to create an opportunity to interact with the insights.

Business Questions:

The business questions that we are trying to answer through this project are:

- Which aspects of a person affect the healthcare costs the most?
- Which aspects of a person moderately affects the healthcare cost?
- Which aspects of a person have the least effect on their healthcare costs?
- Which age groups spends the most on health care?

Assumptions:

The first assumption that we have made was to determine the basis on which we consider the cost to be "expensive" and "inexpensive". For this, we have decided on choosing the costs in the top 22% of the dataset to be considered as "expensive". This means that any individual with a cost greater than $5,353.00 will be contributing to the expensive section of the dataset and the remaining 78% will be contributing to the inexpensive section. The reason for choosing the top 22% is because, when we started with the project, we first decided on choosing the median as the threshold, but with that condition almost 75% of the dataset is falling in the expensive section and that is the reason why we went with choosing the 78^{th} percentile.

The second assumption that we made was to consider each observation – a unique observation, irrespective of the fact that there can be some identical attributes. This assumption is based on the fact that the variables that we are dealing with in the dataset are general in nature and there is a high potential for identical traits or characteristics with unique individuals in the dataset.

The final assumption is to stick to the consideration of gender and age variables contribute to biases in algorithms but because we are dealing with the data set that is related to the healthcare industry, gender and age play a pivotal role in determining the expensive and inexpensive analysis. Therefore, these two variables are taken into consideration throughout the project.

Data Exploration:

The dataset had 14 different variables and it is important to understand the behavior of the each of the variable to predict or analyze the dependencies for next year.

The first variable we tried to explore was the education level, in which we figure out that people with a bachelor's degree are of higher frequency contributing to nearly half of the dataset followed by people with a master's degree.

Var1 <fctr>	Freq <int>
Bachelor	4578
Master	1533
No College Degree	759
PhD	712

The second variable was based on the location type, in which we see that 5679 which accounts to nearly 74.9% of people live in the city whereas the remaining live on the country side.

Var1 <fctr>	Freq <int>
Country	1903
Urban	5679

Based on being a smoker in the past year, nearly 80.4% of the people in the data set were non-smokers from the past year, whereas as the rest were smokers from the past year.

Var1 <fctr>	Freq <int>
no	6103
yes	1479

Based on the gender, both male and female contribute around equally to the dataset.

Var1 <fctr>	Freq <int>
female	3662
male	3920

Next, we see around 66.7% are married whereas the remaining group of people are unmarried.

Var1 <fctr>	Freq <int>
Married	5060
Not_Married	2522

The frequency of people having no children is dominant compared to people having 1,2,3,4,5 children off which people with 5 children is the least.

Var1 <fctr>	Freq <int>
0	3259
1	1772
2	1367
3	942
4	130
5	112

Looking at the number of people from different locations, we see that a majority of people were residing in Pennsylvania and the number of people from Massachusetts is the least of the lot. However, if we observe the distribution contributed from other places is more or less the same.

Var1 <fctr>	Freq <int>
CONNECTICUT	611
MARYLAND	747
MASSACHUSETTS	465
NEW JERSEY	498
NEW YORK	547
PENNSYLVANIA	4010
RHODE ISLAND	704

Coming to the data that contributes to people having their yearly physical with any professional, maximum number of them haven't booked when compared to the people that have booked their yearly physical with a professional.

