Health Management Organization

Final Project

05/09/2023

IST 687

Group 6:

Francesca Fan

Shubham Gaikwad

Sindhuja Maheswaran

Priyasha Sinha Roy

Yuqing Zhang

Project Objectives

- Understand the key drivers for why some people are more expensive (i.e., require more health care), as well as predict which people will be expensive (in terms of health care costs).

- Predict people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs).

- Provide actionable insight to the HMO, in terms of how to lower their total health care costs, by providing a specific recommendation on how to lower health care costs.

Project Agenda

For finishing this project, we have five phases.

★ Phase one is to clean the original data since we find out that there are some not available values in the dataset. We want to make sure that the results of the project are more accurate after cleaning.

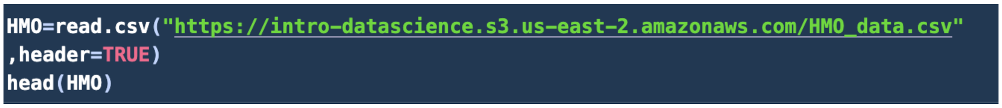
★ Phase two is exploring the data by histograms and boxplots of numeric.

★ Phase three is an overview of the important variables. In this data, we have 12 independent variables in total, but we want to know which variables are affecting the total cost more or the most. Finding out the important variables can help us to do the suggestions effectively.

★ Phase four is doing the data modeling and trends for predicting if a person is "expensive". The last phase is giving an accurate conclusion and recommendations.

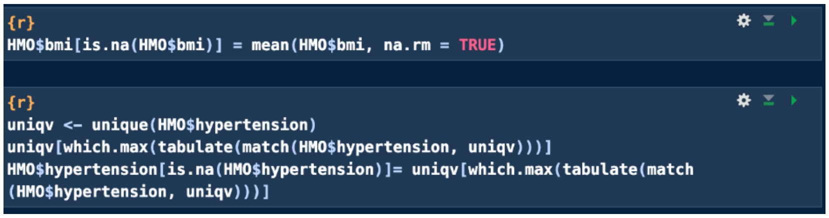
→ Data Cleaning

First, we import the original data into R studio by using "read.csv", then we use "head ()" to check the data again.



```
HMO=read.csv("https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv"
,header=TRUE)
head(HMO)
```

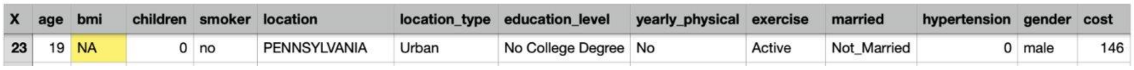
Second, we use "which(is.na)" to check each variable if there any unavailable values. Then we find out that there are two variables contain unavailable values, which are "bmi" and "hypertension". For bmi, we used average value of others in bmi to replace unavailable values. For hypertension, we used mode value of others in hypertension to replace the unavailable values since it is a dummy variable.



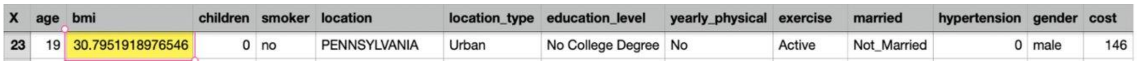
```
{r}
HMO$bmi[is.na(HMO$bmi)] = mean(HMO$bmi, na.rm = TRUE)
```

```
{r}
uniqv <- unique(HMO$hypertension)
uniqv[which.max(tabulate(match(HMO$hypertension, uniqv)))]
HMO$hypertension[is.na(HMO$hypertension)]= uniqv[which.max(tabulate(match
(HMO$hypertension, uniqv)))]
```

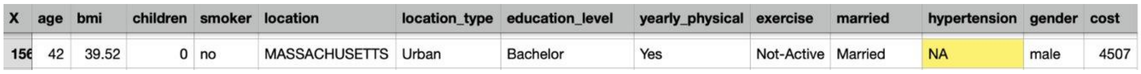
In the end, we check the data again to make sure that all the unavailable values are replacing by the average values and mode values.



