**IST 687 INTRODUCTION TO DATA SCIENCE**

Lab Section M002 | Group 4

**DATA ANALYSIS FOR HEALTH MANAGEMENT ORGANIZATION**

School of Information Studies, Syracuse University  
Prof. Erik Anderson  
December 8th, 2022

Submitted By:

Choijamts Bataa | Komal Regmi | Noah Goldie | Pralhad Vaishnav | Shriya Gawade

1. **Description**

The project revolves around inspecting the data that was collected from a health insurance plan in the United States. We will use the survey results to see which features will likely impact the health care cost in regard to increasing or decreasing total cost.

1. **Project Background**

The dataset that we are using for our project is obtained from a Health Management Organization. The file includes details of an individual with respect to their age, exercise activity, smoking habits, location, bmi, yearly physical activity, hypertension, gender, education level, marriage status, number of children, and past healthcare cost. The dataset was provided to us by our instructors. It has 7582 rows and 14 columns.

The link for the data file: <https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv>

1. **Objectives**

In analyzing the HMO dataset, we believe we will be able to gain more insight on which individual or group of people would likely spend more money on health care costs. Our team will use various techniques to understand the key drivers with respect to the expenses and the contributing factors associated with it. Additionally, we have identified trends and provided suggestions as to which corrective actions aids in lowering the health cost for our clients.

1. **Data Cleaning**

* The “X” column in the dataset didn’t supply us any context. Thus, we decided to remove it.

Code below:

hmo = hmo %>% dplyr::select(-X)

* As we were scanning, we noticed that two of the columns, BMI and Hypertension were missing some values. Total number of the missing values were 158 with 78 from BMI and 80 from Hypertension. Due to the low percentage of the data, we decided to omit the rows.

Code below:

hmo = hmo %>% filter(!is.na(bmi), !is.na(hypertension))

* We decided to turn all the categorical variables into factors to better understand.

Code below:

hmo = hmo %>% mutate(location = as.factor(location),

location\_type = as.factor(location\_type),

education\_level = as.factor(education\_level),

yearly\_physical = as.factor(yearly\_physical),

exercise = as.factor(exercise),

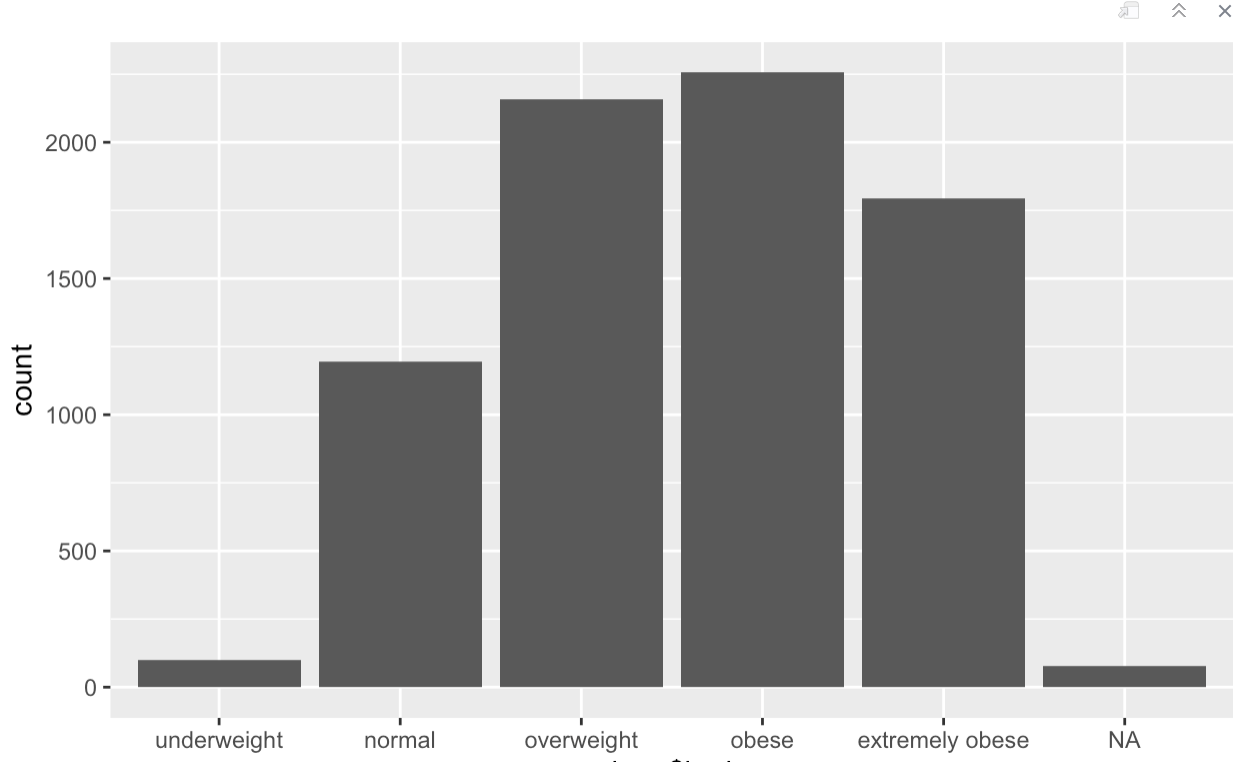
married = as.factor(married),

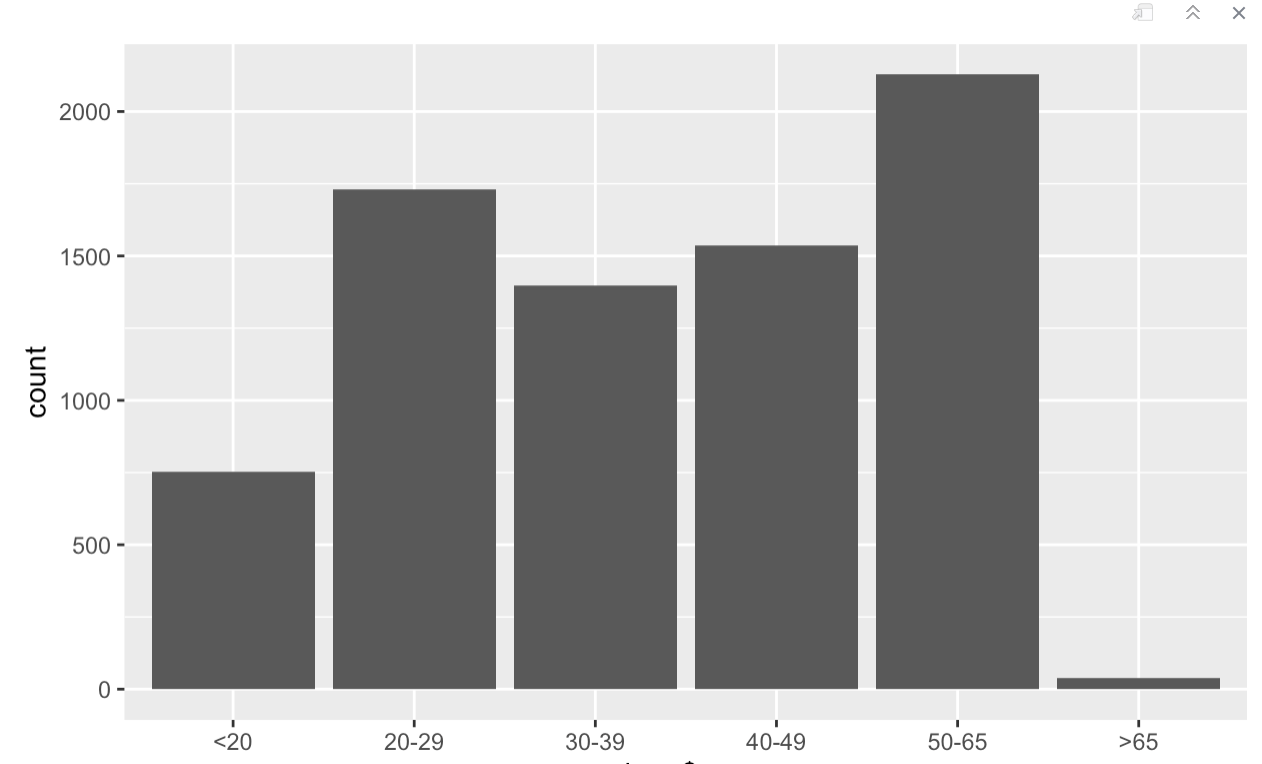
gender = as.factor(gender))

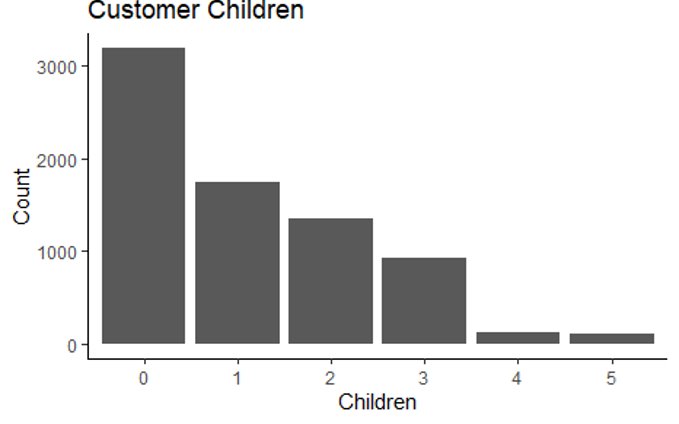
1. **Understanding Data**

For comparison

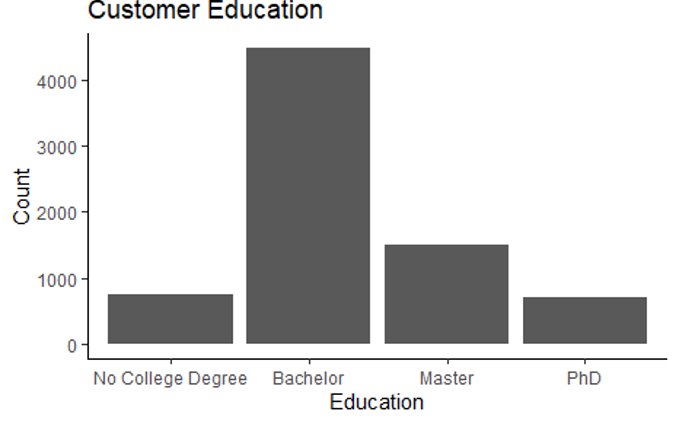
* We decided to classify the BMI into categories for instance, for an individual with a BMI of 18.5 or under we put them as underweight.



* For the age column we labeled the ages with ranges such as 20-29 or 50-65. Doing this gave us a clear view in organizing our data pieces. 
* Similarly, we looked at the number of children the customer had. Mostly, there were no children and at the second most, the customer had one child.

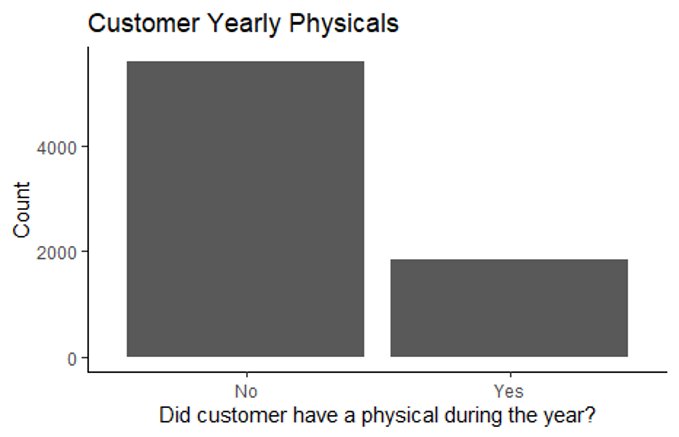


* The level of education that frequently appeared in our data was customers that had bachelor's degree which reveals the customer had some level of education for a large part.