Var1 <fctr>	Freq <int>
No	5699
Yes	1883

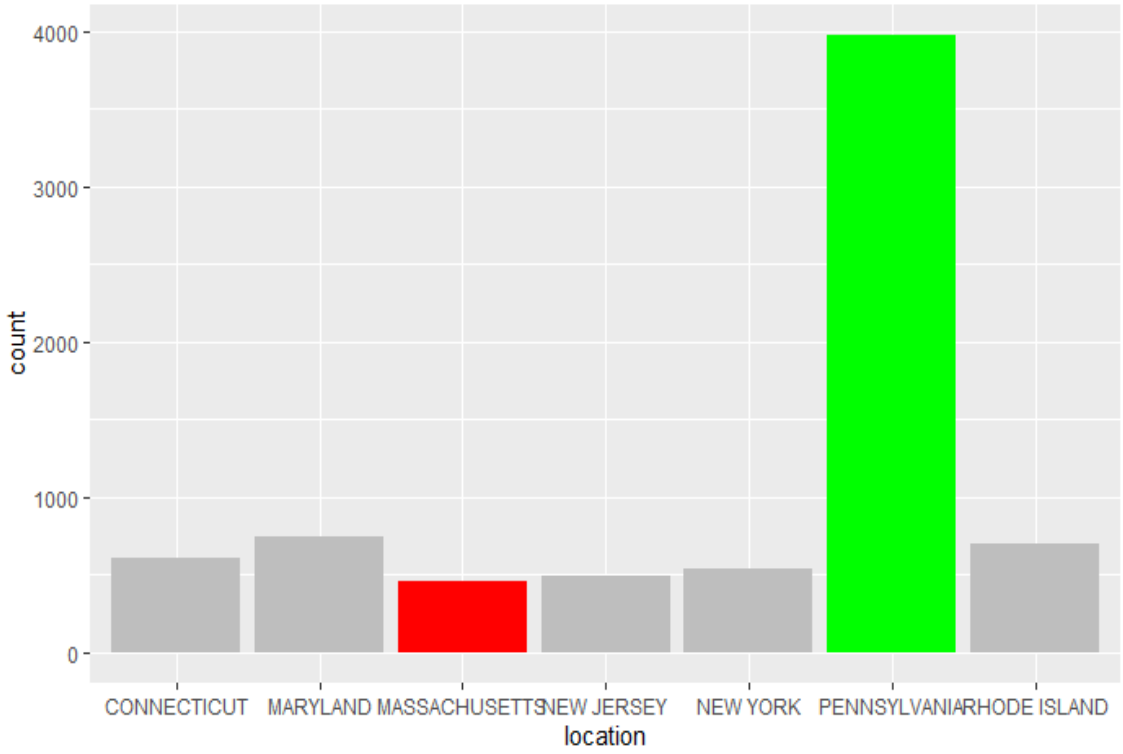
People with hypertension that constituted to the dataset is around 1,504 whereas those without hypertension were around 5,998

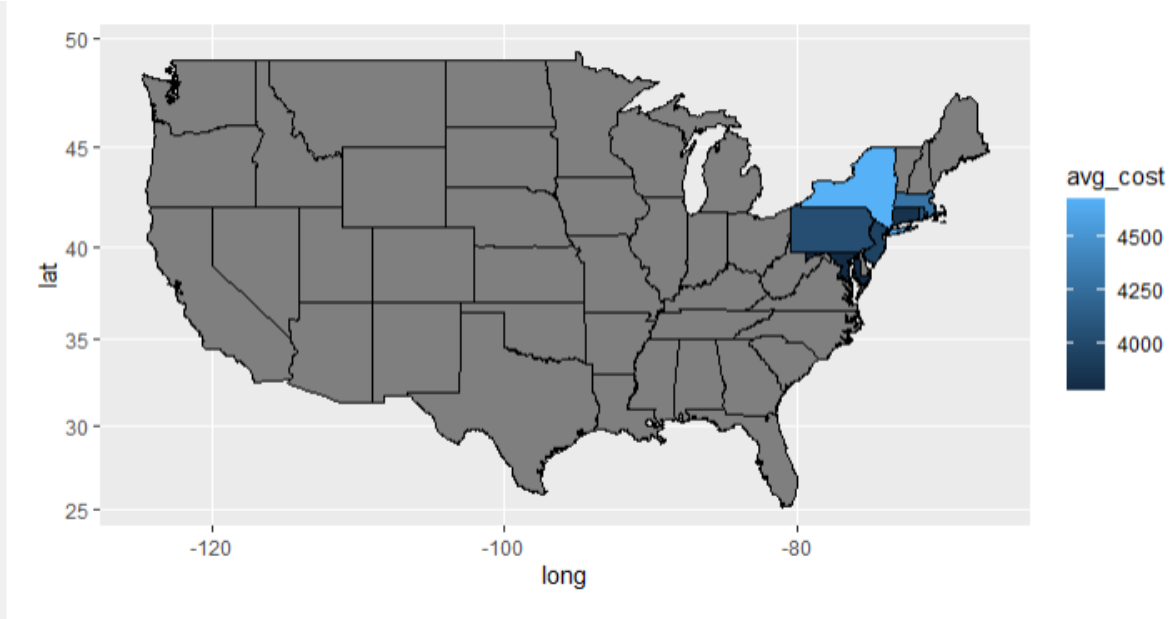
Var1 <fctr>	Freq <int>
0	5998
1	1504

Nearly, 75% of the individuals contributing to the dataset are people that have been active members when it comes to exercising from the past year, whereas the remaining 25% are people are not active.