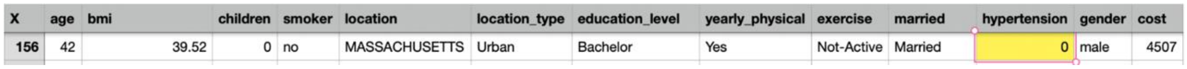
X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
23	19	NA	0	no	PENNSYLVANIA	Urban	No College Degree	No	Active	Not_Married	0	male	146



X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
23	19	30.7951918976546	0	no	PENNSYLVANIA	Urban	No College Degree	No	Active	Not_Married	0	male	146

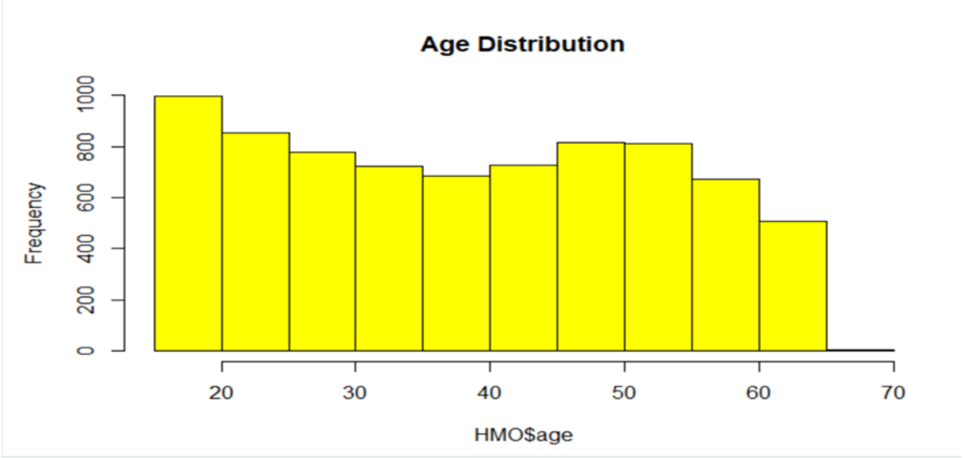


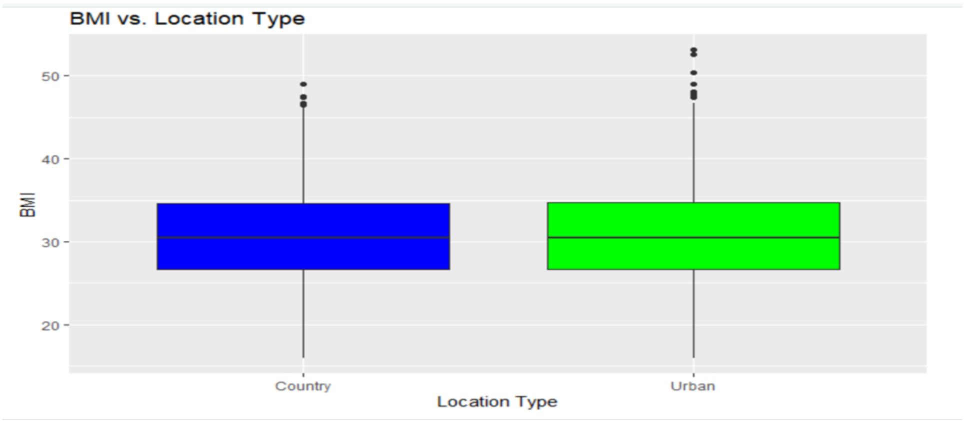
X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
156	42	39.52	0	no	MASSACHUSETTS	Urban	Bachelor	Yes	Not-Active	Married	NA	male	4507



X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
156	42	39.52	0	no	MASSACHUSETTS	Urban	Bachelor	Yes	Not-Active	Married	0	male	4507

- **Exploratory Of Data – Histograms**





- ## Overview of Important Variables

1. Age: Age of the person will play major role in analysis here as Senior people have more health issues than young ones.

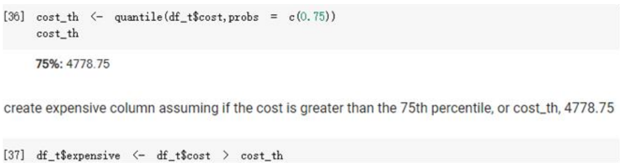
2. **BMI**: Body Mass Index defines a person's height with respect to height. This is one of the important factors in finding how a person is doing in the health aspect.

3.**Smoker:** Smoking leads to many health issues hence it is important to know if the person is into smoking or not.

4.**Exercise:** Active exercise people have a lower chance of facing medical issues than non-active exercise people.