For better context and knowledge of the data, we performed comparative analysis to see the differences or similarities.

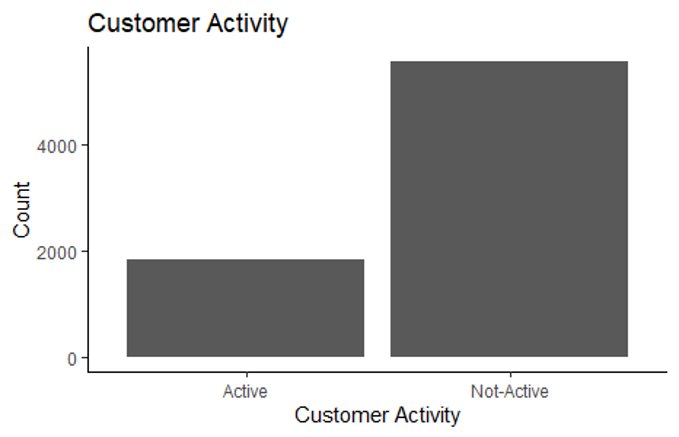
* We also looked at customer yearly physical. As displayed in the bar graph below. Most of our customer regardless of their gender did not have an annual physical exam.



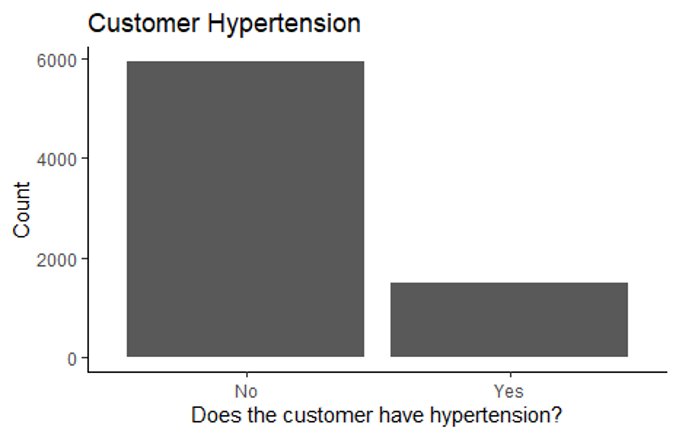
* When inspecting customer smoking habits and it appears that most of them are nonsmokers.



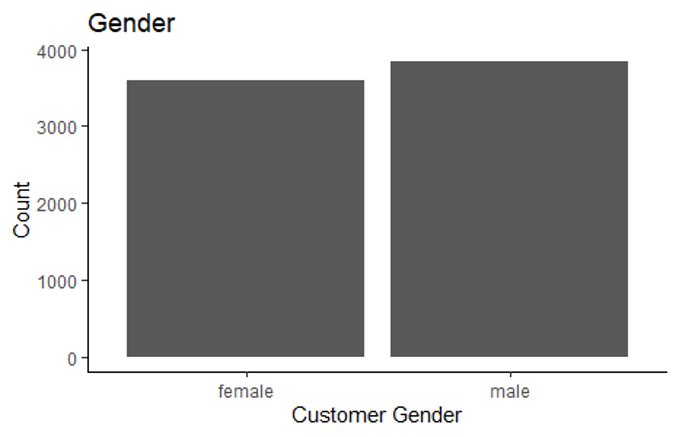
* As for physical activity level, more than 4,000 of our customers were non-active.



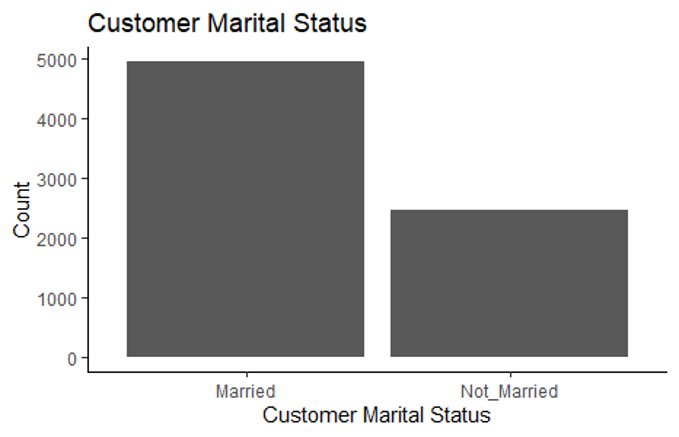
* A great number of the customers did not have hypertension or high blood pressure.



* The male and female ratio are close to one another which sets impartial view in which we can treat the data equally.



* It appears that close to 5,000 customers are married and around 2,500 customers are unmarried.



1. **Defining cost threshold**

Multiple group members used 4625 as their threshold to define expensive and some used 3230. After multiple attempts, we decided to use 5000. In our analysis, we deleted the cost column and replaced it with the threshold we set for defining expensive cost. Cost greater than 5000 is expensive.

Code below:

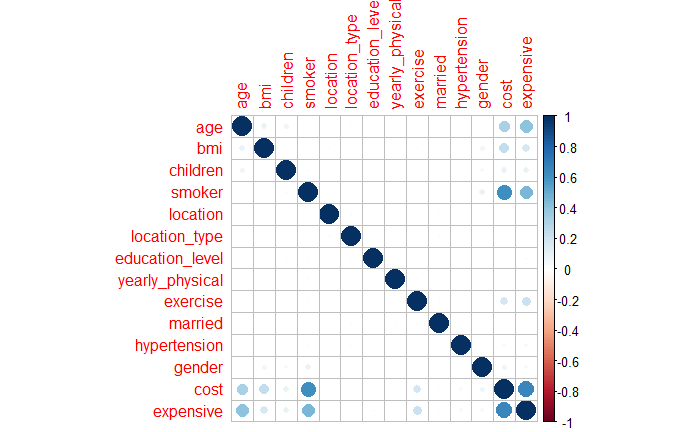
cost\_threshold = 5000

hmo = hmo %>%

mutate(expensive = factor(ifelse(cost > cost\_threshold, "yes", "no")))

