THE iSCHOOL
Syracuse University

Analyzing Healthcare Cost Information from a Health Management Organization

Parth Gulavani
Rutu Waghela
Anurag Paradkar
Kishan Rathor
Whitaker Ellis

Introduction

Objective:

- We serve as a consulting firm for HMOs (Health Management Organizations), which are medical insurance groups that offer health services in exchange for a set annual charge.
- Our objective is to identify: the main factors behind why some people need more medical attention than others, those who will spend a lot of money on healthcare in the upcoming year, and finally offer the HMO specific advice on how to cut costs to lower their overall health care expenses.

Business Objectives

Predict people who will spend a lot of money on health care next year (i.e., which people will be considered “expensive”).

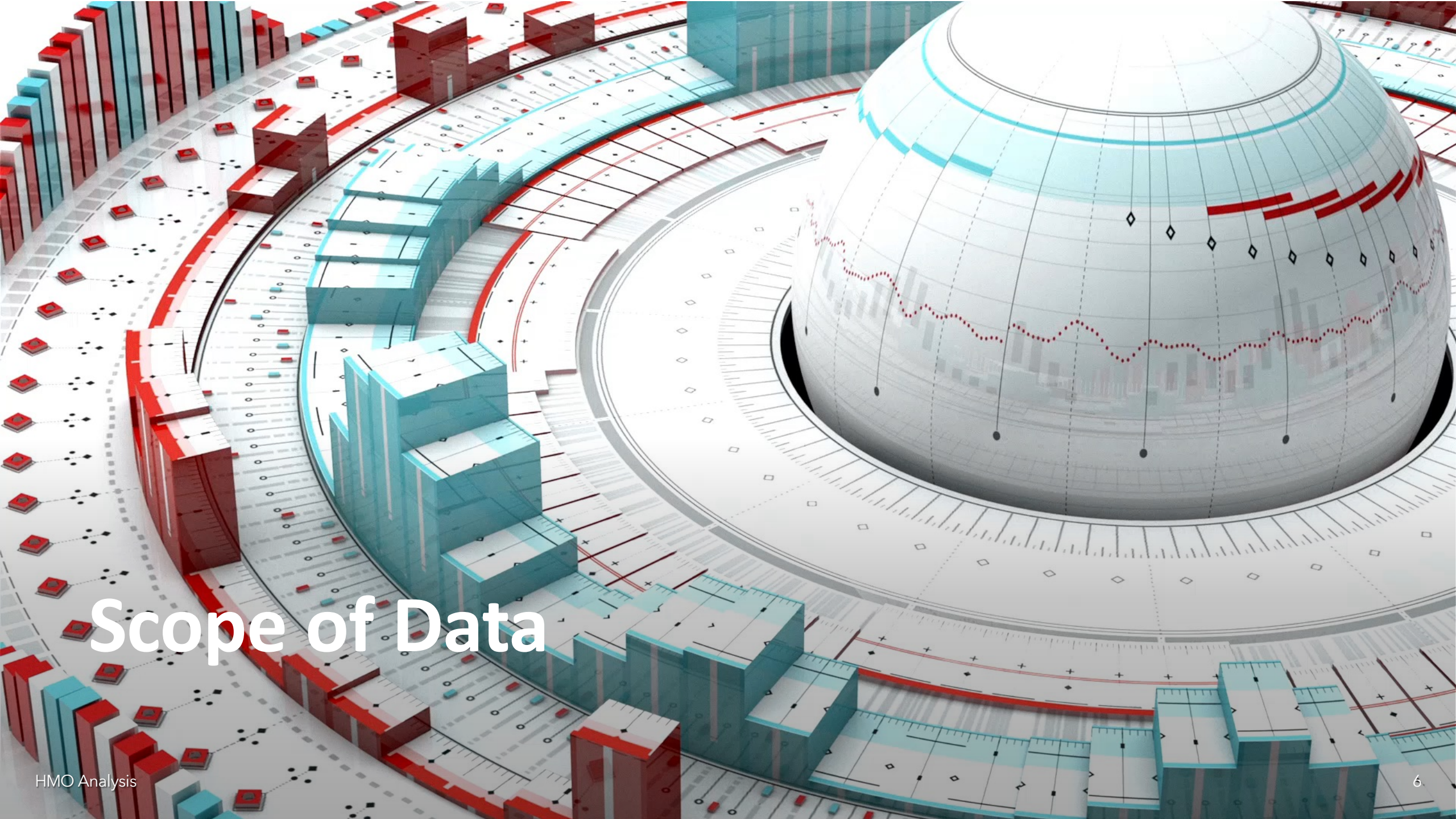
Provide actionable insight to the HMO on how to lower their total health care costs. We will do this by providing a specific recommendation that will help them accomplish this goal.

Approach

- Load the data
- Explore the data
- Clean the data
- Data Modeling and Trends
- Conclusion and Recommendations

Loading Data

- Using the dataset provided to us by the professor, we copied the dataset and created a .csv file out of it.
- The dataset contains healthcare cost information from an HMO (Health Management Organization).
- This data set has 14 variables and 7,582 observations.
- We imported the dataset into R Studio using the read.csv ( ) function, and stored it in a new data frame named “hmodata” which helps us ultimately observe the new data frame



Scope of Data

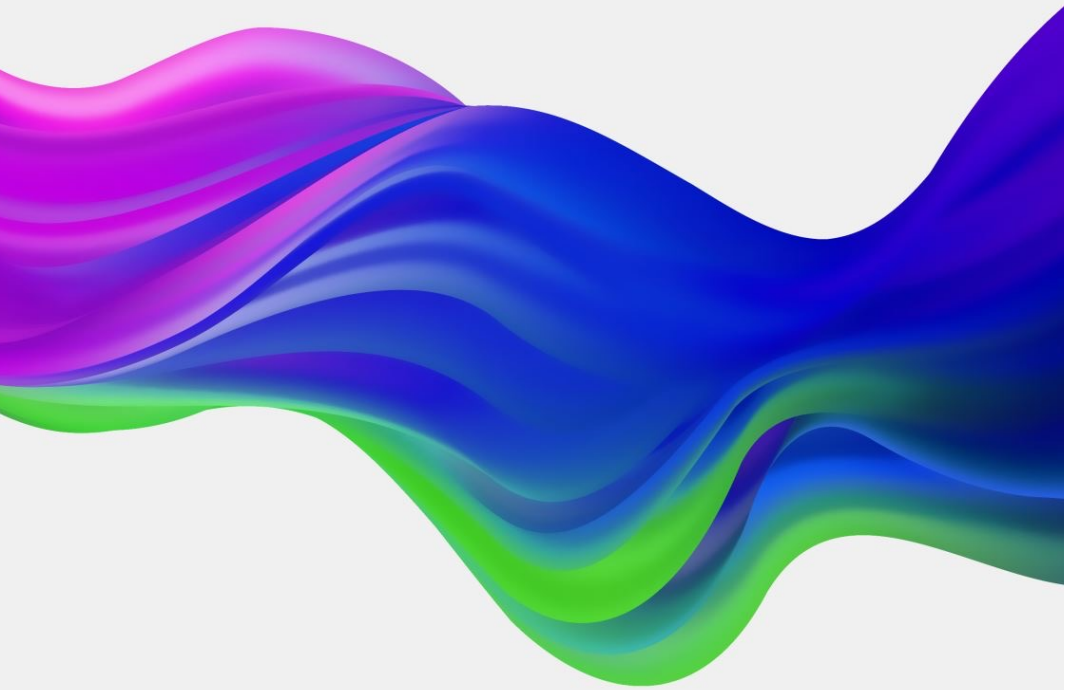
- First, we chose columns from the given data set and categorized them into three different sections:

Individuals Basic Info	
Variable	Description
x	Unique Identifier
age	Age of the person at the end of the year
Gender	Gender of the person
education_level	The amount of College Education
married	Marital Status of the individual
num_children	Number of children

Individuals Geographical Information	
Variable	Description
location	US States
location_type	Urban or Country

Individual Health Information	
Variable	Description
exercise	If the person exercises actively or not
smoker	If the person smokes or not
hypertension	If the person has hypertension or not
bmi	Body Mass Index of the person
yearly_physical	If the person visited their doctor during the year or not
cost	Total healthcare cost for that person, during the past year

Cleaning the Data

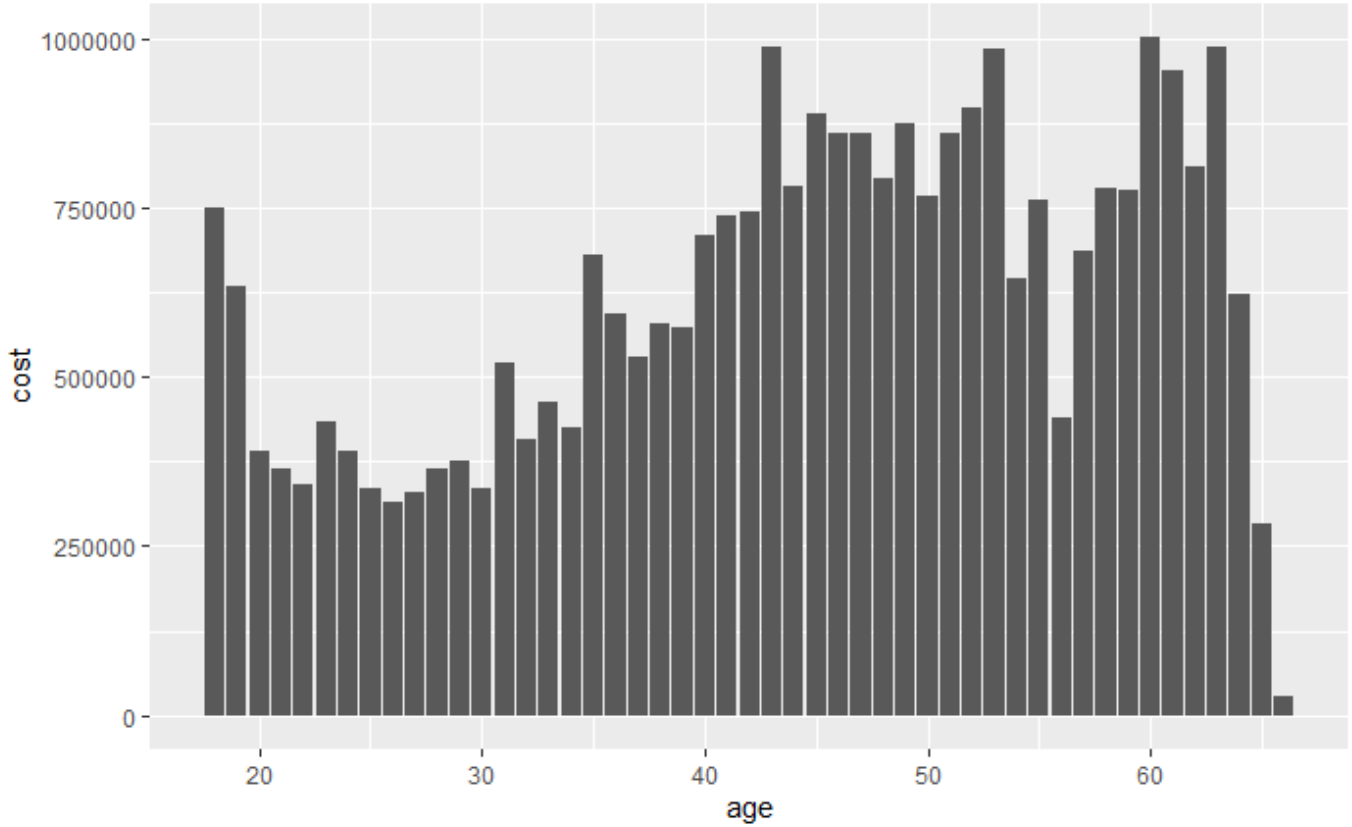


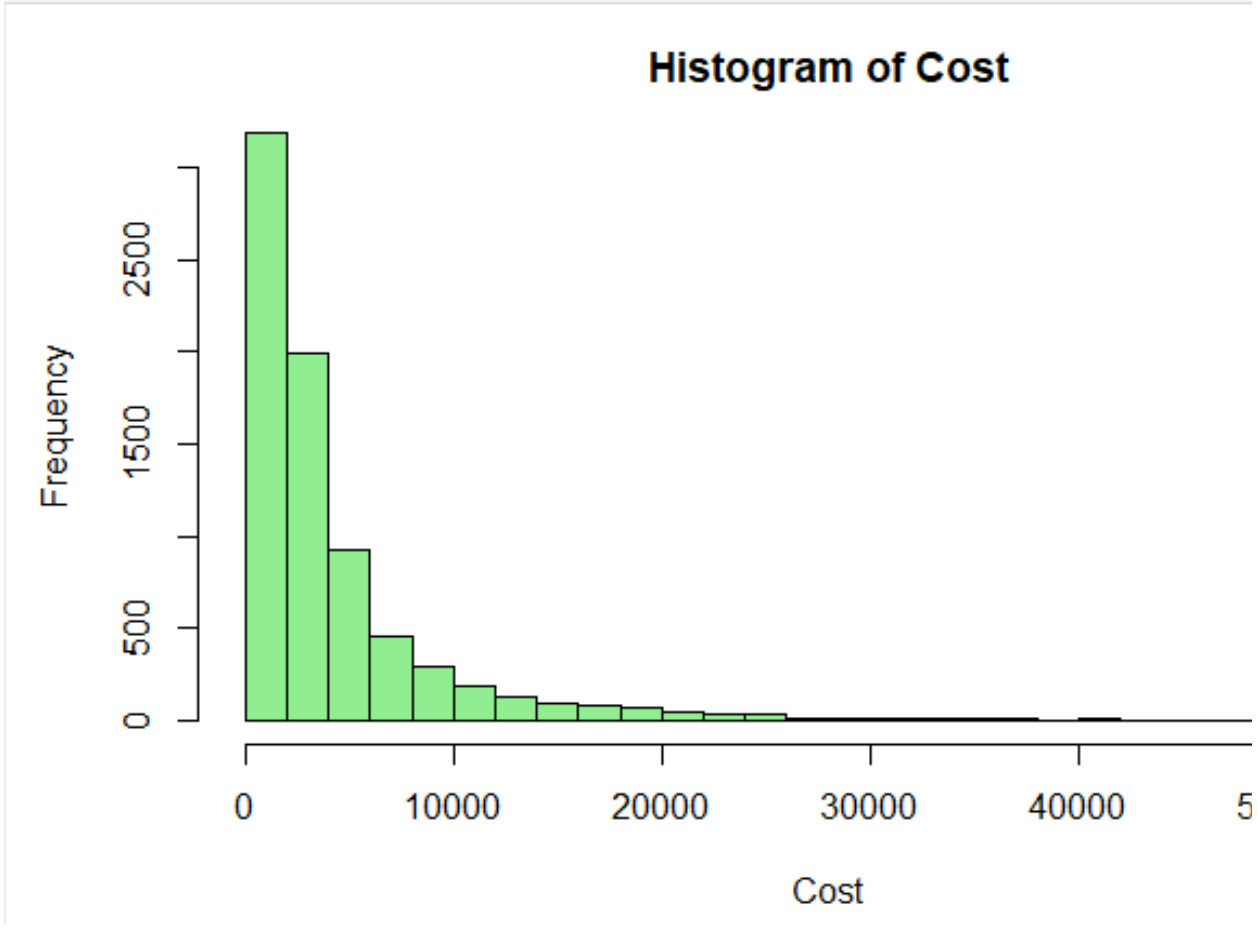
- In the dataset we were provided, we observed there are multiple missing values in the **“bmi”** and **“hypertension”** columns.
- In order to solve this issue we used na_interpolation function to eliminate null values in the columns that needed it.

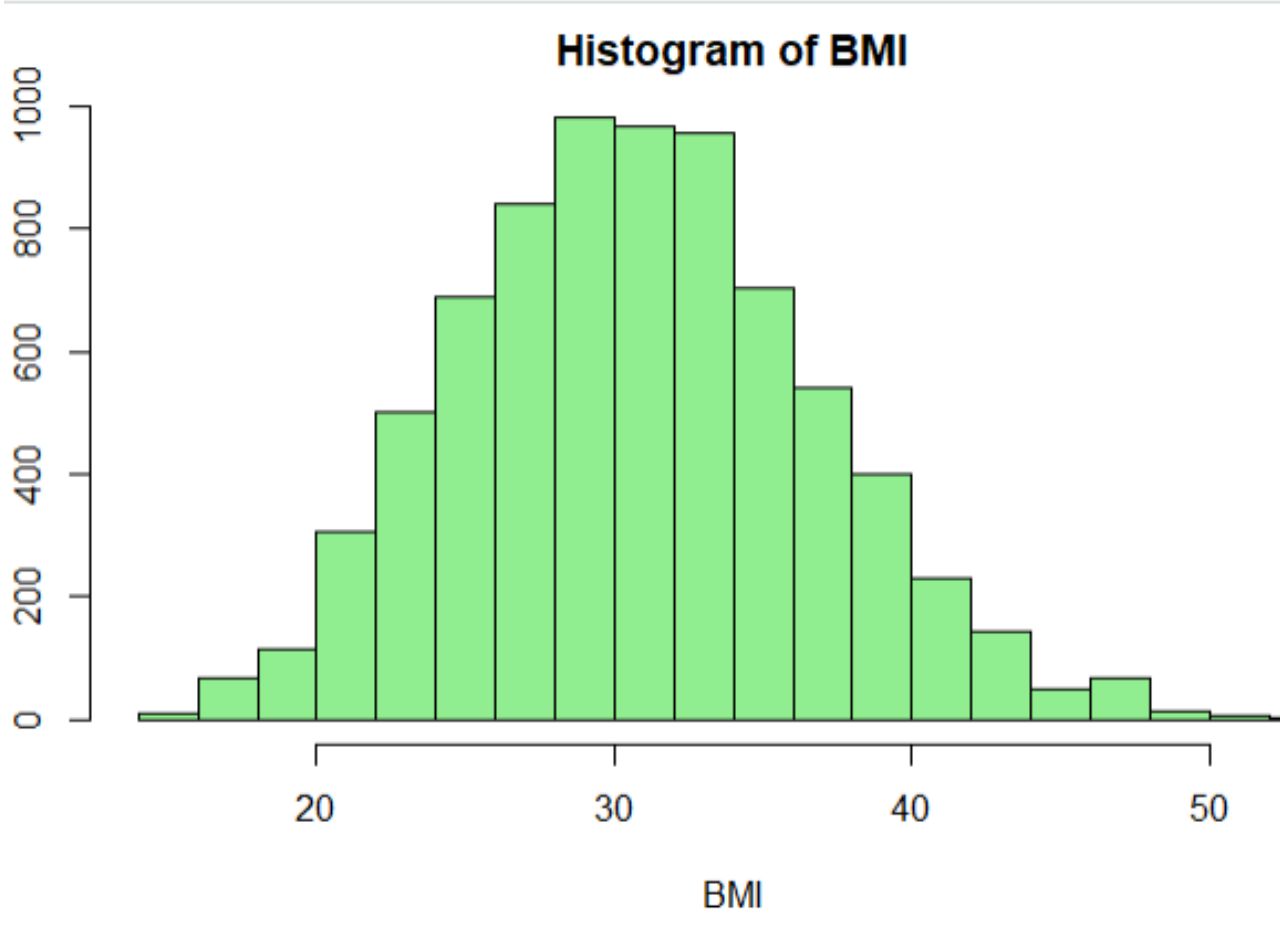
Data Exploration and Visualization

Bar Plot of Categorical Variables

- Comparing cost of healthcare for individuals based on age

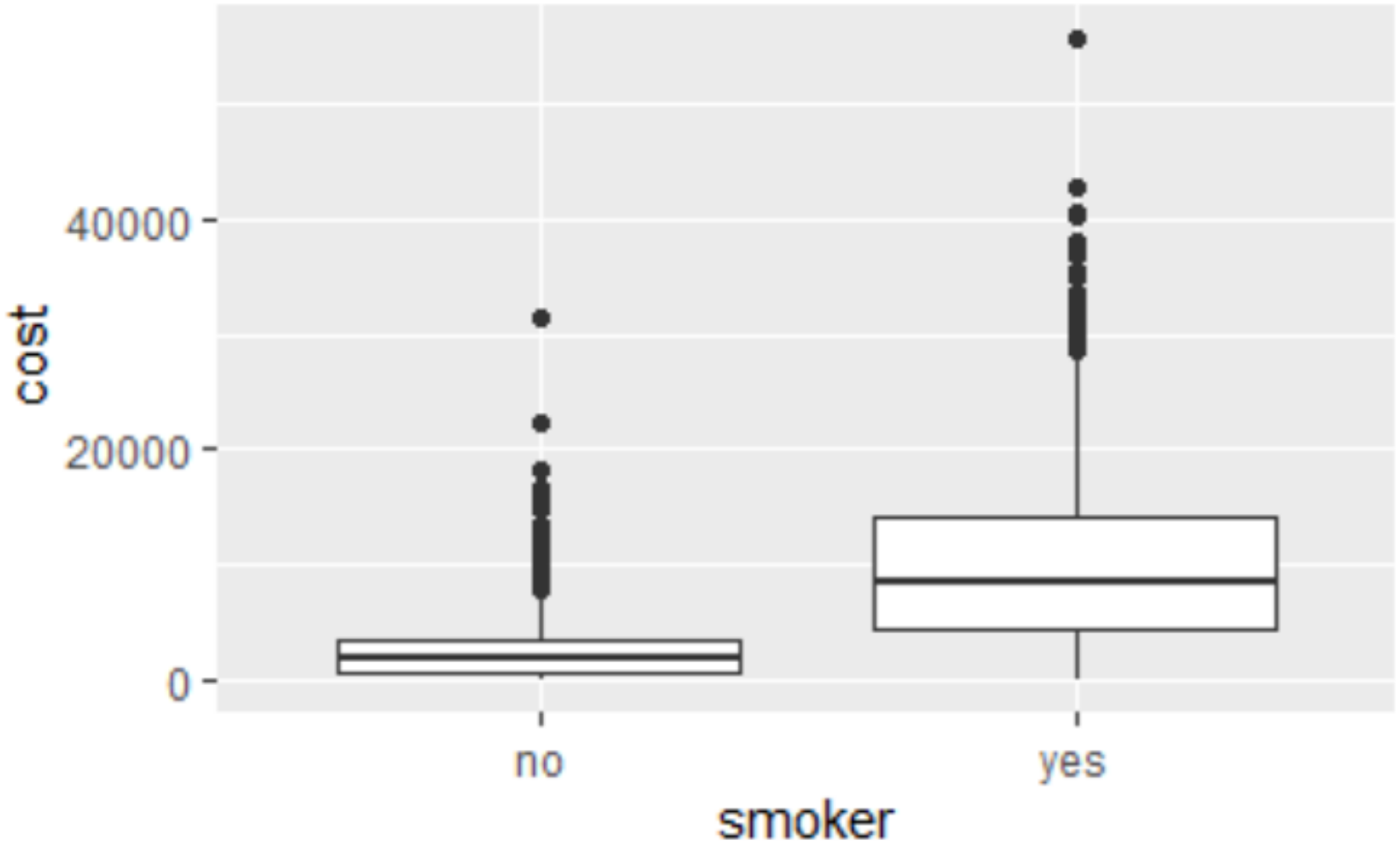




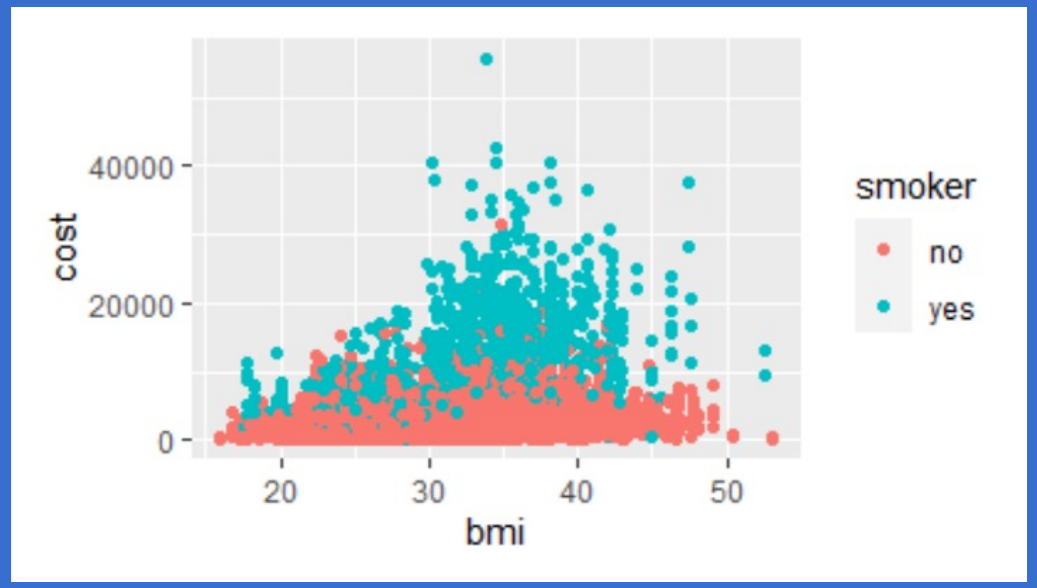


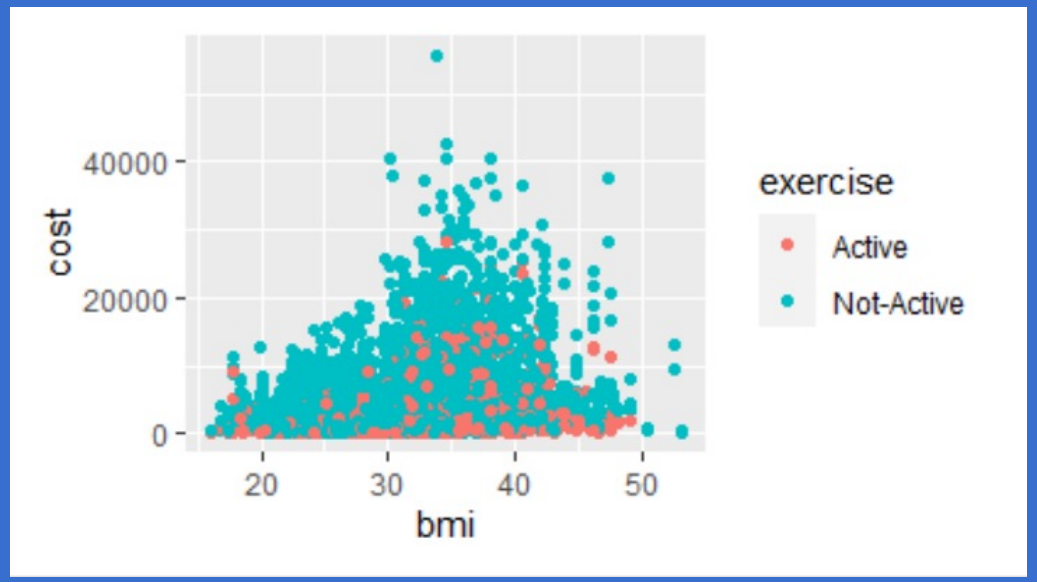
Histograms Observing Distribution of Quantitative Variables

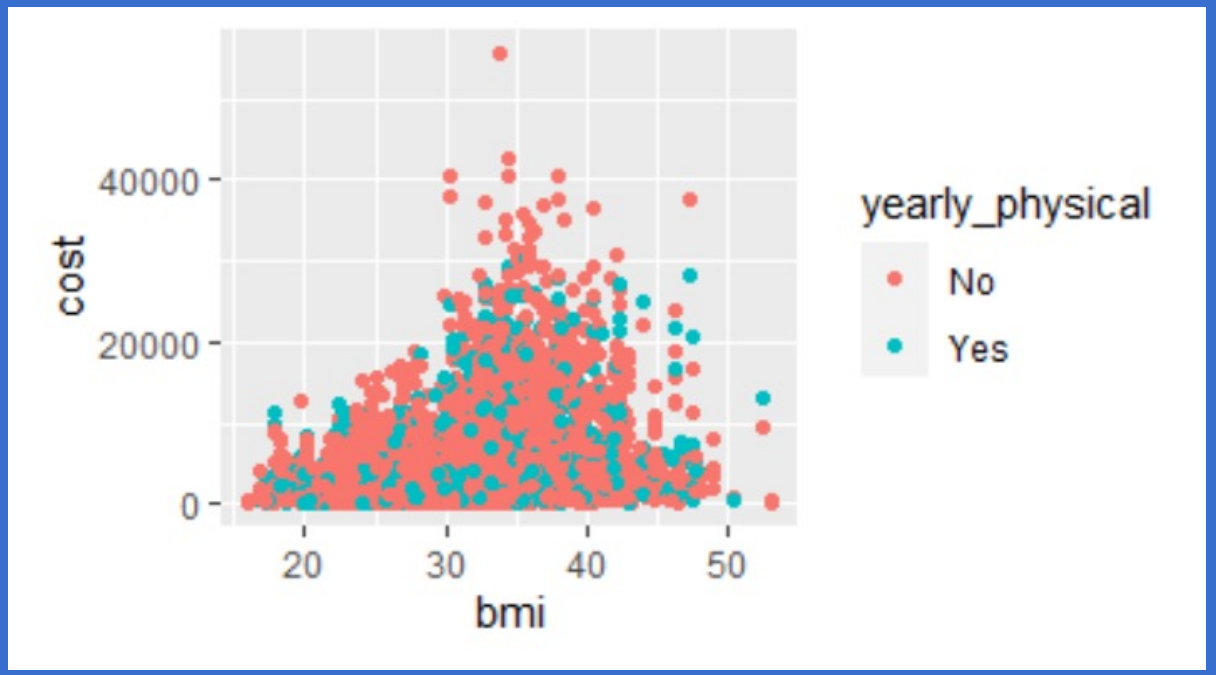
Box Plot



Scatter Plots








Machine Learning
Models

SVM Model

```
hmodata_svm1 <- train(cost_status ~
X+age+bmi+children+smoker+location_type+education_level+yearly_physical+exercis
e+married+hypertension+gender, data = trainSet ,method =
"svmRadial",trControl=trainControl(method ="none"), preProcess = c("center",
"scale"))
```

- From this we can say that for our SVM model we consider these attributes from our dataset to predict cost status.
- We implemented a SVM radial method using the general SVM function.
- With this model we get:
  - Accuracy of 85.88%
  - Sensitivity of 96.84%

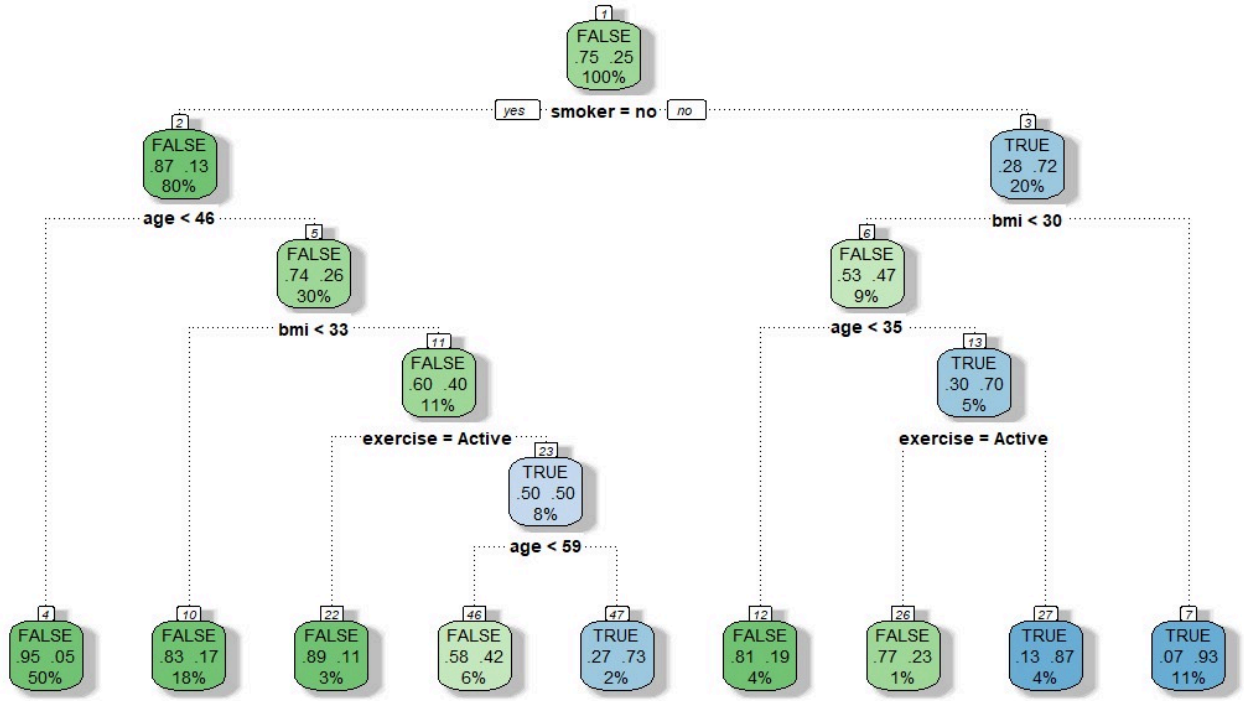
KSVM Model

```
hmodata_ksvm1<-ksvm(data=
trainSet,cost_status~X+age+bmi+children+smoker+location_type+education_level+ye
arly_physical+exercise+married+hypertension+gender, C=5, cross=3,
prob.model=TRUE)
```

- A regression model to predict how an output is predicted based on other variables in the data set
- We used KSVM for our prediction model.
- This resulted in a model sensitivity of 97.66% and a model accuracy of 87.73%

R-Part Tree Model

- We used the rpart and rpart.plot packages to create this model
- This model had 88.3% accuracy and 97.48% sensitivity.




```
170  Treeplot<-rpart(cost_status ~
     X+age+bmi+children+smoker+location_type+education_level+yearly_physical+exercis
     e+married+hypertension+gender, data = trainSet)
```

Actionable Insights and Conclusion

- From the earlier graphs, we can see that young adults have a higher cost, We could provide free health checkups for them so that they would join HMO
- As we have seen from our scatterplots, individuals who are smokers and don't exercise have significantly higher costs on healthcare when compared with costs for non-smokers. Therefore, we can make a recommendation to suggest smokers join a partnered gym giving heavy discounted rates, motivating individuals to adopt a healthy lifestyle and minimize their cigarette consumption.
- From our analysis we can see that people who undergo yearly physical exams have comparatively lower costs hence we can suggest mandatory yearly physical health checkups to get an understanding of any factors that could be harmful to individuals and help prevent any such circumstances from occurring beforehand

THE iSCHOOL
Syracuse University

Thank You for Listening