Var1 <fctr>	Freq <int>
Active	1888
Not-Active	5694

Analyzing the data over the location and finding out which states contribute the most in the dataset and which state is the most expensive, we found that Pennsylvania contributes the highest data in the dataset and Massachusetts contribute the least. Coming to the expensive relation, we find that New York is significantly more expensive than other states. The visualizations are as below:


4000
3000
2000
1000
0
count
CONNECTICUT
MARYLAND
MASSACHUSETTS
NEW JERSEY
NEW YORK
PENNSYLVANIA
RHODE ISLAND
location


50
45
40
35
30
25
lat
-120
-100
-80
long
avg_cost
4500
4250
4000

Data Cleaning:

After exploring the data, the next stage is to clean the data. To begin with this, we started using functions like head() , tail(), str() and summary() to have a complete clear picture on the data set. Then, cleaning of the data is done using the following cleaning methods:

- First, we dealt with the null values using the na_interpolation from the imputeTS library for the BMI column.
- We cannot use mean or na_interpolation to fill in the missing values because the column for hypertension is of a binary data type. Therefore, we will delete any rows when the hypertension property has missing values.
- Next, we factorized the columns using as. factor method to categorize them because it will be easy to analyze later.

Business Question 1:

Which aspects of a person affect the healthcare costs the most?

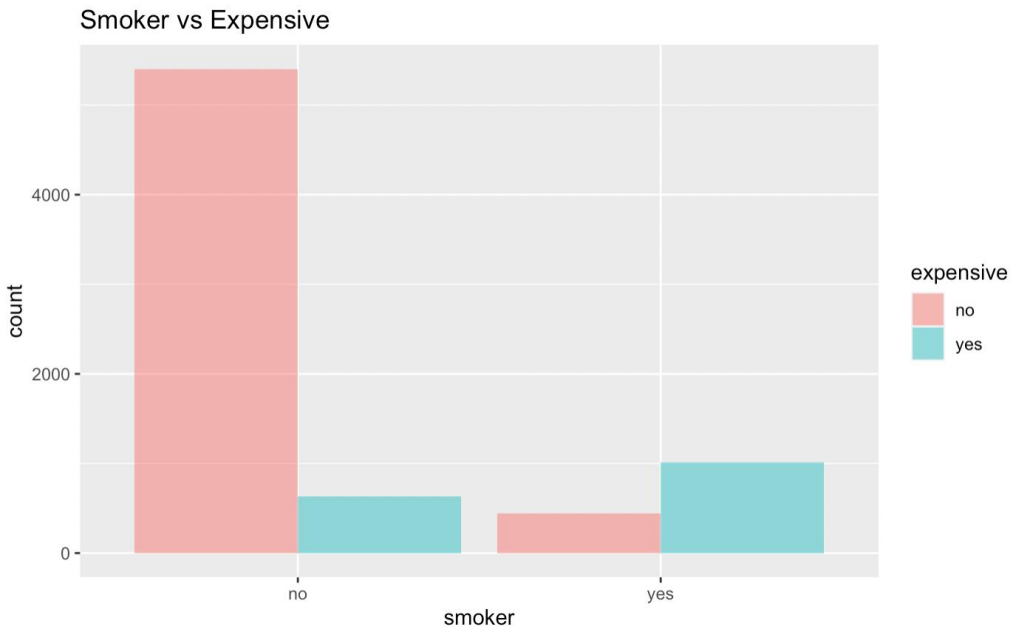
From the prediction models we have achieved, we found that that the factors that affect the healthcare cost most are:

1)Smoker

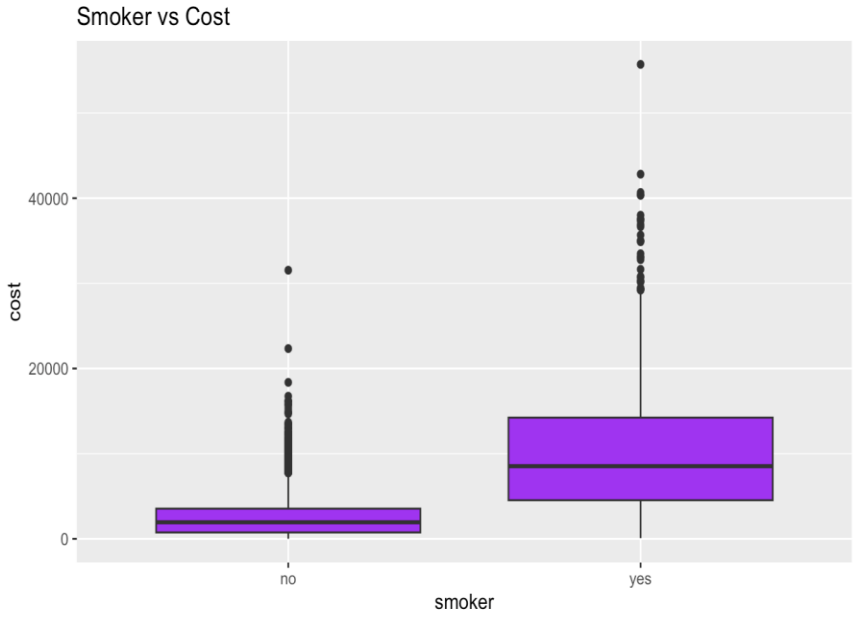
2)BMI

3)Exercise

Smoker:

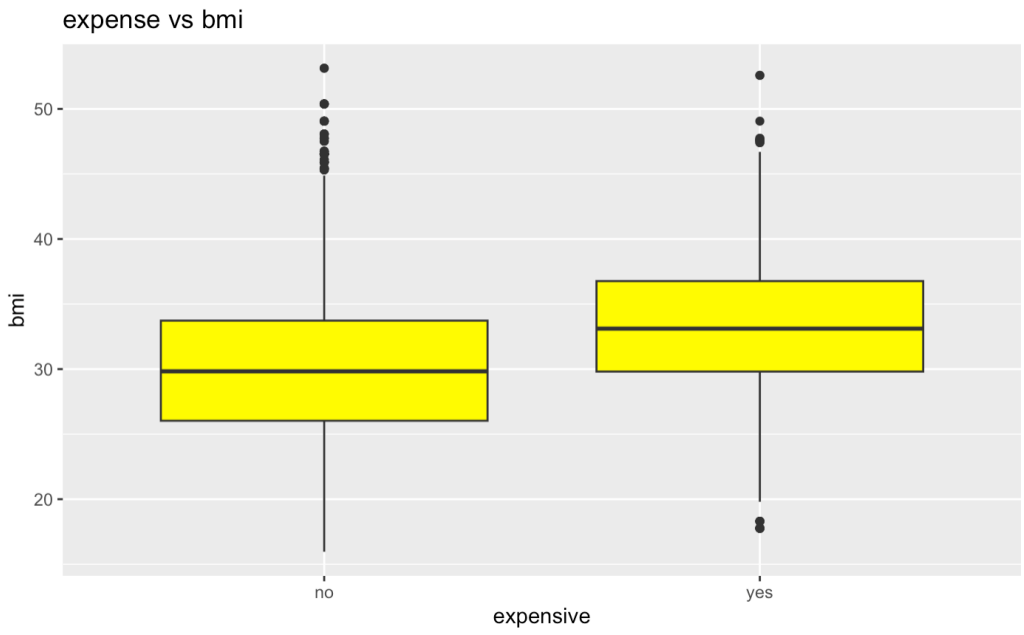


From the graph above we see that on the right where the smoker = “yes” category is represented, the expensive= “yes” dominates the expensive = “no”, which in simpler words mean that the healthcare costs are relatively higher for those who are into the habit of smoking.