1. **Data analysis**

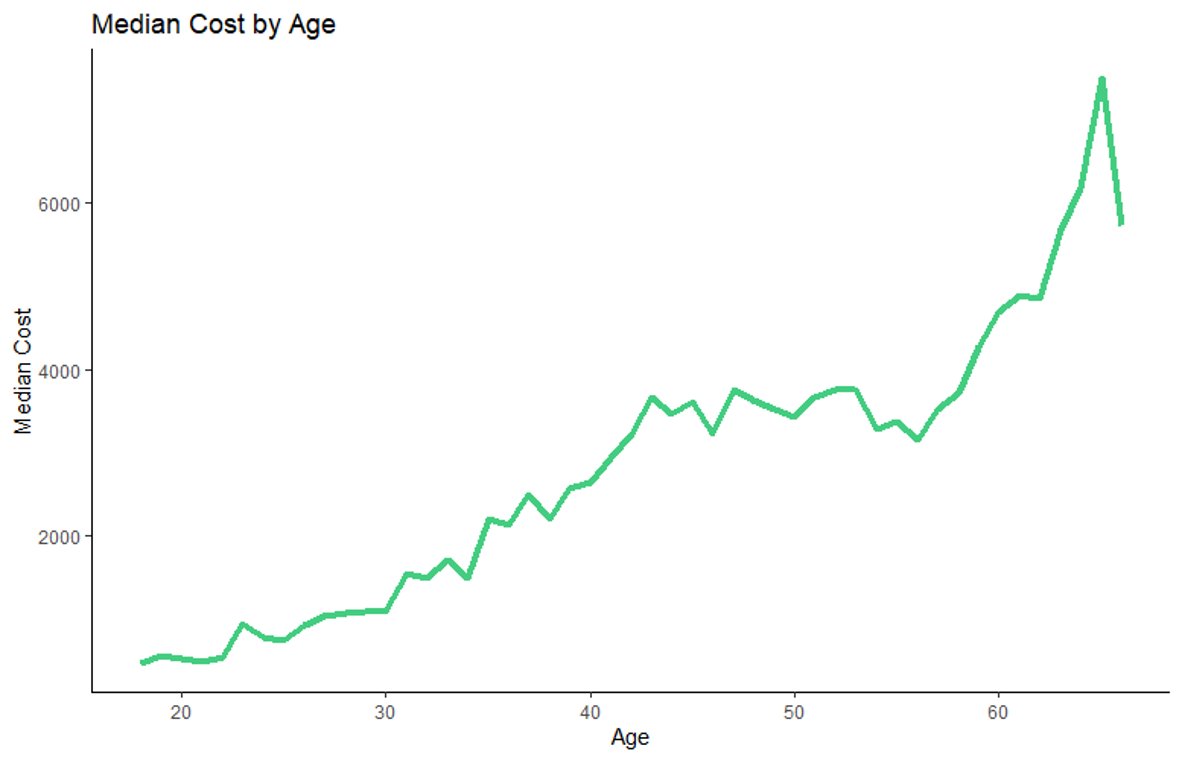
In order to understand the relationship between the variables, we decided to generate a correlative matrix plot.



Per our correlation plot, the variables that have a relationship with the cost variable are

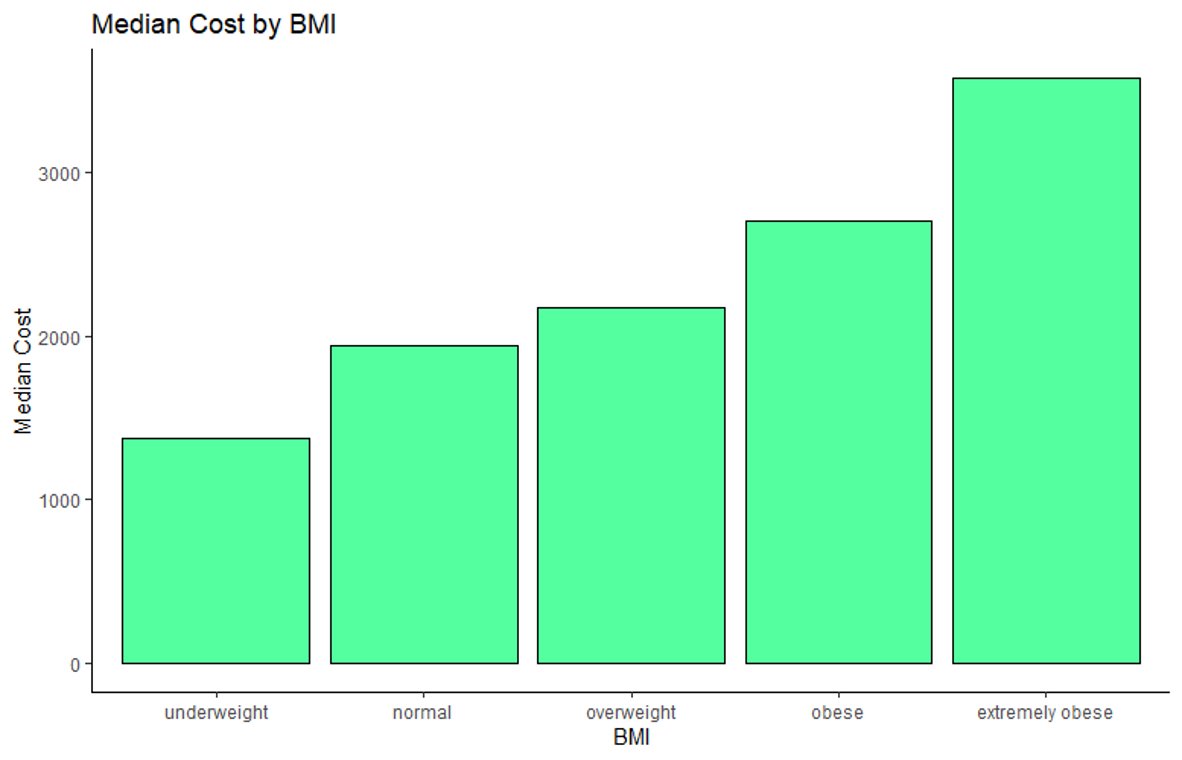
* Age
* BMI
* Smoker
* Exercise

To obverse the customers' age and the cost trends, we agreed to utilize a line graph.