- ## Data Modeling and Trends

- Based on the fact that over 75th percentile, there is a very long tail in the box plot
- We model expensive:
  - True, if the cost>= 75th percentile, or 4778.75
  - False, the other condition

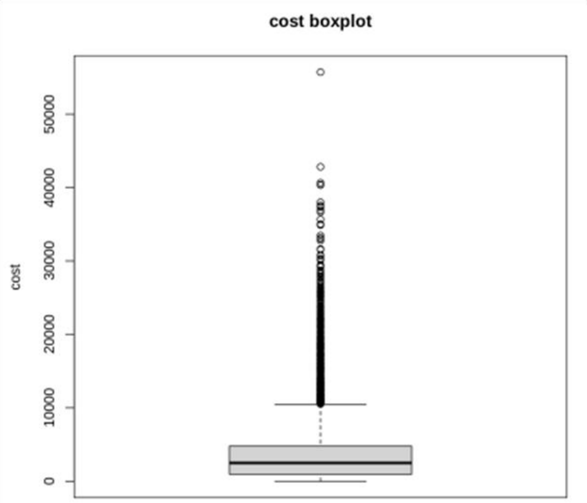


```
[36] cost_th <- quantile(df_t$cost,probs = c(0.75))
     cost_th

     75%: 4778.75
```

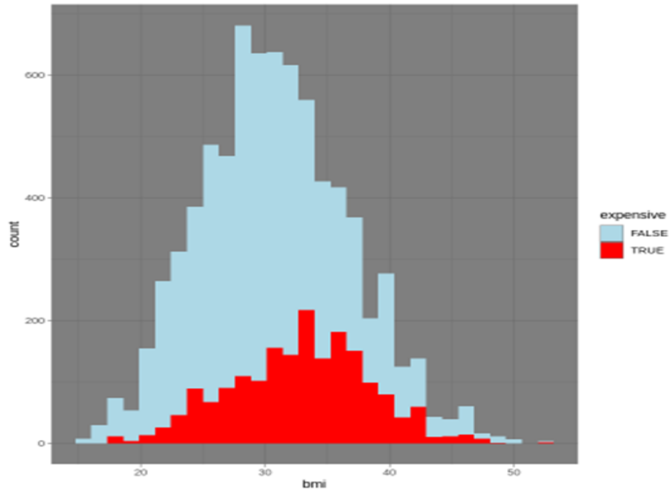
create expensive column assuming if the cost is greater than the 75th percentile, or cost_th, 4778.75

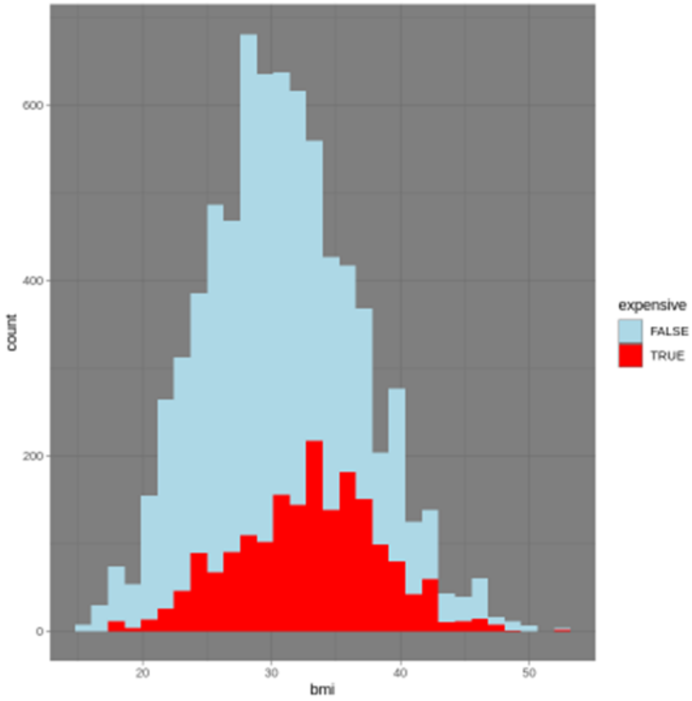
```
[37] df_t$expensive <- df_t$cost > cost_th
```



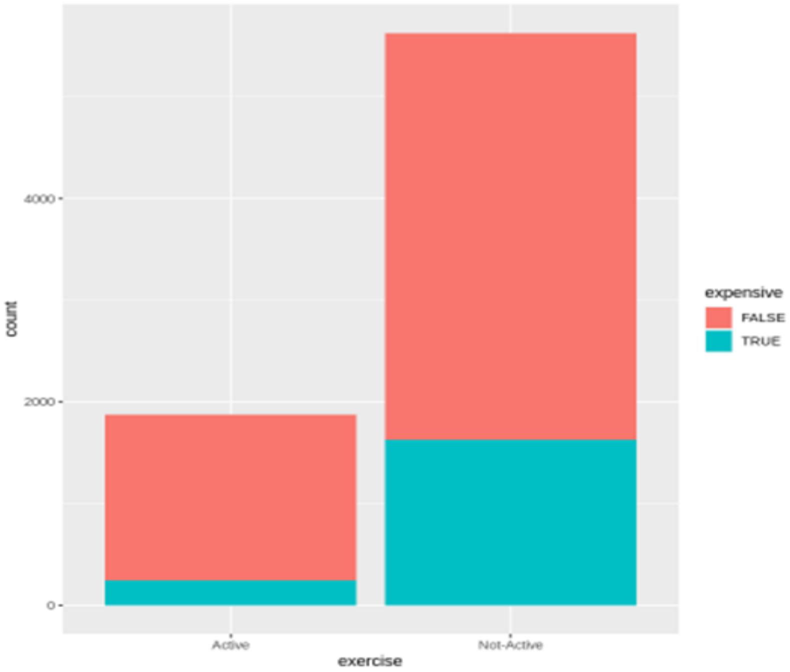
Histogram compares with expense

→ BMI

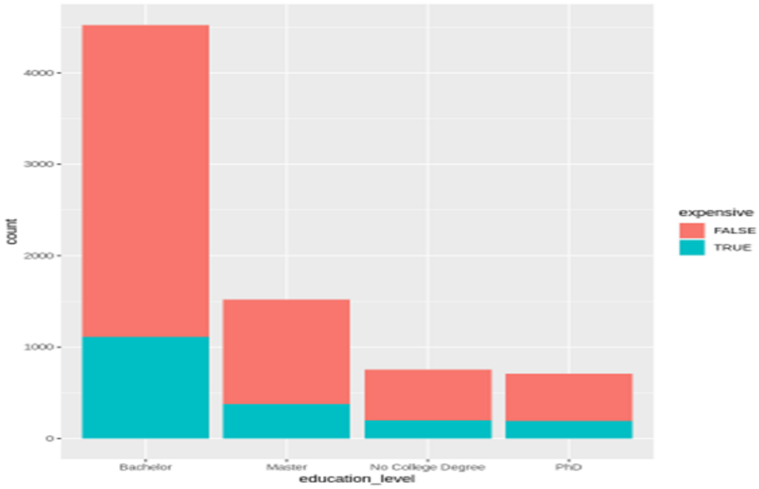




→ EXERCISE



→ EDUCATION LEVEL



- **Machine Learning Model 1, SVM**

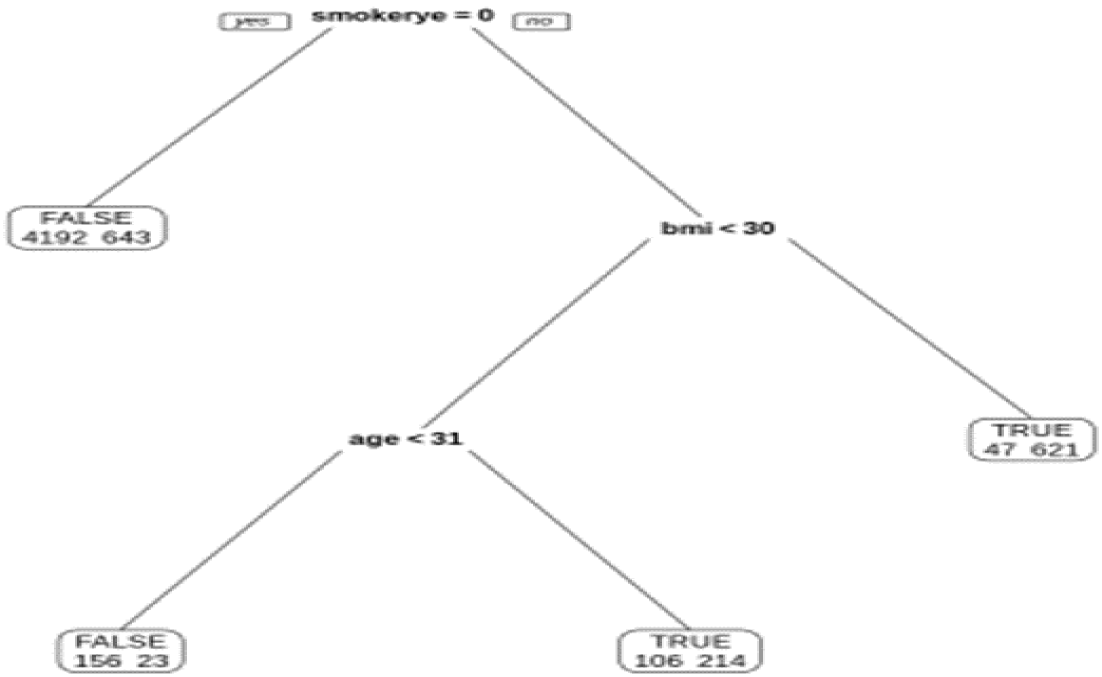
● "Support Vector Machine" (SVM) is a supervised learning machine learning algorithm, mostly used in classification problems.