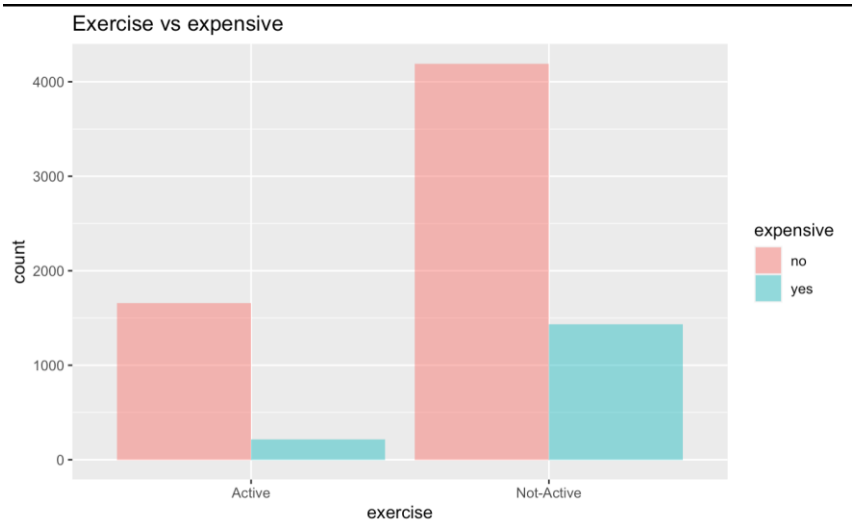
As we can see from the box plot, the median cost for smokers is way higher than that for non-smokers. This is same for the 25th and 75th quartile value. There are a few data points which are outliers and are hence exceptions.

Body Mass Index:

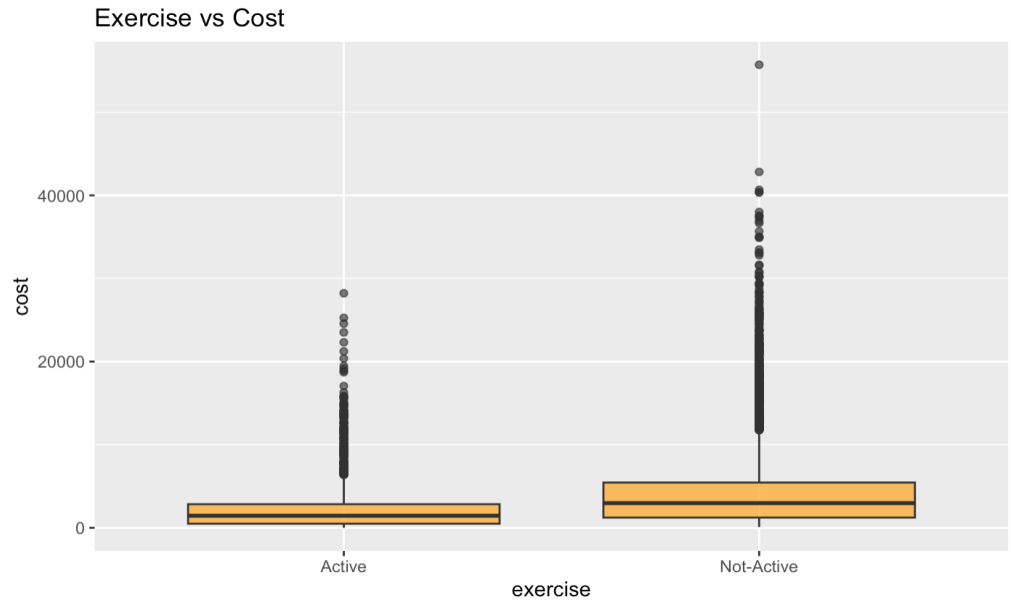


According to the previous box plot, almost 50% of expensive people have higher BMI than 50% of the non-expensive people.

Exercise:



From the above visualization people that are not active members of exercising are relatively more expensive than those who are active members of exercising.



Here, the median value of the non-active set of people have higher cost than that of the active people. This shows that almost 50% of non-active people pay more than 50% of active people.

Business Question 2:

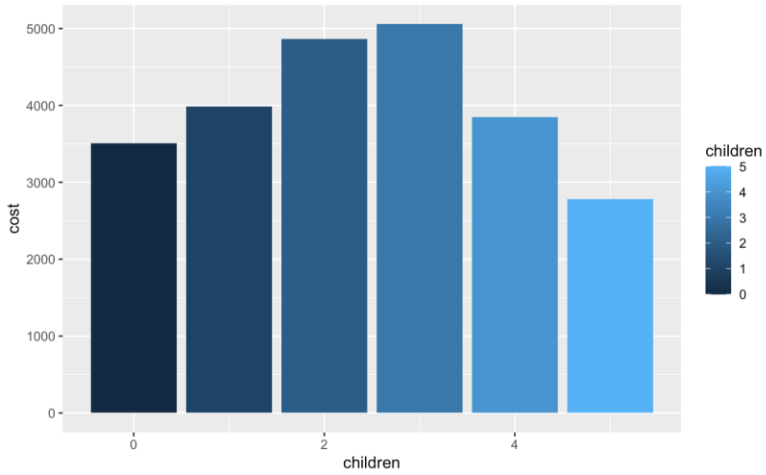
Which factors have the moderate effect on healthcare?