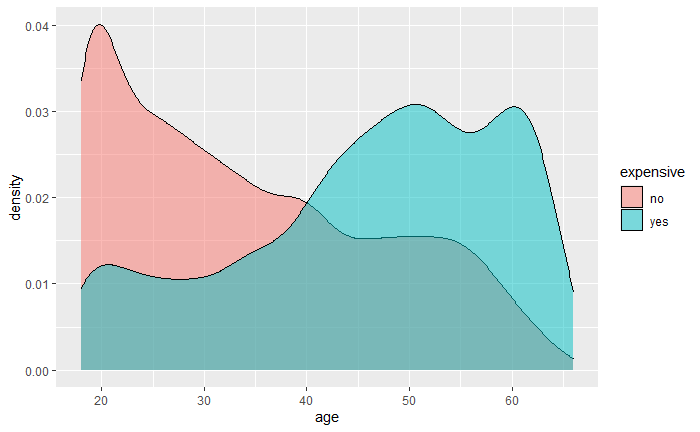


The line graph above clearly conveys that as age increases, the median cost will likely increase. Older adults will have higher health costs versus younger adults.

Another trait that impacted an individual's health care cost was their BMI, body mass index. As demonstrated below, it is evident that with a higher BMI, the median cost of healthcare will only rise. The normal cost for someone whose BMI is over 25 is around 2,000$. The highest annual cost is for extremely obese customers who pay over 3,000$.



The density plot displays the distribution of age and how expensive health care cost would be.



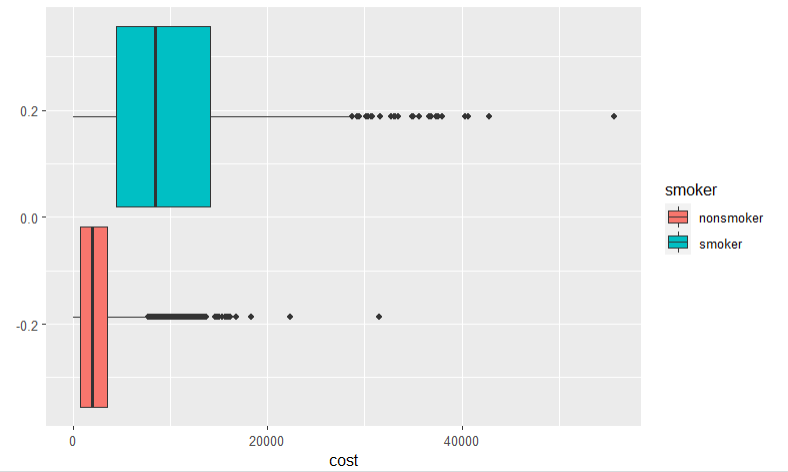
As shown above, the ages of expensive vs inexpensive customers are exhibited very differently:

* + - Expensive customer skew much older (In BLUE)
    - Inexpensive customers skew much younger (In PINK)

1. **Data Visualization**

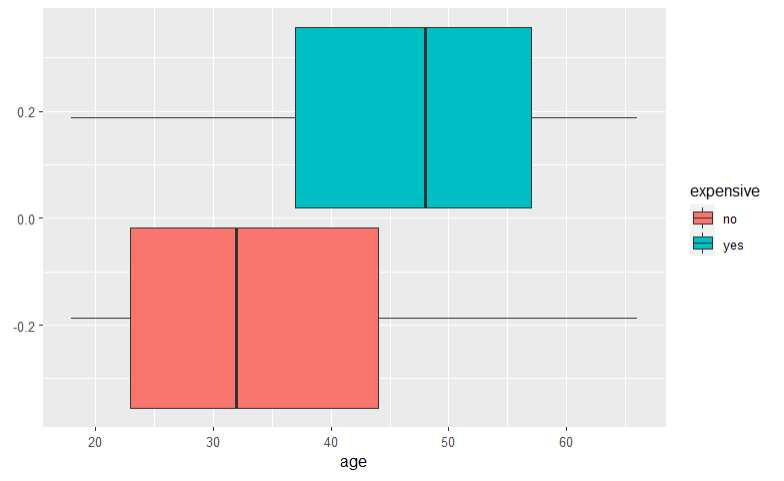
To begin with, we visualized our dataset using plots and maps. By applying simple plotting, we could summarize and study relationships between two continuous variables in our dataset. The core idea was to obtain expensive and non-expensive groups based on following parameters:

1. Smoker
2. Age
3. Bmi
4. Exercise



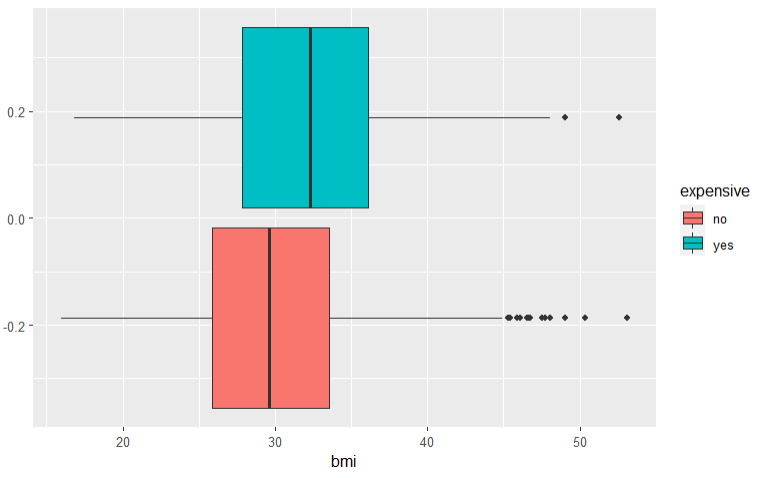
**Fig: Box plot graph Cost vs Smoker**

Insight from the graph: It is clearly evident from the box plot graph that smokers tend to cost more than the ones who are non-smokers. The average non-smoker costs lower than 3000$ whereas the average smoker cost more than 9000$ where we see more of the population of smokers above 12000-15000$



