IST 687 M002

Group 2

PROJECT REPORT

HEALTHCARE COSTS INSIGHTS, ANALYSIS AND RECOMMENDATIONS

Team: Satwik Belaldavar, Lahari Chowtoori, Ximeng Deng, Jackson Hett, Supraja Ramachandran

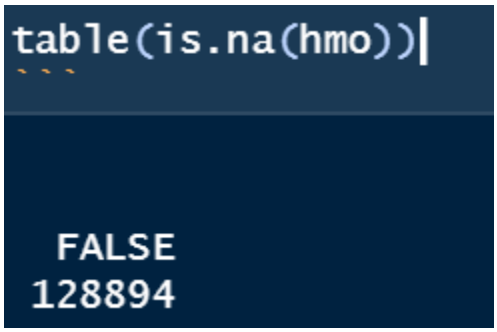
The dataset for this project contains healthcare cost information from a Health Management Organization (HMO). The goal of this project was to accurately predict which customers will be expensive and determine which factors lead to a higher healthcare cost. Additionally, we were assigned the task of creating actionable insights for a healthcare company. To do this, we used the skills developed in the course to perform data analysis.

The data used for this project is from a url of csv, and the read_csv function was used to bring in the data.

```
hmo <- read_csv("https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv")
```

	X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender	cost
1	1	18	27.900	0	yes	CONNECTICUT	Urban	Bachelor	No	Active	Married	0	female	1746
2	2	19	33.770	1	no	RHODE ISLAND	Urban	Bachelor	No	Not-Active	Married	0	male	602
3	3	27	33.000	3	no	MASSACHUSETTS	Urban	Master	No	Active	Married	0	male	576
4	4	34	22.705	0	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	1	male	5562
5	5	32	28.880	0	no	PENNSYLVANIA	Country	PhD	No	Not-Active	Married	0	male	836
6	7	47	33.440	1	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0	female	3842
7	9	36	29.830	2	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	1304
8	10	59	25.840	0	no	PENNSYLVANIA	Country	Bachelor	No	Not-Active	Married	1	female	9724
9	11	24	26.220	0	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	201
10	12	61	26.290	0	yes	CONNECTICUT	Urban	No College Degree	No	Active	Married	0	female	4492
11	13	22	34.400	0	no	MARYLAND	Urban	Bachelor	No	Not-Active	Married	0	male	717
12	14	57	39.820	0	no	MARYLAND	Urban	Bachelor	Yes	Not-Active	Married	0	female	4153
13	15	26	42.130	0	yes	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	0	male	5336
14	16	18	24.600	1	no	PENNSYLVANIA	Country	No College Degree	Yes	Not-Active	Not_Married	0	male	382
15	18	23	23.845	0	no	MASSACHUSETTS	Urban	No College Degree	No	Active	Married	0	male	294
16	19	57	40.300	0	no	PENNSYLVANIA	Urban	Bachelor	Yes	Active	Not_Married	0	male	1382
17	20	31	35.300	0	yes	PENNSYLVANIA	Urban	PhD	No	Not-Active	Married	0	male	15058
18	21	60	36.005	0	no	PENNSYLVANIA	Urban	PhD	No	Active	Married	0	female	3384
19	22	30	32.400	1	no	PENNSYLVANIA	Urban	Master	No	Active	Married	0	female	761

This is how the data looked initially after reading it in but before any data cleaning. There were 7,852 observations of 14 variables. Before doing any data analysis, we needed to deal with NA values using the imputeTS package. We decided to repair the NA values using interpolation, as we did in class several times. There were 186 NA values before data cleaning began.



Using the str() function, we examined the dataset in more detail.

```
spc_tbl_ [7,582 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ X              : num [1:7582] 1 2 3 4 5 7 9 10 11 12 ...
 $ age            : num [1:7582] 18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num [1:7582] 27.9 33.8 33 22.7 28.9 ...
 $ children       : num [1:7582] 0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : chr [1:7582] "yes" "no" "no" "no" ...
 $ location       : chr [1:7582] "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS"
"PENNSYLVANIA" ...
 $ location_type  : chr [1:7582] "Urban" "Urban" "Urban" "Country" ...
 $ education_level: chr [1:7582] "Bachelor" "Bachelor" "Master" "Master" ...
 $ yearly_physical: chr [1:7582] "No" "No" "No" "No" ...
 $ exercise       : chr [1:7582] "Active" "Not-Active" "Active" "Not-Active" ...
 $ married        : chr [1:7582] "Married" "Married" "Married" "Married" ...
 $ hypertension   : num [1:7582] 0 0 0 1 0 0 0 1 0 0 ...
 $ gender         : chr [1:7582] "female" "male" "male" "male" ...
 $ cost           : num [1:7582] 1746 602 576 5562 836 ...
 - attr(*, "spec")=
  .. cols(
  ..   X = col_double(),
  ..   age = col_double(),
  ..   bmi = col_double(),
  ..   children = col_double(),
  ..   smoker = col_character(),
  ..   location = col_character(),
  ..   location_type = col_character(),
  ..   education_level = col_character(),
  ..   yearly_physical = col_character(),
  ..   exercise = col_character(),
  ..   married = col_character(),
  ..   hypertension = col_double(),
  ..   gender = col_character(),
  ..   cost = col_double()
  .. )
```

As we can see, there are 8 numerical columns, 1 of which is a categorical variable (hypertension). There are also 8 character columns which are strings. We then used the summary() function to further explore the data.

```
       X                  age              bmi            children
 Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000
 1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000
 Median :    24916   Median :39.00   Median :30.50   Median :1.000
 Mean   :   712602   Mean   :38.89   Mean   :30.80   Mean   :1.109
 3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.70   3rd Qu.:2.000
 Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000
    smoker            location         location_type      education_level
 Length:7582        Length:7582        Length:7582        Length:7582
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


 yearly_physical      exercise           married           hypertension
 Length:7582        Length:7582        Length:7582        Min.   :0.0000
 Class :character   Class :character   Class :character   1st Qu.:0.0000
 Mode  :character   Mode  :character   Mode  :character   Median :0.0000
                                                          Mean   :0.2005
                                                          3rd Qu.:0.0000
                                                          Max.   :1.0000
    gender              cost
 Length:7582        Min.   :    2
 Class :character   1st Qu.:  970
 Mode  :character   Median : 2500
                    Mean   : 4043
                    3rd Qu.: 4775
                    Max.   :55715
```

As seen here, the ages in the sample ranged from 18 to 66 with a mean of around 39. For bmi, values ranged from about 16 to 53 with a mean of about 31. The insurance cost has a wide range, spanning from only $2 to $55,715 with a mean of $4,043. This means that the cost variable is skewed to the right by some extremely high values.

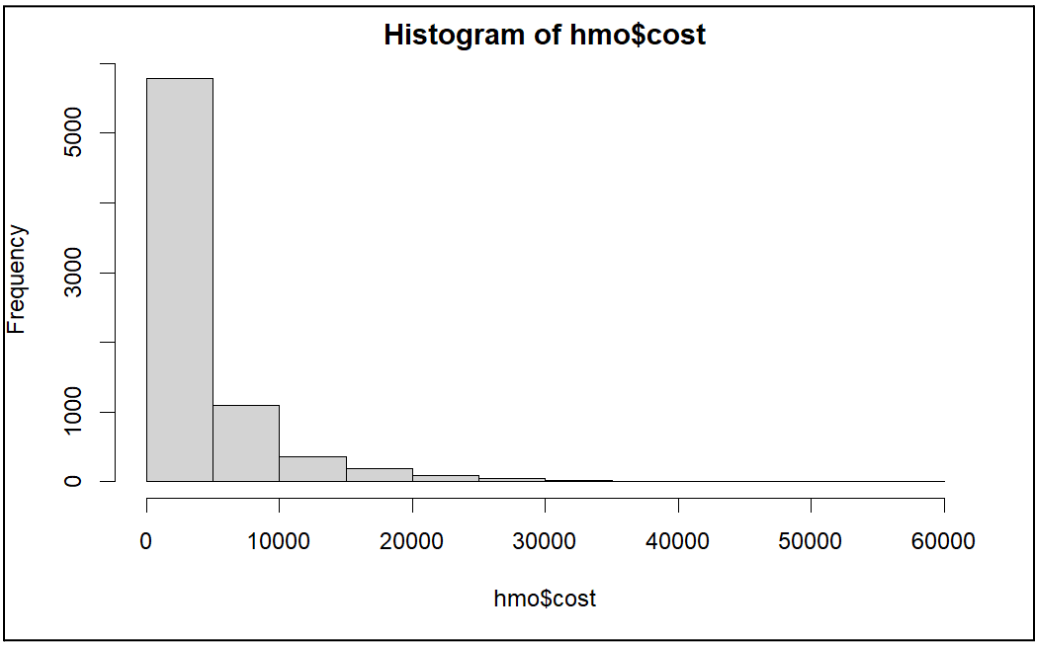
The smoker column is a binary column that denotes if that specific person is a smoker. Location type can be either urban or country. Education level has several options. These include no college degree, bachelor, master, and PHD. Yearly physical and exercise are also binary

columns with options of yes, no, active, and not-active. Hypertension is a binary column where 1 indicates a person with hypertension. Gender and married are self-explanatory binary variables as well. Cost is the amount of money a person spent on healthcare in the past year. This is our variable of interest.