Factors effecting the healthcare cost moderately are:

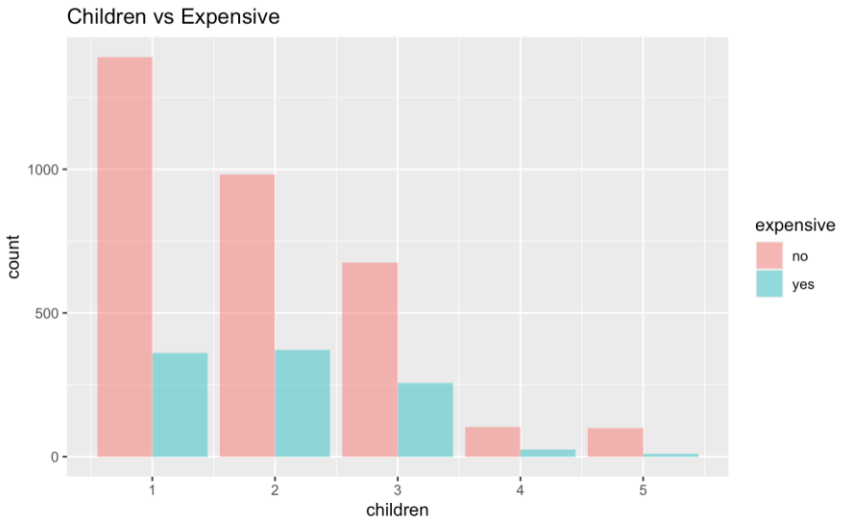
1.) Number of children a person has

2.) Hypertension

Visualizations based on the number of children a person has are as below:

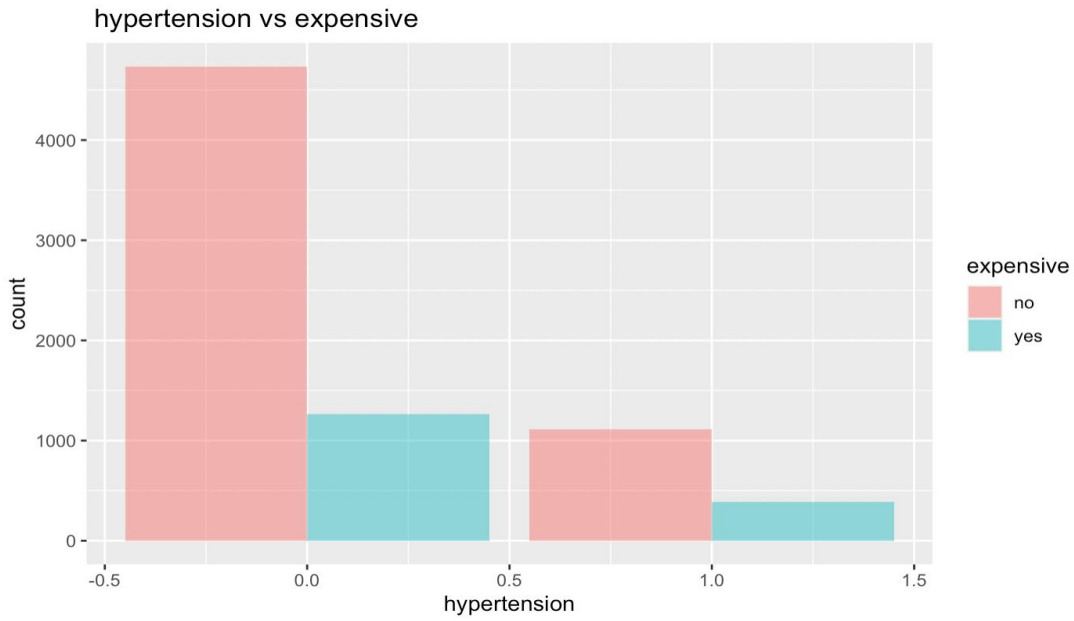


We see that the cost is more when the number of children is 3. The expensiveness of people with different number of children is showed in the graph below.