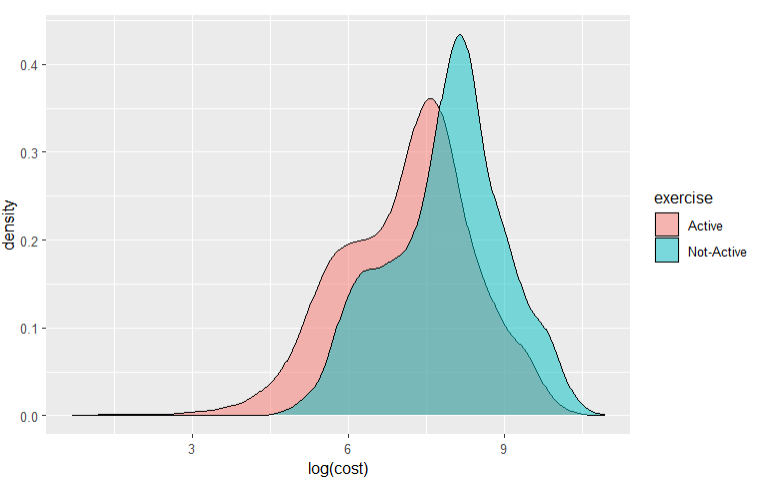
**Fig: Box plot graph Age vs Expensive**

Insight from the graph: The above graph depicts how the cost gets affected by age. We see that most of the people between age of 18 to 35 tends to be non-expensive whereas the population between age more than 48 tends to be expensive. We see that we have a mix of both expensive ones and non-expensive ones between age 35-48 years



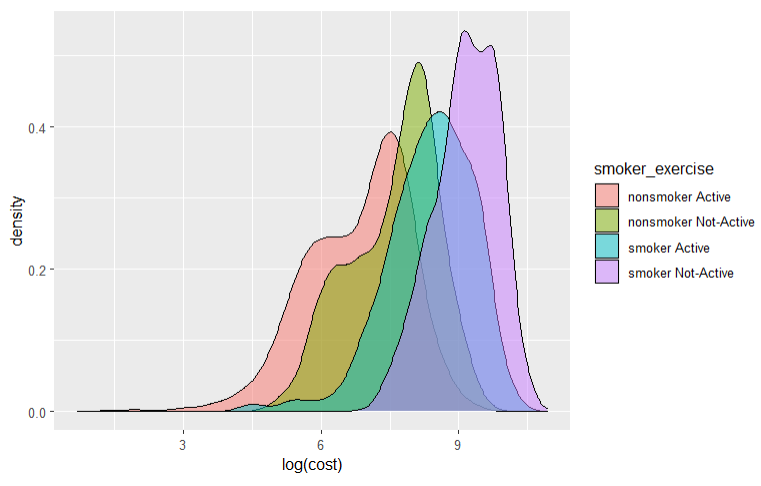
**Fig: Box plot graph Bmi vs Expensive**

Insight from the graph: The above graph shows how much of the bmi is important in differentiating the expensiveness. We see that people with normal bmi index are non-expensive whereas people who are overweight (Bmi 25-30) has a mix of both expensive and non-expensive. People who have obesity with BMI > 30 tend to be more expensive.



**Fig: Logarithmic Graph Cost vs Exercise**

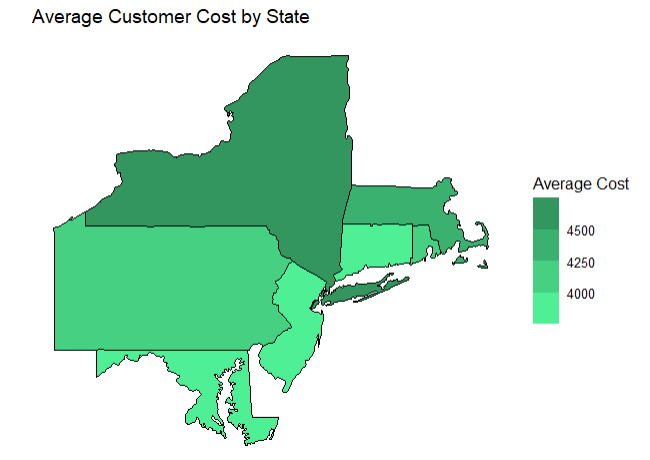
Insight from the graph: The above graph shows how much of the cost varies when it comes to population who are active and non-active on exercises. It is clear that people who are not actively exercising have a high risk factor in terms of expensiveness.



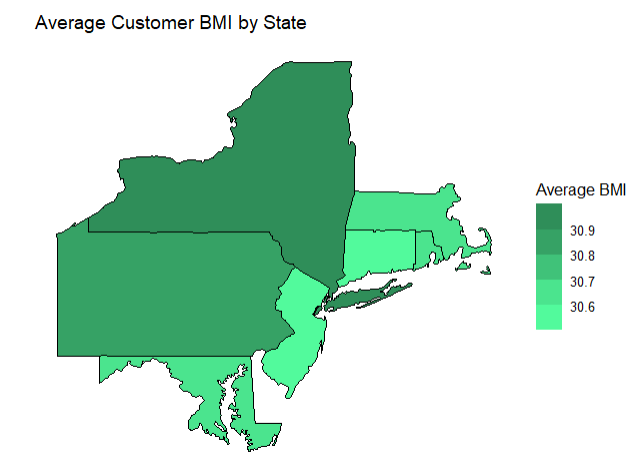
**Fig: Logarithmic Graph Cost vs Smoker\_Exercise**

Insight from the graph: The above graph shows how much of the cost varies when it comes to population who are both active and non-active on exercises and smoking. We can infer from the above graph that population who are non-active on exercising and an active smoker has the highest risk factor of expensiveness, whereas population with active on exercising and a non- active smoker has the least risk factor. Non-smokers who are non-active on exercise have the second highest risk factor followed by smoker active population.

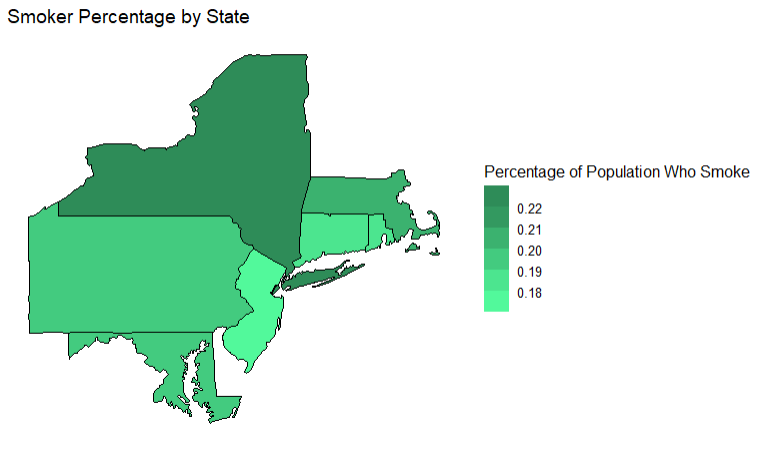
1. **Maps**



**Fig: Map graph Average Customer Cost by State**



**Fig: Map graph average customer BMI by State**



**Fig: Map graph Smoker Percentage by State**

Graph Insight: If we look at the three map graphs, we can find a similarity between all where New York state has the maximum percentage of Smokers with high BMI and cost which tends to infer that population of New York will be more expensive than compared to the other states.

The least expensive population will be Connecticut where we see that the people are less on smoking and BMI indexes.

1. **Modeling techniques:**

We have used various methods to obtain many models and better understand the bigger picture.

Model Assessment methodology

* Estimates for accuracy and sensitivity are obtained by resampling and refitting
* The training data partition is set at 0.7
* In each iteration, the data is partitioned and then fit with a model
* 100 iterations are run, with the accuracy and sensitivity of each iteration being recorded

To assess our accuracy and sensitivity in the models that we will be carrying out, we created a function. This is appropriate for us in order to rank and weigh the models.

Code below:

CI = function(acc,sens) {

mu\_acc = round(mean(acc), digits = 4)

sigma\_acc = round(sd(acc), digits = 4)

mu\_sens = round(mean(sens), digits = 4)

sigma\_sens = round(sd(sens), digits = 4)

print("Accuracy")

print(paste(" Mean:", mu\_acc))

print(paste(" 95% CI: [", mu\_acc - 1.96\*sigma\_acc, ',', mu\_acc + 1.96 \* sigma\_acc, "]", sep = ''))

print("Sensitivity")

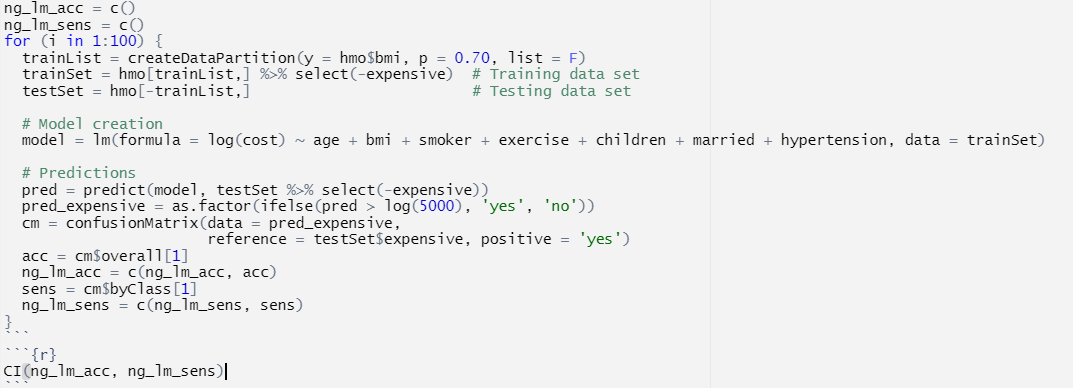
print(paste(" Mean:", mu\_sens))

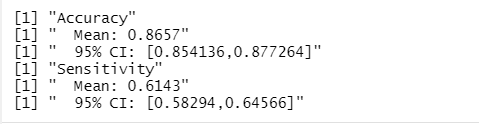
print(paste(" 95% CI: [", mu\_sens - 1.96\*sigma\_sens, ',', mu\_sens + 1.96 \* sigma\_sens, "]", sep = ''))

}

**9.1 Linear Model**

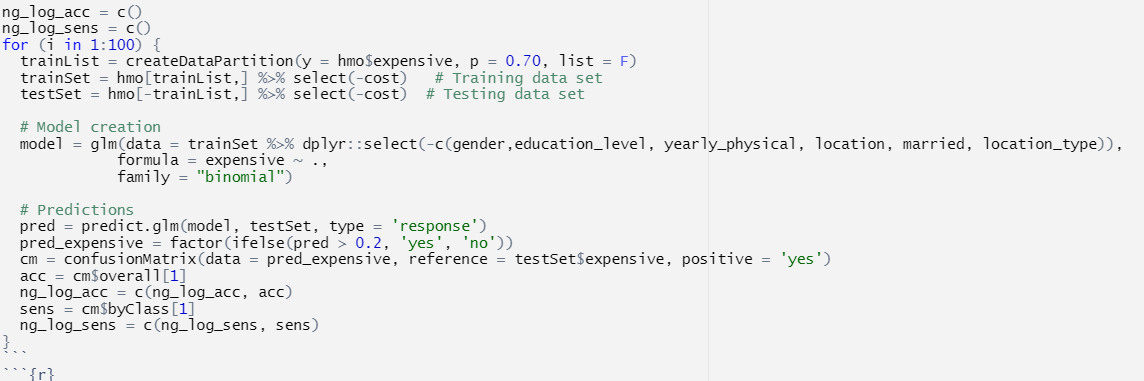
Code:



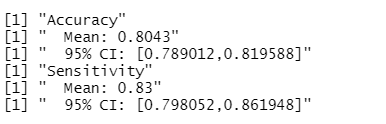
Output:  


**9.2 Logistic Regression**

Code:



Output:



**9.3 Support Vector Machine**

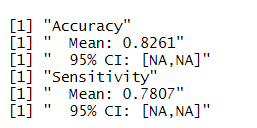
Support vector machines (SVM) are powerful techniques that let you validate your result and how accurate they are.

We split the dataset into training and testing to examine the validity.

Code:



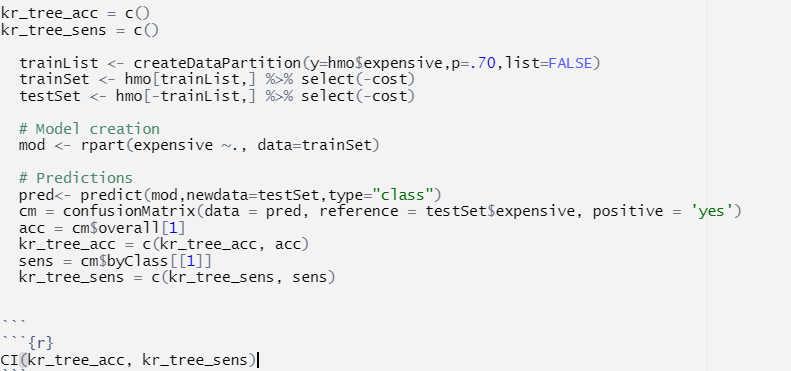
Output:



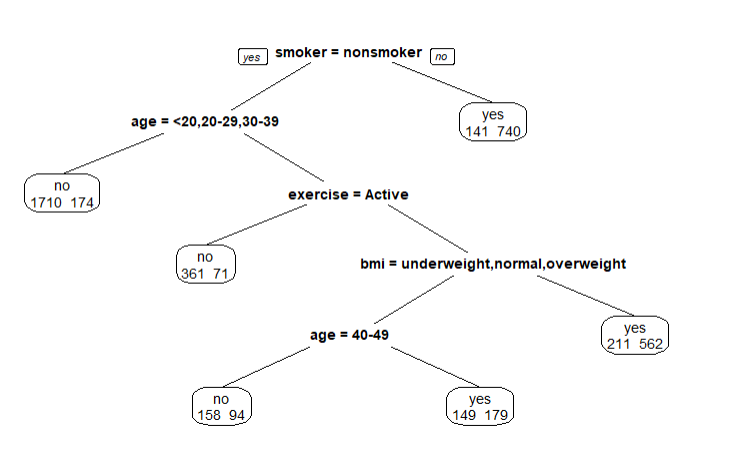
**9.4 Decision Tree**

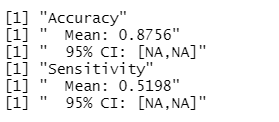
Decision Tree aids viewers in seeing the flow of the model and the possible outcomes.

Code:

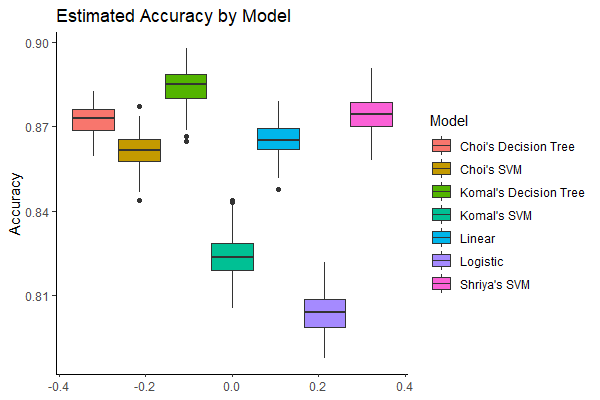


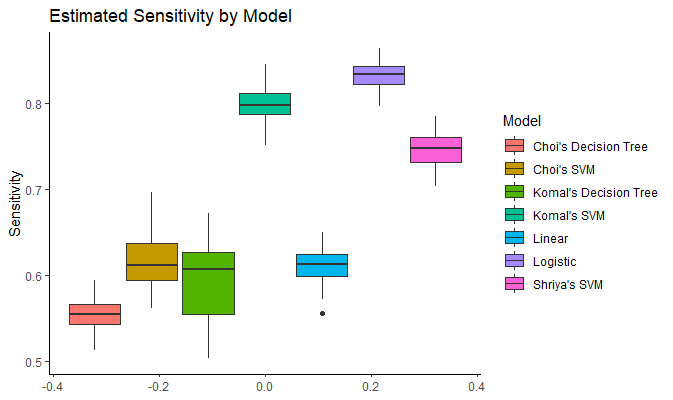
Output:





After evaluating the results, we decided to compare our models with regards to which one produces our desired outcome in both accuracy and sensitivity.





1. **Interpretations**

* Customers who smoke are far more likely to be expensive
* Customers get more expensive as they age
* Adjusting for other factors, state and location type do not meaningfully affect cost
* Customers who exercise regularly cost less
* Cost increases as BMI increases

1. **Recommendations**

* Hold or encourage fitness classes to promote exercise
* Offer discounts to customers that do not smoke
* Offer better coverage for smoking addiction therapy
* Encourage yearly checkups to increase customer awareness of common health factors
* Adjust premiums based on smoking habits, activity, BMI, and age
* Run advertisement campaigns targeting younger people to lower average customer cost