We then had to create a categorical variable quantifying an “expensive person”. To do this, we ran descriptive statistics on the cost variable and found the quantile values. We chose to define expensive as a person that has a healthcare cost of $4,776 or more. The reason we picked this value was because $4,775 is the 75th percentile.

We created several other variables to further analyze the data. We created age categories including young adults, middle aged, and older adults. We did this to investigate differences in healthcare costs among different age brackets. We also created bmi categories using the defined ranges of bmi. These categories include underweight, healthy, overweight, obese, and extremely obese.

Following the exploratory data analysis and initial cleaning, we decided to further explore the dataset using visualizations. First, we made a simple histogram of the cost variable to observe the heavy skew to the right of the data. This confirmed our initial hypothesis. Most of the data falls between $0 and $5,000, but there are several outliers above $20,000.



Group by State

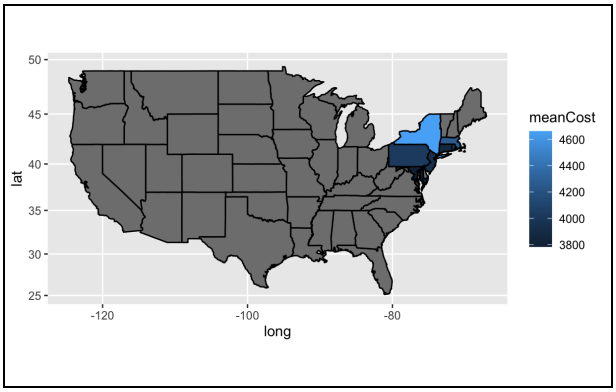
Following this, we grouped the data by state and created a plot showing the average healthcare cost for each state in our dataset.

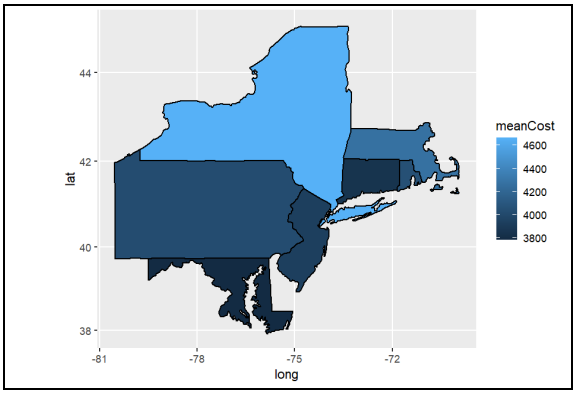
```
statemean <- hmo %>%
  group_by(location) %>%
  summarise(mean_cost = mean(cost))
```

```
bar <- ggplot(statemean, aes(x=location, y=mean_cost, fill=location)) +
  geom_bar(stat="identity")+theme_minimal() + geom_jitter(width=0.15)+
  theme(axis.text.x = element_text(angle = 45, hjust=1)) + theme(legend.position =
"none")
```

This shows that people living in New York have the highest healthcare costs, on average. New York is followed by Massachusetts. The remaining states all have negligible differences in the average healthcare costs. Another visualization we created shows the healthcare costs by states by using a map.

```
stateHmo <- hmo %>%
  group_by(location) %>%
  summarise(meanCost = mean(cost))

stateHmo <- data.frame(stateHmo)
stateHmo$state <- tolower(stateHmo$location)

us <- map_data("state")
mergeHmo <- merge(stateHmo,us,by.x="state",by.y="region",all.x=T)
mergeHmo <- mergeHmo %>% arrange(order)

mapHmo <- ggplot(mergeHmo) +
  geom_polygon(color="black",aes(x=long,y=lat,group=group,fill=meanCost)) +
  coord_map()
mapHmo
```