Visualizations based on hypertension are as below:

In the visualization we see below, we understand that the number of people with lower hypertension are more in the dataset and they are comparatively less expensive that the people with higher hypertension.



Business Question 3:

Which factors have the least effect on healthcare?

From our analysis, we found that the following five variables do not show as much effect as the factors discussed above on the health care costs.

- Location
- Gender
- Yearly physical
- Marriage
- Education level

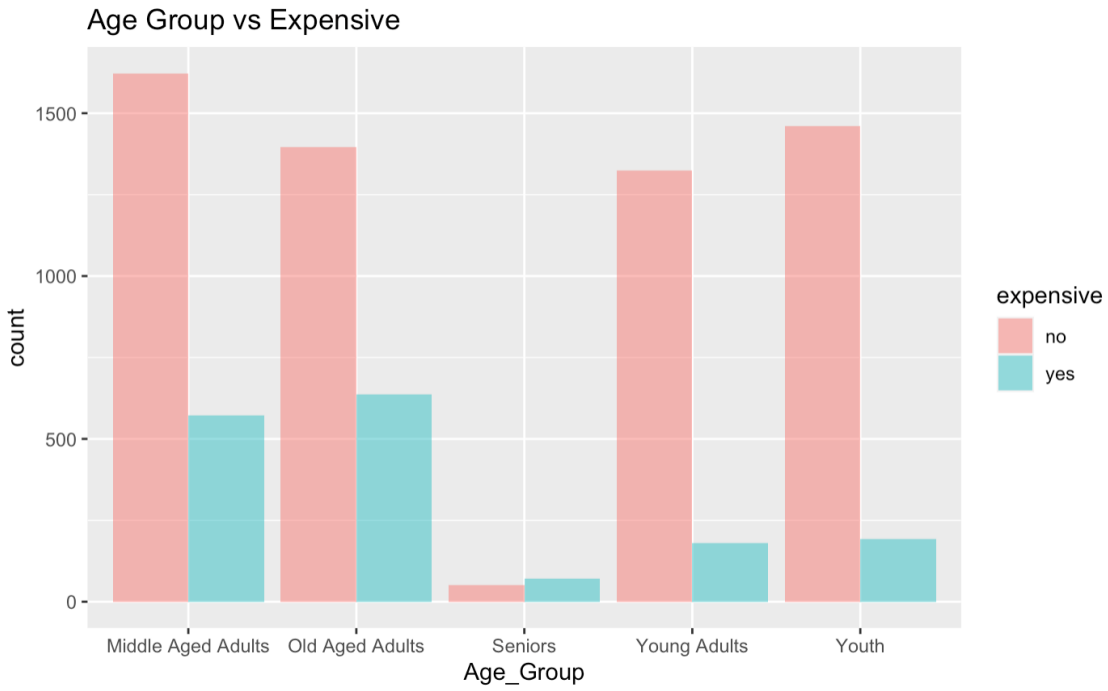
Business Question 4:

Which age group spends the most on the health care cost?

To analyze and answer this question, we have first classified the age groups separately as below:

- Age group >=15 and <25 – Youth
- Age group >=25 and <35 – Young Adults
- Age group >=35 and <50 – Middle-aged adults
- Age group >=50 and <64 – Old-aged adults
- Age>=64 – Seniors

After that, we then visualized the expenses against the age groups and saw that of all the age groups, seniors is the only age group in which the expenses “yes” is dominating the expenses “no” part. Therefore, we conclude that the seniors age group is the most expensive age group on the health care cost.