●In this case, we feed all the input, build, and test the SVM classifier:

- expensive **TRUE OR FALSE**
- Model accuracy: 88.08%

- Prediction accuracy: 86.68%
- 95% prediction accuracy CI: 84.85% to 88.36%

- **Machine Learning Model 2, CART**



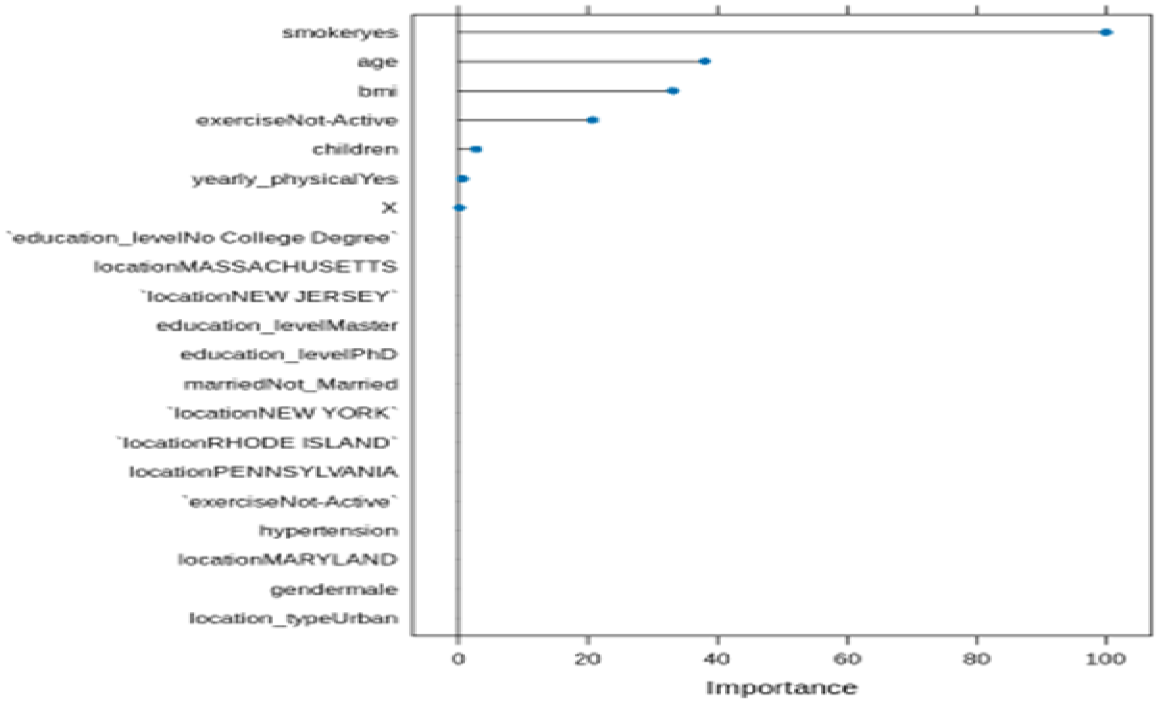
●Classification And Regression Trees (CART) is another widely used Machine Learning model employee decision.

●Model accuracy: 88.4%

●Prediction accuracy: 86.07%

●95% prediction accuracy CI: 84.21% to 87.78%

- Machine Learning Model 2, CART Result



●The interpretations of the decision tree can be found in the plot on the right

●The following four factors contribute most to high expensive cost:

- Active smoke
- Senior people
- High bmi
- Non-active

Conclusion and Recommendations

Based on our analysis and understanding as future data scientists, if we want to reduce the health care cost, for HMO we need to first narrow our focus to our target audience (the ones who are the end of the highest medical expenditure) and then locate the causes behind the expenditures. Once we did the above, we found the below 4 causes as the most prominent ones. Hence, we suggest the following 4 categories of people as the desired group for our analysis.

○**Active smokers: We can morally help them by suggesting them to find more constructive ways to de-stress. It can be working out in any form or just finding a passion.**

○**Senior people: Since ageing cannot be stopped and is a very natural thing. This category of people needs the utmost care and respect for their service thus far.**

○**High BMI: We can recommend them to go to therapy and find their actual reason of stress so that they can first work on their mental health and then it should automatically help them to work on their health.**

○**Non-active: Get up and Work!**