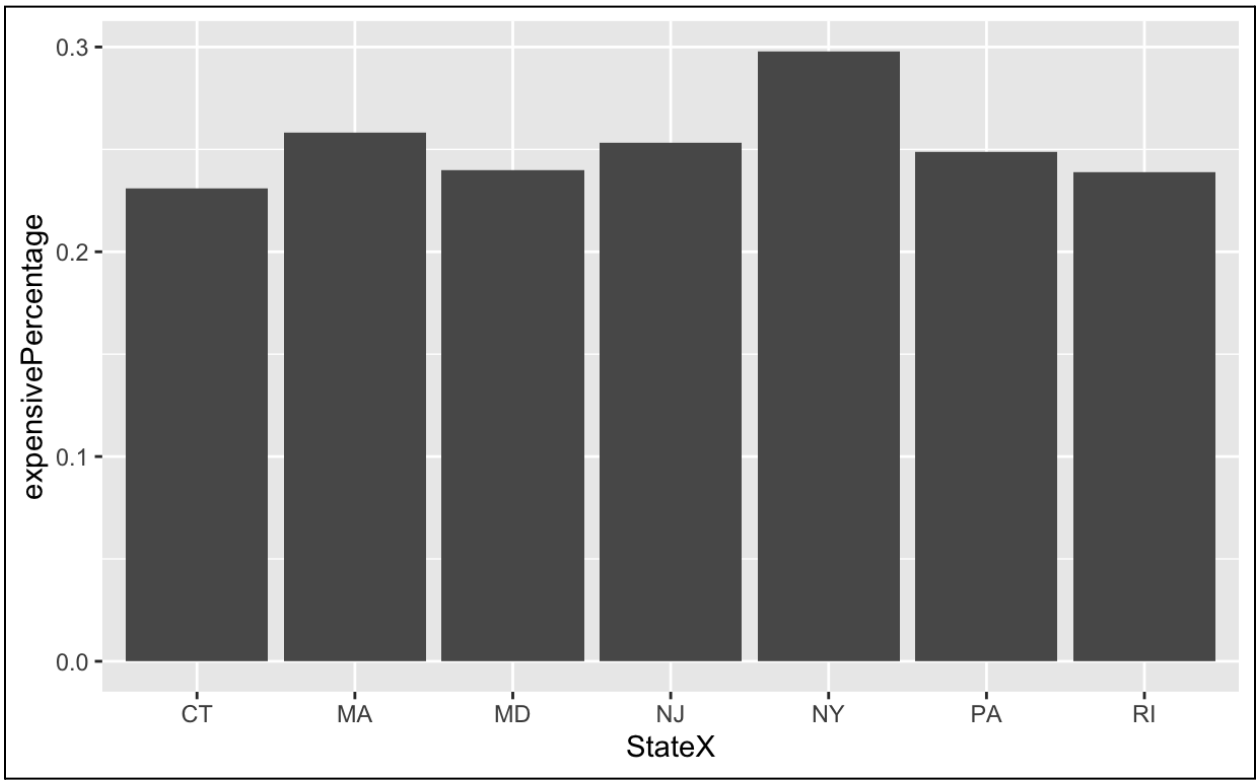




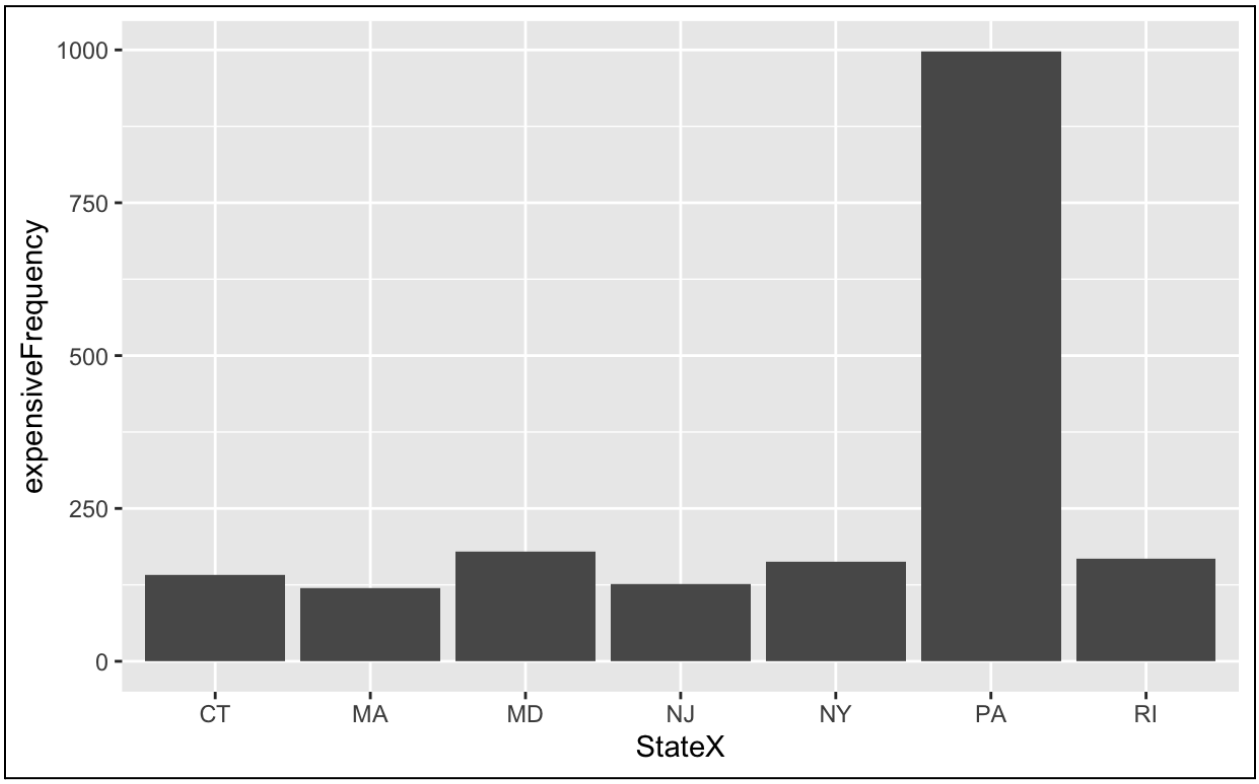
Besides, we also grouped the data by state and created 2 plots showing the percentage of being expensive and the frequency of being expensive on the healthcare cost for each state in our dataset.

```
hmo%>% group_by(location)%>%
  summarise(table(expensive))
```

Visualizing the expensive percentage by state:



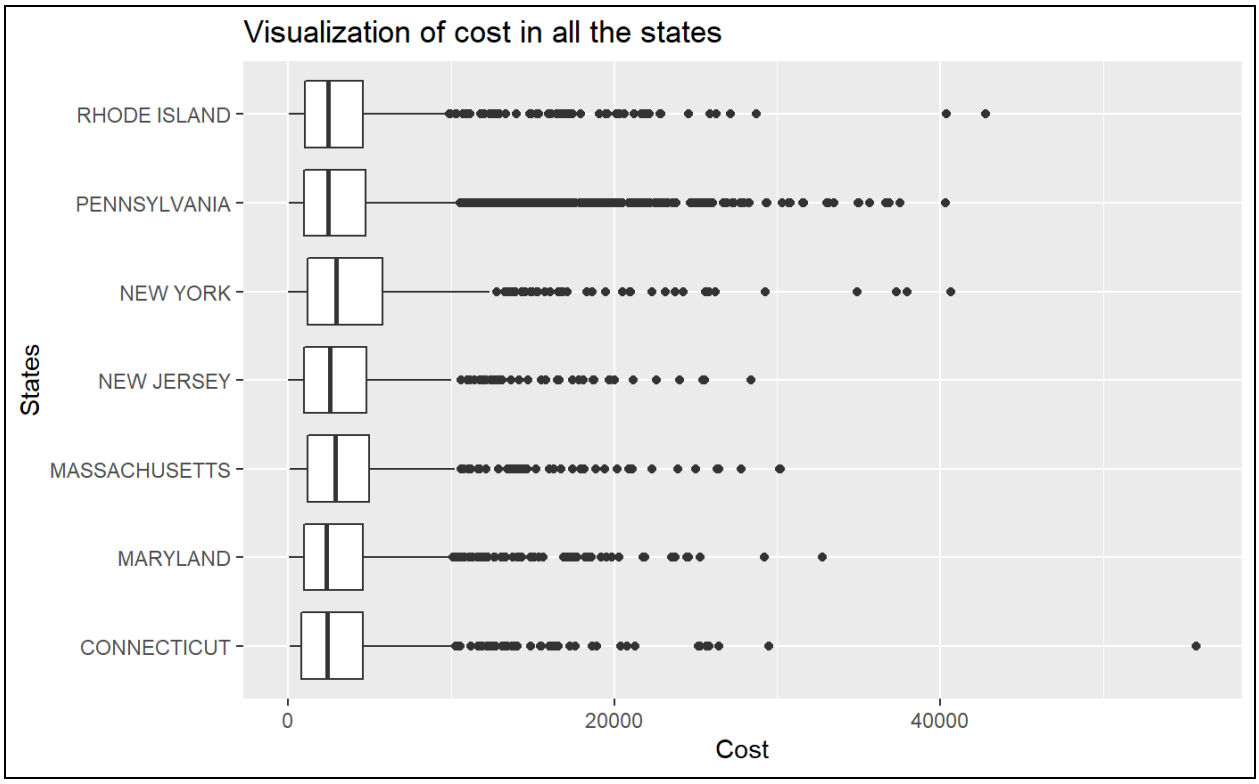
Visualizing the expensive frequency by state:



This shows that people living in New York have the highest percentage of expensive healthcare costs. New York is followed by Massachusetts. The remaining states all have negligible differences in the percentage of being expensive on healthcare costs.

Because the sample size of the data varies from state to state, the frequency of being expensive varies widely, but this has little impact on our analysis and is not important data, so we decided to skim over.

```
boxplot_state <- ggplot(hmo)+aes(x=cost,y=location)+geom_boxplot() +
xlab("Cost")+ylab("States")+ggtitle("Visualization of cost in all the states")
boxplot_state
```

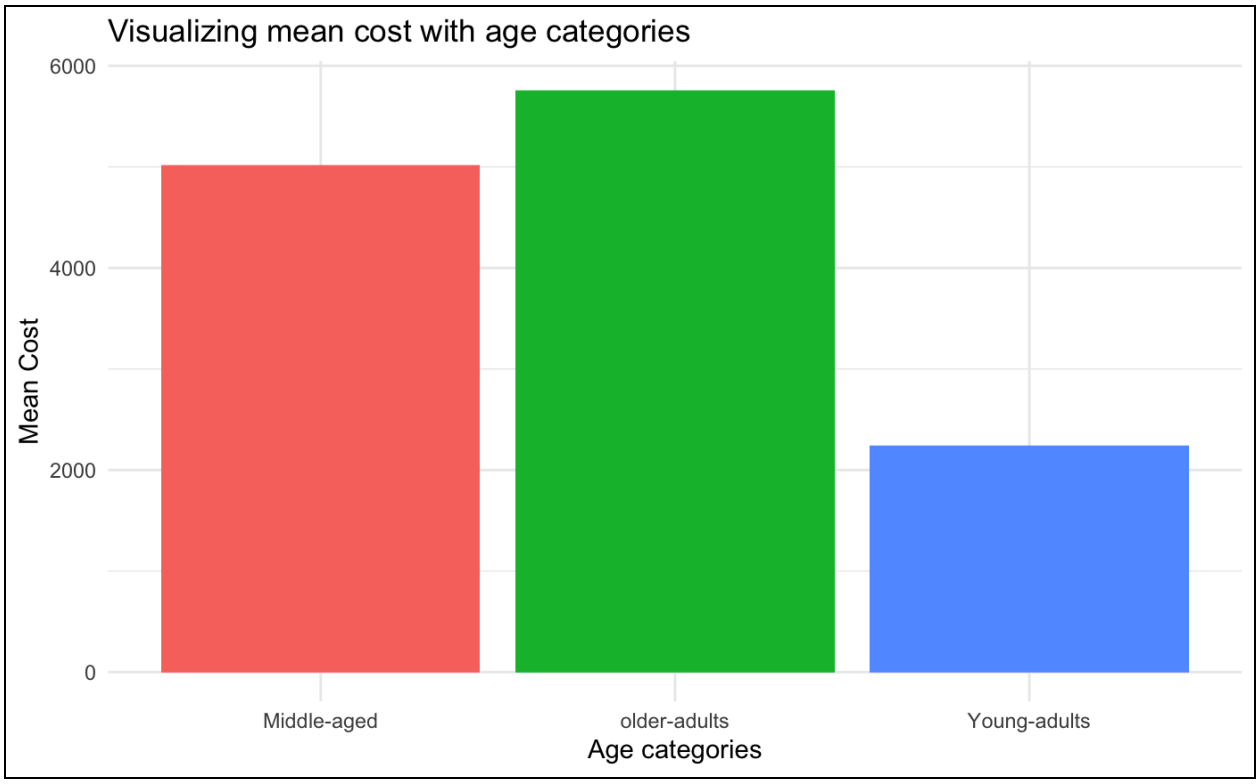


Group by Age Category

Next, we grouped the data by age category to determine what impact age has on cost.

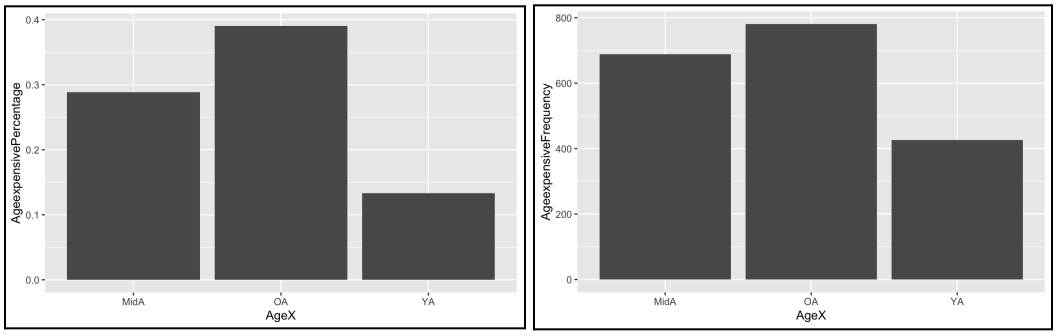
```
statemeanage <- hmo %>%
  group_by(agecategory) %>%
  summarise(mean_cost = mean(cost))
```

```
bar2 <- ggplot(statemeanage, aes(x=agecategory, y=mean_cost, fill=agecategory)) +
geom_bar(stat="identity")+theme_minimal() + theme(legend.position = "none")
```



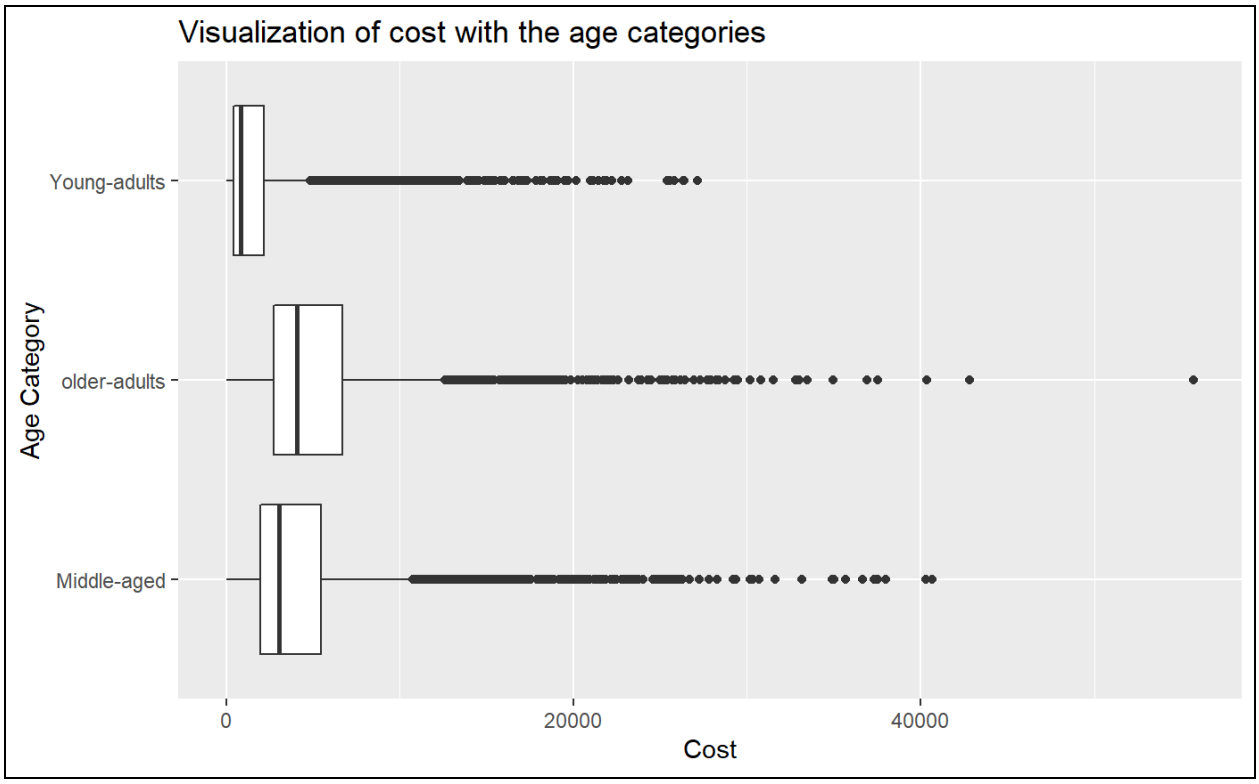

```
hmo%>% group_by(agecategory)%>%
  summarise(table(expensive))
```

```
YA <- hmo[hmo$agecategory=="Young-adults",]
expensivePercentageYA <- sum(YA$expensive=="1")/nrow(YA)

MidA <- hmo[hmo$agecategory=="Middle-aged",]
expensivePercentageMidA <- sum(MidA$expensive=="1")/nrow(MidA)

OA <- hmo[hmo$agecategory=="older-adults",]
expensivePercentageOA <- sum(OA$expensive=="1")/nrow(OA)

AgeX <- c("YA","MidA","OA")
AgeexpensivePercentage <- c(expensivePercentageYA,
                            expensivePercentageMidA,
                            expensivePercentageOA)

AgeexpensiveFrequency <- c(sum(YA$expensive=="1"),
                           sum(MidA$expensive=="1"),
                           sum(OA$expensive=="1"))

AgeexpensivePercentageDF <-
data.frame(AgeX,AgeexpensivePercentage,AgeexpensiveFrequency)

bar11 <- ggplot(AgeexpensivePercentageDF, aes(x=AgeX, y=AgeexpensivePercentage)) +
geom_bar(stat="identity")
bar11

bar12 <- ggplot(AgeexpensivePercentageDF, aes(x=AgeX, y=AgeexpensiveFrequency)) +
geom_bar(stat="identity")
bar12
```



It is clearly evident that the average cost, the percentage of being expensive, and the frequency of being expensive for older adults is significantly higher than those for middle aged adults, which is significantly higher than those for young adults.

```
boxplot_age<-ggplot(hmo)+aes(x=cost,y=agecategory)+geom_boxplot() +
xlab("Cost")+ylab("Age Category")+ggtitle("Visualization of cost with the age categories")
boxplot_age
```

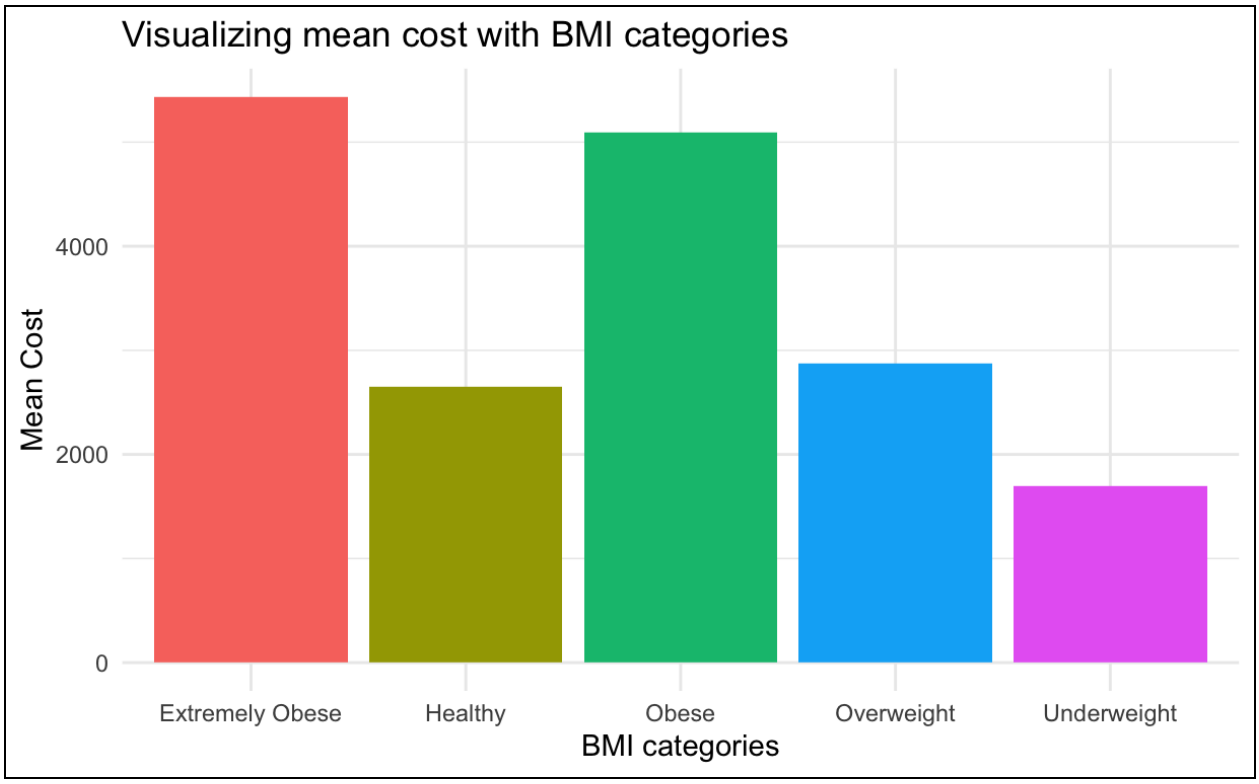


Group by BMI Category

We then grouped the data by bmi category to analyze the impact of bmi on cost.

```
meanbmi <- hmo %>%
  group_by(bmicategory) %>%
  summarise(mean_cost = mean(cost))
```

```
bar3 <- ggplot(meanbmi, aes(x=bmicategory, y=mean_cost, fill=bmicategory)) +
geom_bar(stat="identity")+theme_minimal() + theme(legend.position = "none")
```



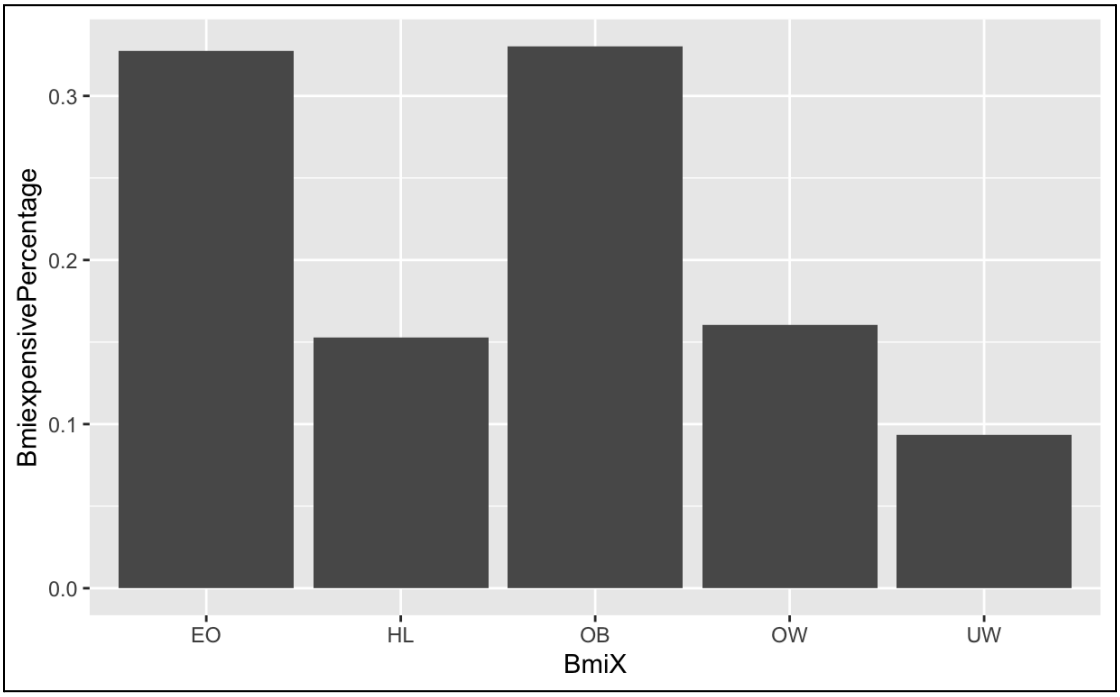
As you can see, the average cost for extremely obese and obese people is well over $4,000. On the other hand, the cost for overweight and healthy people is similar while the cost for underweight is the lowest.

When we looked at the percentage and frequency of being expensive grouped by the bmi category, we can see that the cost for extremely obese and obese people is more likely to cost more on health care.

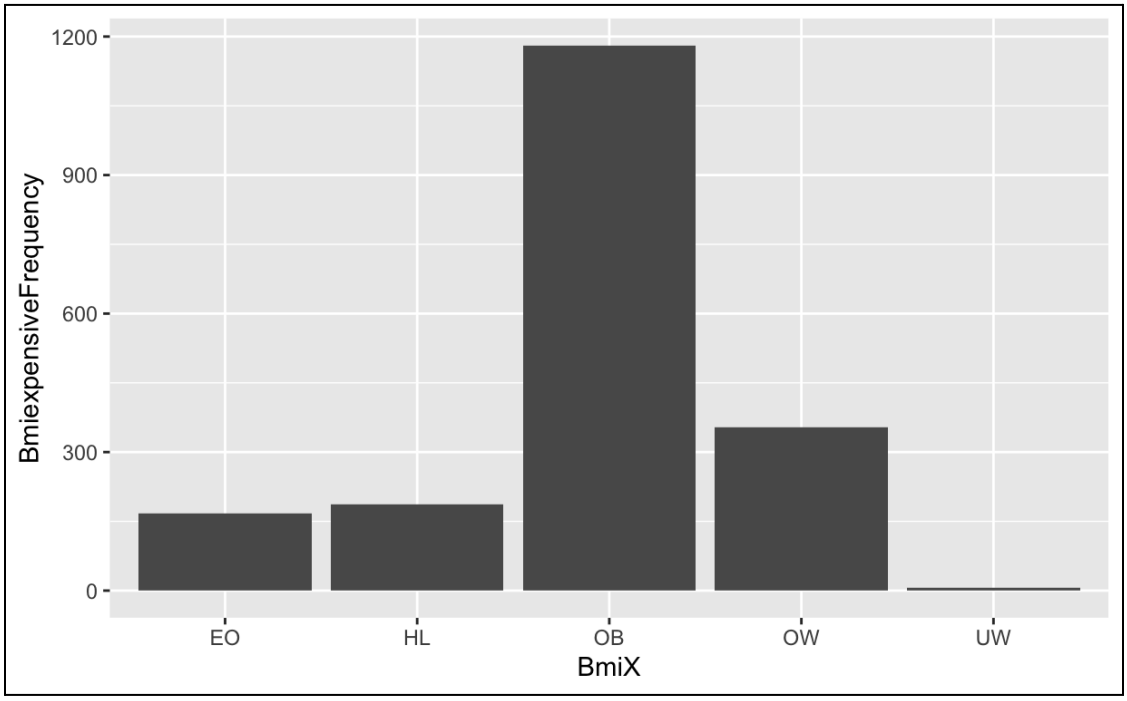
```
hmo%>% group_by(bmicategory)%>%
  summarise(table(expensive))
```

```r
EO <- hmo[hmo$bmicategory=="Extremely Obese",]
expensivePercentageEO <- sum(EO$expensive=="1")/nrow(EO)

HL <- hmo[hmo$bmicategory=="Healthy",]
expensivePercentageHL <- sum(HL$expensive=="1")/nrow(HL)

OB <- hmo[hmo$bmicategory=="Obese",]
expensivePercentageOB <- sum(OB$expensive=="1")/nrow(OB)

OW <- hmo[hmo$bmicategory=="Overweight",]
expensivePercentageOW <- sum(OW$expensive=="1")/nrow(OW)

UW <- hmo[hmo$bmicategory=="Underweight",]
expensivePercentageUW <- sum(UW$expensive=="1")/nrow(UW)

BmiX <- c("EO","HL","OB","OW","UW")
BmiexpensivePercentage <- c(expensivePercentageEO,
                            expensivePercentageHL,
                            expensivePercentageOB,
                            expensivePercentageOW,
                            expensivePercentageUW)

BmiexpensiveFrequency <- c(sum(EO$expensive=="1"),
                           sum(HL$expensive=="1"),
                           sum(OB$expensive=="1"),
                           sum(OW$expensive=="1"),
                           sum(UW$expensive=="1"))
BmiexpensivePercentageDF <-
data.frame(BmiX,BmiexpensivePercentage,BmiexpensiveFrequency)

bar9 <- ggplot(BmiexpensivePercentageDF, aes(x=BmiX, y=BmiexpensivePercentage)) +
geom_bar(stat="identity")
bar9

bar10 <- ggplot(BmiexpensivePercentageDF, aes(x=BmiX, y=BmiexpensiveFrequency)) +
geom_bar(stat="identity")
bar10
```

Expensive Percentage by BMI Category:



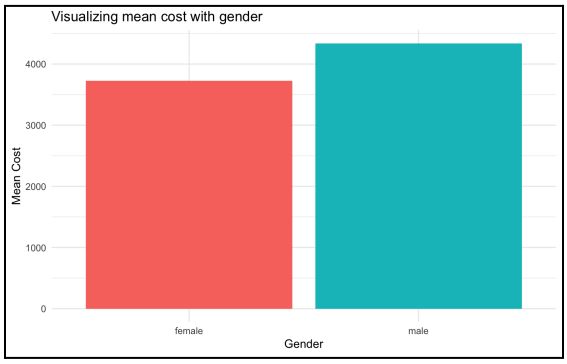
Expensive Frequency by BMI Category:

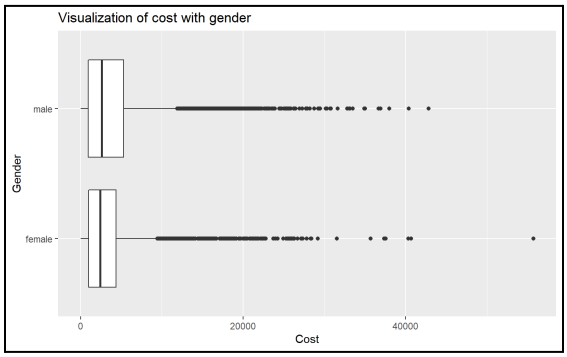


Other Visualizations

When we look at the bar graph for gender, we can see that men spend more on healthcare than women do, but the difference is not significant.

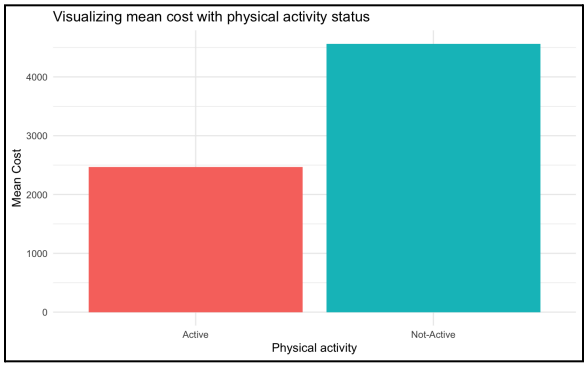
```
boxplot_gender <- ggplot(hmo) + aes(x=cost,y=gender)+geom_boxplot() +
xlab("Cost")+ylab("Gender")+ggtitle("Visualization of cost with gender")
boxplot_gender
```

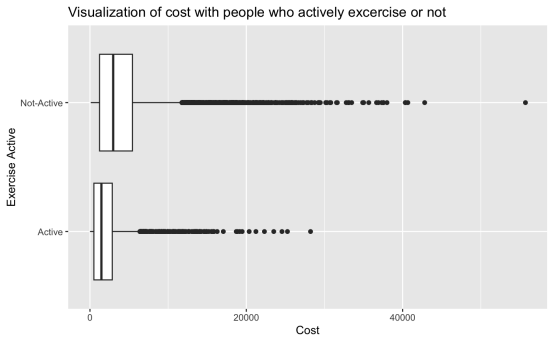




It is clear that the mean cost for someone that is not active doing physical exercise is much higher than the mean cost for someone that is active doing exercise. Therefore, active exercise people have a lower chance to be expensive than someone who is not active.

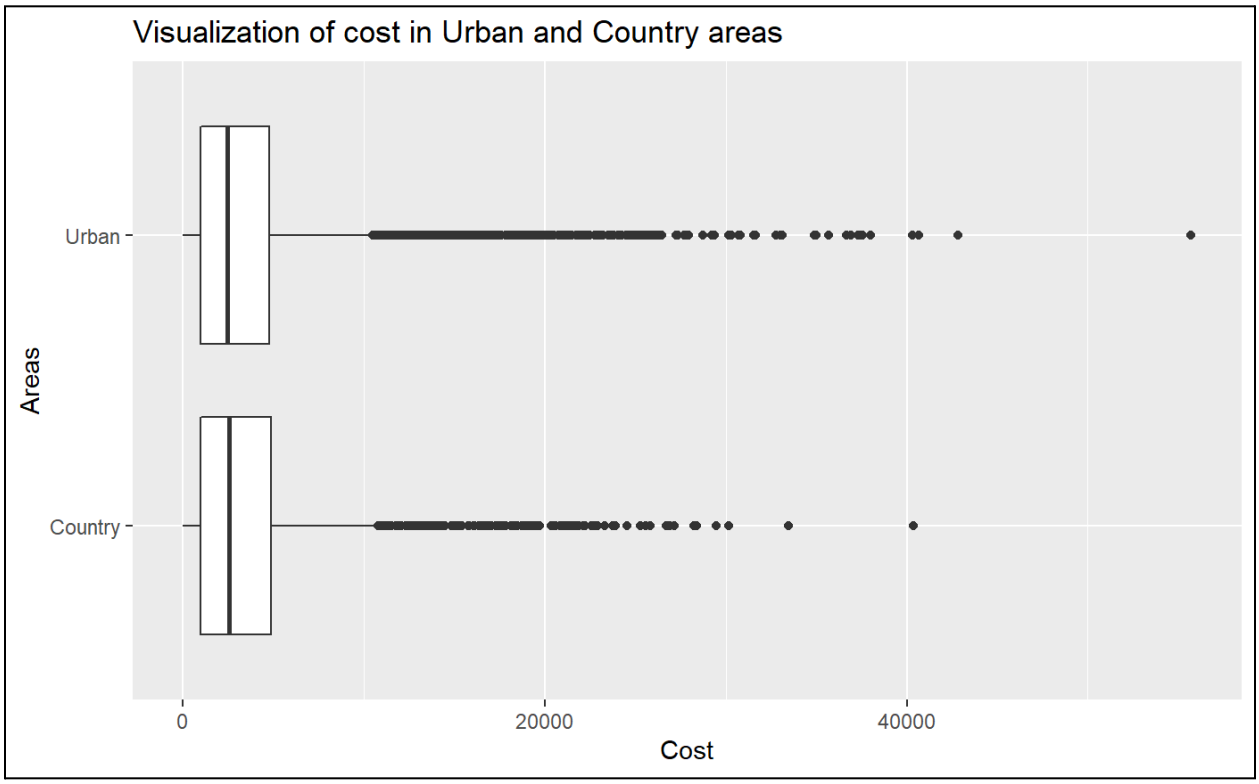
```
boxplot_exercise <- ggplot(hmo) + aes(x=cost,y=exercise)+geom_boxplot() + xlab("Cost")+ylab("Physically
Active")+ggtitle("Visualization of cost with people who actively excercise or not")
boxplot_exercise
```





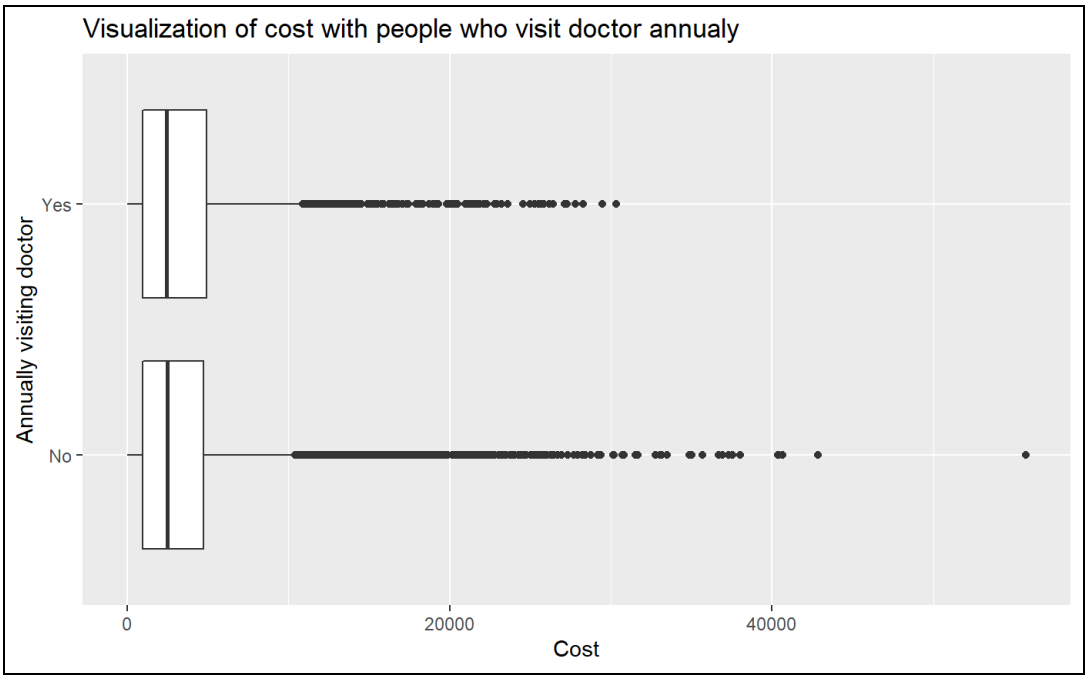
Additionally, we looked at urban vs country locations and the correlation to health care costs. We did not find a significant relationship. However, there are several high outliers in the urban sample compared to the country sample.

```
boxplot_location_type<- ggplot(hmo)+aes(x=cost,y=location_type)+geom_boxplot()+
xlab("Cost")+ylab("Areas")+ggtitle("Visualization of cost in Urban and Country areas")
boxplot_location_type
```

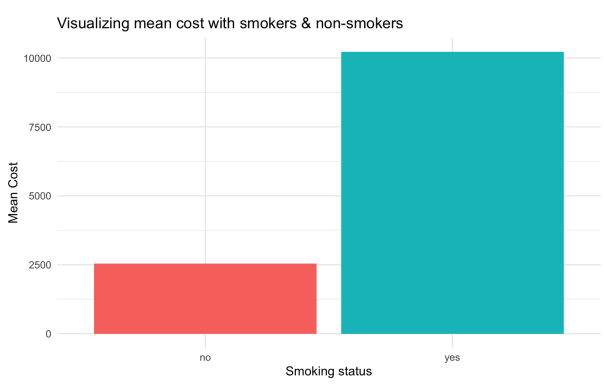


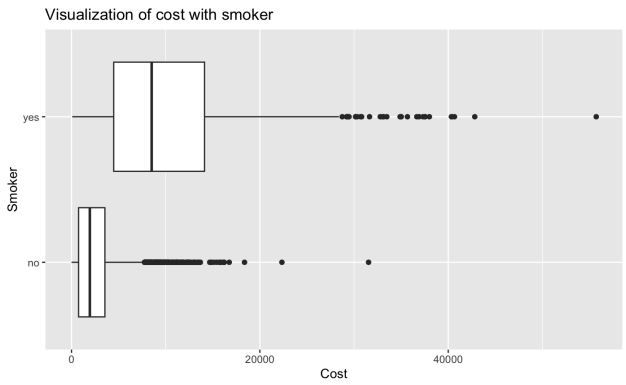
We explored the impact of annual physicals on the healthcare cost for an individual. This was clear in that it showed much higher outliers for people that do not have an annual physical and the average price for someone who does not have a physical is higher than someone who does.

```
boxplot_physical <- ggplot(hmo) + aes(x=cost,y=yearly_physical)+geom_boxplot() +
xlab("Cost")+ylab("Annually visiting doctor")+ggtitle("Visualization of cost with people who visit doctor
annualy")
boxplot_physical
```

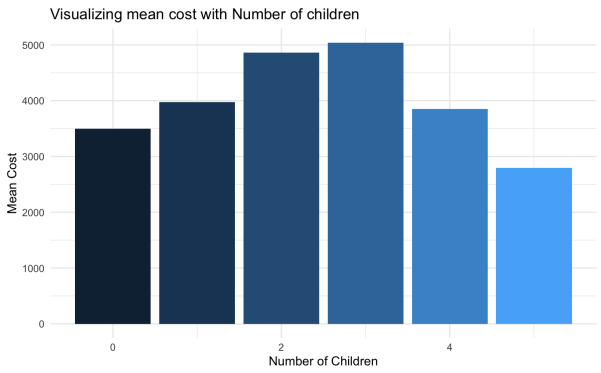


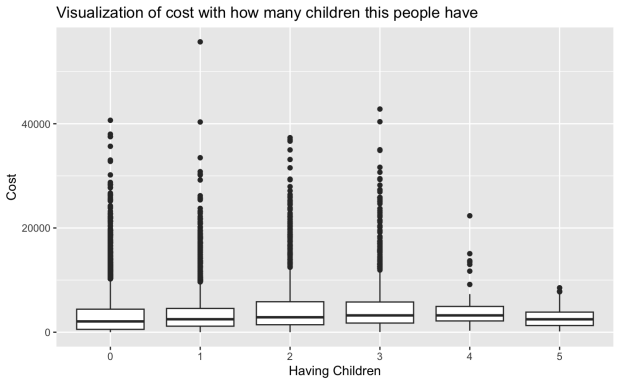
Then, we visualize the correlation between cost and smoker & non-smoker, it is clear that the mean cost for someone that is a smoker is much higher than the mean cost for someone that is not a smoker. Therefore, non-smoker people have a lower chance to be expensive than someone who is a smoker.





Lastly, as the number of children increases, the cost of health care increases, but when there are four, the cost of healthcare decreases, and when there are five children, the cost of healthcare is lower than if there were no children.





Linear Model

After the initial data visualization, we began creating a model to solve the problem that we faced. First, we created a multiple linear regression model using all variables in the dataframe to model cost. We found that age, bmi, children, smoker, and being from New York were positive and statistically significant at the 0.05 level. This means that an increase in age, bmi, or children leads to an increase in insurance costs. Being a smoker or being from New York also leads to higher costs. We then ran a model with only statistically significant variables and the results are shown below.

```
hmo_1 <- select(hmo, -expensive)
modelAll_1 <- lm(cost ~., data = hmo_1)
summary(modelAll_1)

#model with significant predictors
modelSignificant <- lm(cost ~ age + bmi + children + smoker + location + education_level + exercise +
married + hypertension, data = hmo)
summary(modelSignificant)
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -9101.452    259.363 -35.092  < 2e-16 ***
age                            102.399      2.629  38.952  < 2e-16 ***
bmi                            181.439      6.221  29.164  < 2e-16 ***
children                       233.250     30.444   7.662 2.06e-14 ***
smokeryes                     7665.446     93.437  82.039  < 2e-16 ***
locationMARYLAND              -129.468    175.706  -0.737 0.461238
locationMASSACHUSETTS            8.140    198.350   0.041 0.967268
locationNEW JERSEY             108.036    194.511   0.555 0.578621
locationNEW YORK               470.417    189.706   2.480 0.013170 *
locationPENNSYLVANIA            14.409    139.922   0.103 0.917982
locationRHODE ISLAND           118.118    178.143   0.663 0.507317
education_levelMaster          -97.611     95.102  -1.026 0.304741
education_levelNo College Degree  42.000    126.283   0.333 0.739453
education_levelPhD            -234.554    129.865  -1.806 0.070937 .
exerciseNot-Active            2261.702     85.619  26.416  < 2e-16 ***
marriedNot_Married             132.345     78.548   1.685 0.092051 .
hypertension                   341.360     92.739   3.681 0.000234 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3220 on 7565 degrees of freedom
Multiple R-squared:  0.574,     Adjusted R-squared:  0.5731
F-statistic: 637.1 on 16 and 7565 DF,  p-value: < 2.2e-16
```

The p-value is less than 0.05, we accepted this model. The Adjusted R-squared value for this was 0.5731. This means that 57.31% of the variation in health cost can be explained by the variables included in the model.

SVM Model

We then decided to create an SVM model, which is a supervised modeling technique. To do this, we used the createDataPartition() function from the caret package. We first tried splitting the data into test and training datasets with .2 and .8 partitions, but that model is not as accurate as the model we created when we split the data into test and training datasets with ⅔ of the data being used as the training data.

```r
library(caret)
library(kernlab)
hmo$expensive <- as.factor(hmo$expensive)

HMO <- select(hmo,-cost,-X)

set.seed(687)
trainList <- createDataPartition(y=HMO$expensive,p=.67,list=FALSE)
training <- HMO[trainList,]
testing <- HMO[-trainList,]
```

Next, we created an SVM model using all the variables with expensive being the variable of interest. This was created using the train() function. A second SVM model was made using the ksvm() function. Then, we predicted the test data with both models and created a Confusion matrix for each.

```
svm.model1 <- train(expensive ~ ., data = training,
method = "svmRadial",
trControl=trainControl(method = "none"),
preProcess = c("center", "scale"))
svm.model1

# another way to train the model
svm.model2 <- ksvm(expensive ~., data=training,
C=5,cross=3, prob.model = TRUE)
svm.model2

svmPred1 <- predict(svm.model1, newdata=testing)
confusionMatrix(svmPred1, testing$expensive)

svmPred2 <- predict(svm.model2, newdata=testing)
confusionMatrix(svmPred2, testing$expensive)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1825  267
         1   51  358

               Accuracy : 0.8729
                 95% CI : (0.8592, 0.8857)
    No Information Rate : 0.7501
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6167

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

            Sensitivity : 0.9728
            Specificity : 0.5728
         Pos Pred Value : 0.8724
         Neg Pred Value : 0.8753
             Prevalence : 0.7501
         Detection Rate : 0.7297
   Detection Prevalence : 0.8365
      Balanced Accuracy : 0.7728

       'Positive' Class : 0
```

(Model 1)

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1806  220
         1   70  405

               Accuracy : 0.884
                 95% CI : (0.8708, 0.8963)
    No Information Rate : 0.7501
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6638

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

            Sensitivity : 0.9627
            Specificity : 0.6480
         Pos Pred Value : 0.8914
         Neg Pred Value : 0.8526
             Prevalence : 0.7501
         Detection Rate : 0.7221
   Detection Prevalence : 0.8101
      Balanced Accuracy : 0.8053

       'Positive' Class : 0
```

(Model 2)

The P-Value for both models is less than 0.05, the no information rate is 0.7501. The accuracy of the first model was 0.8729 while the second model was 0.884. However, the

sensitivity of the first model was 0.9728, higher than the sensitivity of the second model 0.9627. We then decided to use the first model to predict the test sample.

```
testData <- read_csv("HMO_TEST_data_sample.csv")

testData$agecategory[18<=testData$age & testData$age<=34] <- "Young-adults"
testData$agecategory[35<=testData$age & testData$age<=50] <- "Middle-aged"
testData$agecategory[51<=testData$age & testData$age<=66] <- "older-adults"

testData$bmicategory[testData$bmi <= 18 ] <- "Underweight"
testData$bmicategory[testData$bmi >= 18 & testData$bmi < 25  ] <- "Healthy"
testData$bmicategory[testData$bmi >= 25 & testData$bmi < 30  ] <- "Overweight"
testData$bmicategory[testData$bmi >= 30 & testData$bmi < 40  ] <- "Obese"
testData$bmicategory[testData$bmi >= 40 & testData$bmi < 65  ] <- "Extremely Obese"

testData <- select(testData,-X)

testing2 <- select(testing,-expensive)
total <- rbind(testing2,testData)

svmPred3 <- predict(svm.model1, total)

svmPred3[2502:2521]
testData$expensive_svm <- svmPred3[2502:2521]
```

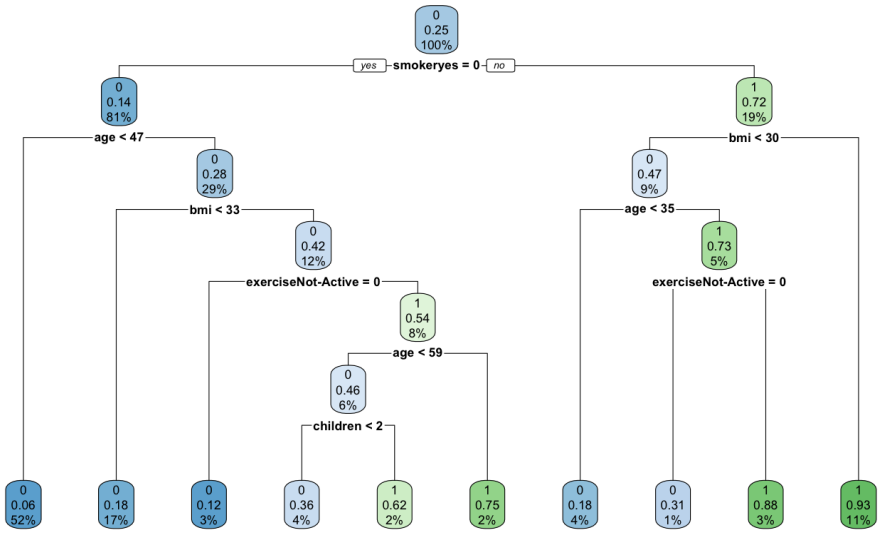
```
[1] 0 0 1 1 1 1 1 1 0 0 1 0 0 0 0 0 0 0 0 0
```

Tree Model

We then generated a tree model.

```r
library(e1071)
# train the model with rpart method
trctrl <- trainControl(method="repeatedcv", number=10)
model.rpart <- train(expensive ~ ., method = "rpart",
data = training,
trControl=trctrl,tuneLength = 50)
# getting the result of model.rpart
model.rpart

library(rpart.plot)
## Loading required package: rpart
library(rpart)
# getting the plot of model.rpart$finalModel
rpart.plot(model.rpart$finalModel)

predictValues <- predict(model.rpart,newdata=testing)
# getting the confusion matrix
# looking at the accuracy, no information rate and the p-value
confusionMatrix(predictValues, testing$expensive)

predictValues2 <- predict(model.rpart,newdata=testData)
# getting the confusion matrix
# looking at the accuracy, no information rate and the p-value
predictValues2
testData$expensive_tree <- predictValues2
```

```
          Reference
Prediction    0    1
         0 1817  211
         1   59  414

               Accuracy : 0.892
                 95% CI : (0.8792, 0.9039)
    No Information Rate : 0.7501
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6866

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

            Sensitivity : 0.9686
            Specificity : 0.6624
         Pos Pred Value : 0.8960
         Neg Pred Value : 0.8753
             Prevalence : 0.7501
         Detection Rate : 0.7265
   Detection Prevalence : 0.8109
      Balanced Accuracy : 0.8155

       'Positive' Class : 0
```



Compared with the SVM models, the Accuracy, the Pos Pred Value, and the Neg Pred Value of the Tree model were a little higher. However, the Sensitivity was lower than the first SVM Model.

We used the tree model to predict the test sample.

```
 [1] 0 0 1 1 1 1 1 1 0 0 1 0 0 0 0 0 0 0 0 0
Levels: 0 1
```

For a test sample of 20 people, the SVM model and the tree model were used to predict whether these people were expensive. Though the first SVM model is our best model (because it has the highest sensitivity), we got the same results from the 2 models.

expensive_svm	expensive_tree
0	0
0	0
1	1
1	1
1	1
1	1
1	1
1	1
0	0
0	0
1	1
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0

Association Rules

Next, we used association rules mining to predict whether someone would be expensive or not.

```
library(arules)
library(arulesViz)
hmo_new <- data.frame(X=as.factor(hmo$X),
                      agecategory=as.factor(hmo$agecategory),
                      bmicategory=as.factor(hmo$bmicategory),
                      children=as.factor(hmo$children),
                      smoker=as.factor(hmo$smoker),
                      location=as.factor(hmo$location),
                      location_type=as.factor(hmo$location_type),
                      education_level=as.factor(hmo$education_level),
                      yearly_physical=as.factor(hmo$yearly_physical),
                      exercise=as.factor(hmo$exercise),
                      married=as.factor(hmo$married),
                      hypertension=as.factor(hmo$hypertension),
                      gender=as.factor(hmo$gender),
                      expensive=as.factor(hmo$expensive)
)
hmoX <- as(hmo_new, "transactions")
#itemFrequency(hmoX)
itemFrequencyPlot(hmoX,topN=20)
inspect(hmoX[1:10])



ruleset <- apriori(hmoX,
                   parameter = list(support = 0.05,confidence = 0.83),
                   control=list(verbose=F),
 appearance=list(default="lhs",rhs=("expensive=1")))
summary(ruleset)
inspectDT(ruleset)
```

We generated 16 rulesets, and these were the sets of rules with the highest lift:

	LHS	RHS	support	confidence	coverage	lift	count
	All	All	All	All	All	All	All
[11]	{bmicategory=Obese,smoker=yes,yearly_physical=No,exercise=Not-Active}	{expensive=1}	0.051	0.997	0.051	3.991	383.000
[7]	{bmicategory=Obese,smoker=yes,exercise=Not-Active}	{expensive=1}	0.066	0.996	0.066	3.985	500.000
[12]	{bmicategory=Obese,smoker=yes,exercise=Not-Active,hypertension=0}	{expensive=1}	0.052	0.995	0.052	3.981	394.000
[5]	{bmicategory=Obese,smoker=yes,married=Married}	{expensive=1}	0.057	0.946	0.061	3.783	434.000
[2]	{bmicategory=Obese,smoker=yes}	{expensive=1}	0.084	0.940	0.090	3.760	639.000
[4]	{bmicategory=Obese,smoker=yes,education_level=Bachelor}	{expensive=1}	0.053	0.937	0.056	3.748	400.000
[3]	{bmicategory=Obese,smoker=yes,gender=male}	{expensive=1}	0.055	0.937	0.058	3.748	414.000
[8]	{bmicategory=Obese,smoker=yes,yearly_physical=No}	{expensive=1}	0.064	0.934	0.068	3.739	485.000
[9]	{bmicategory=Obese,smoker=yes,hypertension=0}	{expensive=1}	0.065	0.934	0.070	3.736	494.000

If a person is a smoker, with the BMI >= 30 & < 40, and at the same time neither have yearly physical nor active exercise, this person has a higher probability of being expensive in health care cost.

Conclusion

Based on the data exploration and analysis that we performed, we were able to identify patterns and trends that correlate with whether a person is expensive or not expensive in the healthcare industry. For instance, people that are smoker, are obese, extremely obese, not active in doing exercise, and not going to a yearly physical are most likely to be expensive. All of these factors increase the probability that a person is expensive. In addition to this, geographical factors come into play. Living in New York increases the chances that a person is expensive compared to people living in the surrounding states in the northeast. Additionally, as people get older they are more likely to be expensive and an increase in children (but less than 5) leads to an increase in price generally.

Based on the analysis of the data, we recommend that the health care company offer discounts for non-smokers and people with a BMI lower than 30, and offer special plans for people in New York State.