Models Used for Predictions:

The models that we used for the predictions throughout the project are:

1.) Linear regression model was first used to identify the significant predictors for the cost/isexpensive column

2.) These predictors were then used to build the SVM – Support Vector Machine Model

3.) SVM model is a supervised machine learning model which took the following significant factors: age, BMI, smoking, hypertension, and children to predict the target variable i.e., expensive

4.) The confusion matrix ran on the SVM model gave an accuracy of 88% and sensitivity of 97%

LM Model:

The Linear Regression model takes cost as the dependent variable and all other variables as the independent variables. A combination of these variables is used to understand the impact on the dependent variable. The LM model is used to identify the significant predictors. First model was built in order to identify the significant predictors. After finding the significant predictors, the second model is built using these significant predictors as the independent variables. These variables are used to further build the support vector machine model.

```
lm_out <- lm(formula = cost ~ age+bmi+smoker+exercise+children+hypertension,
             data=datafile)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-12216  -1492   -356   1024  41763

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -9078.508    227.405 -39.922  < 2e-16 ***
age                  102.151      2.664  38.342  < 2e-16 ***
bmi                  182.410      6.284  29.028  < 2e-16 ***
smokeryes           7695.696     94.677  81.284  < 2e-16 ***
exerciseNot-Active  2273.202     86.734  26.209  < 2e-16 ***
children             236.355     30.906   7.648 2.31e-14 ***
hypertension         326.699     93.667   3.488  0.00049 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3232 on 7417 degrees of freedom
  (158 observations deleted due to missingness)
Multiple R-squared:  0.5736,    Adjusted R-squared:  0.5732
F-statistic:  1663 on 6 and 7417 DF,  p-value: < 2.2e-16
```

SVM Model:

The Support Vector Machine model is used for working with categorical variables. The SVM model is used to predict the target variable which is the factorized 'expensive' column. To predict this, previously identified significant predictors are factorized and given as input to the model. The SVM model executes with an accuracy of 88% and sensitivity of 97%.

```
SVM_1 <- ksvm(isexpensive ~ age+bmi+smoker+exercise+hypertension,
              data=trainSet,C = 8,cross = 3, prob.model = TRUE)
SVM_1
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2187  303
         1   63  447

               Accuracy : 0.878
                 95% CI : (0.8658, 0.8895)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6358

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

            Sensitivity : 0.9720
            Specificity : 0.5960
         Pos Pred Value : 0.8783
         Neg Pred Value : 0.8765
             Prevalence : 0.7500
         Detection Rate : 0.7290
   Detection Prevalence : 0.8300
      Balanced Accuracy : 0.7840

       'Positive' Class : 0
```

Recommendations to People:

We have seen through the visualizations and predictions that BMI and smoking have the highest stake in determining the cost and expense of a person. Therefore, we have curated recommendations in such a way that following these, people can effectively reduce their expenses in the following year.

Body Mass Index:

BMI is calculated by dividing the weight of a person by the square of their height. Calculation of BMI tells us the risk of being affected by any disease and it is very important to maintain a correct BMI, here are a few ways of achieving it:

- Look out for hidden sugar and try to avoid them to the maximum extent possible.
- Get your heart pumping and that can be through involving into cardio activities.
- Do not believe or bother with crash diets.
- Eat and snack on healthy food.

Smoking:

The act of inhaling and exhaling the fumes of burning plant material is termed as smoking. Tobacco contains nicotine, an alkaloid that is addictive and can have a tranquilizing psychotic effect. Continuous exposure to tobacco can lead to life-threatening diseases. Following are a few measures individuals can take to overcome this addiction and help themselves by reducing the costs of health care:

- Try Nicotine replacement therapy.
- Start with exercising.
- Use Nicotine patches/ gums
- Avoid triggers that instigate/ encourage smoking

Recommendations to HMO:

- Recommendation for the health management organization would be to collaborate with the health rehabilitation centers to help people get rid of their smoking habits.
- HMO can try to come up with health care plans and provide necessary benefits.
- Pricing of the customers ca be based on their profile having a model that considers their BMI, smoking habits, exercising history, etc.

Future Considerations:

To venture more into this dataset and provide more insights that could be more accurate, we might need some additional information from the HMO. If we can collect data of people if they are enrolled in any healthcare plans or taking any healthcare benefits, it can help us with more clear data on how much an individual is spending and then categorize them as expensive or not expensive.

Furthermore, in this project we assumed the threshold for expensive based on our own analysis. If we were given the correct value or constraints over which HMO considers a person is expensive, we could have provided more practical actionable insights.

Links:

Below are the links of all the relevant files:

Link to the shiny App:

HMO vizualization and Prediction (shinyapps.io)

Code for Shiny App:

Final_app.R

Presentation:

Final Presentation IST 687 IDS.pptx

Knit file of all the visualizations:

Visualization file.pdf

Knit file of the models:

Model file.pdf

Screenshot of successful deployment: