IST 687 – Introduction to Data Science

Group 3

DATA ANALYSIS FOR HEALTH MANAGEMENT ORGANIZATION

Project Report By:

Gagan Gupta
Shreyas Tekawade
Chinmay Maganur
Piyush Khedkar

Table of Contents

1. Description

The project revolves around analyzing the medical data collected from a huge number of patients and use those data to provide recommendations to HMO (Health Management Organization). The recommendations answer questions to predict which patients will spend more on healthcare the following year and provide suggestion on how HMO can reduce their total health care costs.

2. Project Scope and Objective

The scope of this project is to analyze and draw insights from the dataset provided to us which contains data regarding different data points of a person such as are they a smoker, do they exercise regularly, do they live in an urban or country location. A total of 13 different attributes were found to be present in the dataset for each person.
We will be concentrating on finding the factors that cause a person to be an “expensive” or “not-expensive” i.e., whether they would be spending more on healthcare or not and possibly what factors are going to be affecting that outcome and get actionable insights by applying statistical techniques.
The objective of this project is to suggest our client, Health Management Organization, the areas where they can improve to decrease their total health care cost. Also, provide them with insights on the category of persons that will be having more health-related spending next year.

3. Project Deliverables

- Ensure that our dataset has no invalid or missing fields by doing data cleaning before continuing analysis.
- Applying linear regression, we can identify the factors that have the greatest impact on the individual's healthcare expenditures, and we can then conduct more analysis into those factors.
- Applying support vector machine to forecast next year's spending and generating actionable insights for HMO.
- Finally, provide suggestions to HMO based on the data analysis and interpretation to enhance and improve their understanding on what factors affect healthcare spending of any individual.

4. Data Acquisition

The data set was made available to us by the course instructors. Before any data munging, this data set consisted of approximately 7583 survey responses of the people and consisted of 14 fields such as age, BMI, number of children, gender etc.
This data was extensively studied to determine the usable variables. After this initial analysis, the data set was forwarded to the preprocessing phase where all the errors in the data were removed to make it usable for further analysis.

5. Data Preprocessing

Before preprocessing, the dataset contained 7582 rows and 14 column variables. We then summarized all the null values present in our dataset as that would cause errors in our models.

```
# Checking for null values
colSums(sapply(df,is.na))

# We have 78 null values in BMI and 80 in hypertension.
```

```
              X             age             bmi        children          smoker        location 
              0               0              78               0               0               0 
  location_type education_level yearly_physical        exercise         married    hypertension 
              0               0               0               0               0              80 
         gender            cost 
              0               0 
```

Fig. Summary of NA values present in the dataset

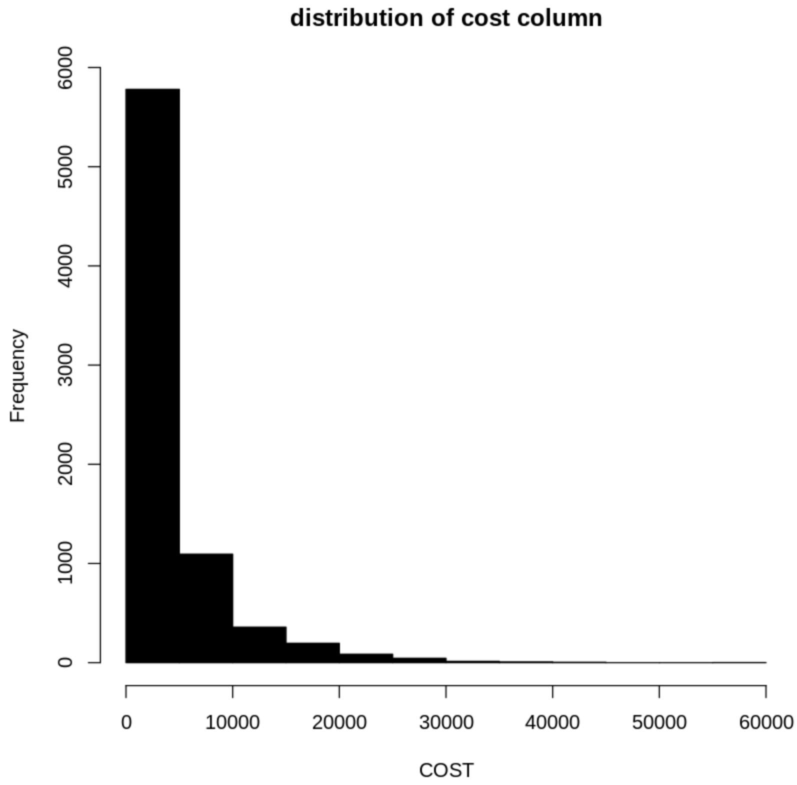
5.1 Expensive or Not Expensive

We then added new column for determining whether the person is "expensive" or "not-expensive". For this we used a threshold as $4775.

```
# Add new column Expensive or not where cost is more than 4775
df <-df %>% mutate(target=if_else(cost>4775,'expensive','not_expensive'))
dim(df)
```

```
[1] 7582   15
```

Fig. Mutating dataframe to add column expensive




```
      cost
 Min.   :    2
 1st Qu.:  970
 Median : 2500
 Mean   : 4043
 3rd Qu.: 4775
 Max.   :55715
```

Fig. Summary of cost in the dataframe.

Fig. Histogram of cost showing highly skewed data

We chose $4775 and not the mean $4043 as our data is skewed. Taking 75% or the 3rd quartile is beneficial as more than 97% of our data lies inside the 3rd standard deviation which gives us a more accurate baseline if we want to categorize people into expensive or not expensive.

5.2 Handling Null Values

```
# Hypertension and BMI have null values. We can use impute to try adding values
# but that returns values such as 0.5 in BMI which is not possible as that column
# denotes that the person either has hypertension or not.
# We are therefore removing those rows from the dataset.

df <- na.omit(df)
colSums(sapply(df,is.na))

# This results in us removing a total of 158 rows.
```

```
             X            age            bmi       children         smoker       location 
             0              0              0              0              0              0 
 location_type education_level yearly_physical     exercise        married   hypertension 
             0              0              0              0              0              0 
        gender           cost         target 
             0              0              0 
```

We had 78 NA values in BMI and 78 in hypertension. To deal with this we first used imputation to plug in missing values. This created a problem as mean for BMI is around 30 and using imputation would put all the NA valued people to have a BMI of 30+ which is Obese, and this would be wrong.

A greater issue was with hypertension attribute as that is only either 0 or 1 indicating either the person has hypertension or not. On using imputation methods this would put values such as 0.5.

The best approach was to just remove the rows in question altogether. This resulted in us removing a total of 158 rows.

After cleaning and creating expensive or not column we ended up with 7424 rows and 15 columns.
We also converted all char columns to factor and for running our machine learning models, we then converted factor to numeric.

```
# Convert all char columns to factor
df <- df %>% mutate_if(is.character, as.factor)
# LM does not accept factor so converted it to numeric
df <- df %>% mutate_if(is.factor, as.numeric)
str(df)
```

```
'data.frame':   7424 obs. of  15 variables:
 $ X              : int  1 2 3 4 5 7 9 10 11 12 ...
 $ age            : int  18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
 $ children       : int  0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : num  2 1 1 1 1 1 1 1 1 2 ...
 $ location       : num  1 7 3 6 6 6 6 6 6 1 ...
 $ location_type  : num  2 2 2 1 1 2 2 1 2 2 ...
 $ education_level: num  1 1 2 2 4 1 1 1 1 3 ...
 $ yearly_physical: num  1 1 1 1 1 1 1 1 1 1 ...
 $ exercise       : num  1 2 1 2 2 2 1 2 1 1 ...
 $ married        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ hypertension   : int  0 0 0 1 0 0 0 1 0 0 ...
 $ gender         : num  1 2 2 2 2 1 2 1 2 1 ...
 $ cost           : int  1746 602 576 5562 836 3842 1304 9724 201 4492 ...
 $ target         : num  2 2 2 1 2 2 2 1 2 2 ...
 - attr(*, "na.action")= 'omit' Named int [1:158] 20 32 93 118 167 231 281 309 320 434 ...
  ..- attr(*, "names")= chr [1:158] "20" "32" "93" "118" ...
```

Fig. Final cleaned dataset ready for analysis

6. Modelling Techniques

The information obtained from the dataset has been accurately modeled using a few different approaches. These models provide a comprehensible representation of the underlying data sets' reality. Specifically, the following models have been used:

6.1 Linear Regression

To begin with, we applied linear modelling to our dataset. By applying Simple linear regression, we could summarize and study relationships between variables in our dataset. The core idea was to obtain a line that best fits the data. The best fit line is the one for which total prediction error are as small as possible.

```
lm_out <- lm(data = df, target ~
age+bmi+children+smoker+location_type+education_level+yearly_physical+exercise+
                                        married+hypertension+gender)
summary(lm_out)
```

```
Call:
lm(formula = target ~ age + bmi + children + smoker + location_type +
    education_level + yearly_physical + exercise + married +
    hypertension + gender, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14911 -0.12754  0.05857  0.20546  0.94669

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.4952874  0.0387171  90.278  < 2e-16 ***
age             -0.0073727  0.0002707 -27.235  < 2e-16 ***
bmi             -0.0126181  0.0006392 -19.740  < 2e-16 ***
children        -0.0115526  0.0031422  -3.677 0.000238 ***
smoker          -0.5952234  0.0096473 -61.699  < 2e-16 ***
location_type    0.0102758  0.0088033   1.167 0.243138
education_level  0.0001895  0.0038518   0.049 0.960759
yearly_physical -0.0226936  0.0088251  -2.571 0.010145 *
exercise        -0.1700949  0.0088091 -19.309  < 2e-16 ***
married         -0.0081198  0.0080861  -1.004 0.315333
hypertension    -0.0330934  0.0095120  -3.479 0.000506 ***
gender          -0.0134879  0.0076709  -1.758 0.078735 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3281 on 7412 degrees of freedom
Multiple R-squared:  0.4272,    Adjusted R-squared:  0.4264
F-statistic: 502.6 on 11 and 7412 DF,  p-value: < 2.2e-16
```

Fig. Linear Regression

In our regression model, we found out several significant variables namely age, bmi, children, smoker, exercise, hypertension. This analysis included the following, checking for a high r squared value that is coefficient of determination, a small p value and by analyzing the residual plots.
Yearly_physical was somewhat important but did not have a major impact. While location_type, education_level, married, gender have little to no impact.

```
# Choosing only relevant columns
df <- data.frame(age = df$age, bmi = df$bmi, children=df$children, smoker=df$smoker, exercise = df$exercise,
hypertension=df$hypertension, yearly_physical = df$yearly_physical, target = as.factor(df$target))

trainList <- createDataPartition(y=df$target,p=.75,list=F)
# Putting 75% data in training and 25% in testing
training <- df[trainList,]
testing <- df[-trainList,]

str(df)
```

```
'data.frame':	7424 obs. of  8 variables:
 $ age            : int  18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
 $ children       : int  0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : num  2 1 1 1 1 1 1 1 1 2 ...
 $ exercise       : num  1 2 1 2 2 2 1 2 1 1 ...
 $ hypertension   : int  0 0 0 1 0 0 0 1 0 0 ...
 $ yearly_physical: num  1 1 1 1 1 1 1 1 1 1 ...
 $ target         : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 1 2 2 ...
```

Fig Splitting into training and testing dataset

We put 75% data in training and other 25% for testing.
In the target attribute, 1 meant that the person was expensive and 2 meant that the person was not expensive. Expensive means that the person would be spending more than $4775 in the subsequent year on medical expenses.

6.2 Support Vector Machine

We use SVM modeling techniques to predict the customer satisfaction by using various significant variables from our model.
After dividing the dataset into training and testing, we can check and validate our results.

```
```{r}
csvm <- ksvm(target~age+bmi+children+smoker+exercise+hypertension+yearly_physical, data=training, type =
"C-svc", C=4, rob.model = T, cross = 3)
csvm
```
```

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 4

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.109781576552456

Number of Support Vectors : 1604

Objective Function Value : -5759.548
Training error : 0.11582
Cross validation error : 0.125516
```

````
```{r}
svmPred <-predict(csvm,testing)
confusionMatrix(svmPred,testing$target)
```
````

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1  261   38
         2  203 1353

               Accuracy : 0.8701
                 95% CI : (0.8539, 0.8851)
    No Information Rate : 0.7499
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6071

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

            Sensitivity : 0.5625
            Specificity : 0.9727
         Pos Pred Value : 0.8729
         Neg Pred Value : 0.8695
             Prevalence : 0.2501
         Detection Rate : 0.1407
   Detection Prevalence : 0.1612
      Balanced Accuracy : 0.7676

       'Positive' Class : 1
```

Fig. Support Vector Machine Output

````
library(arules)
library(arulesViz)
df_x <- as(df,'transactions')

itemFrequency(df_x,type='absolute')
itemFrequencyPlot(df_x,type='absolute')
```
````

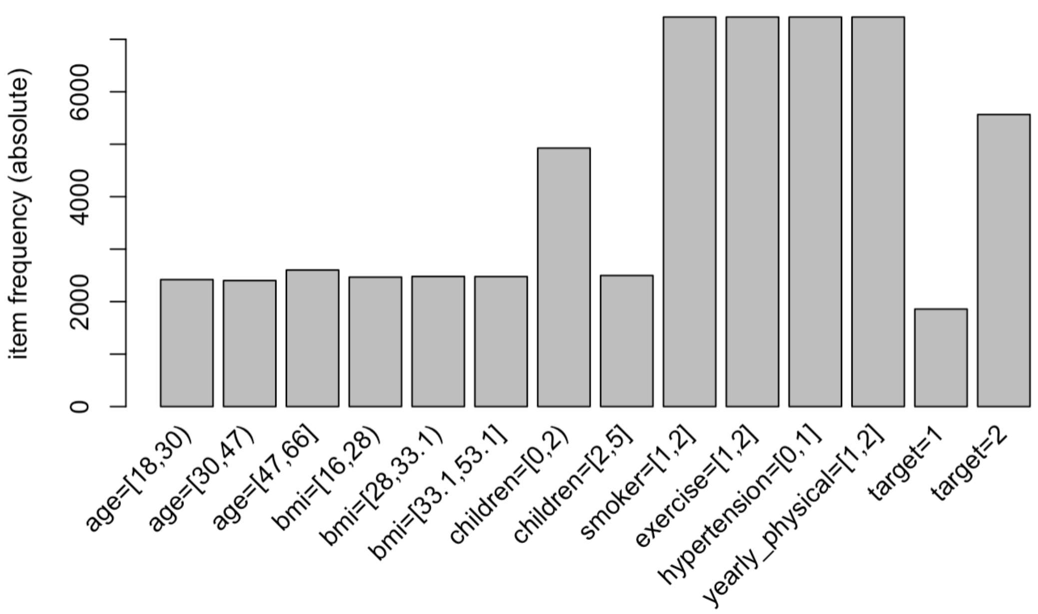


Fig. Item Frequency
````

Supporting Histogram

```
hist(alpha(csvm)[[1]], main="Support Vector Histogram with C=5", xlab="Support Vector Values")
```

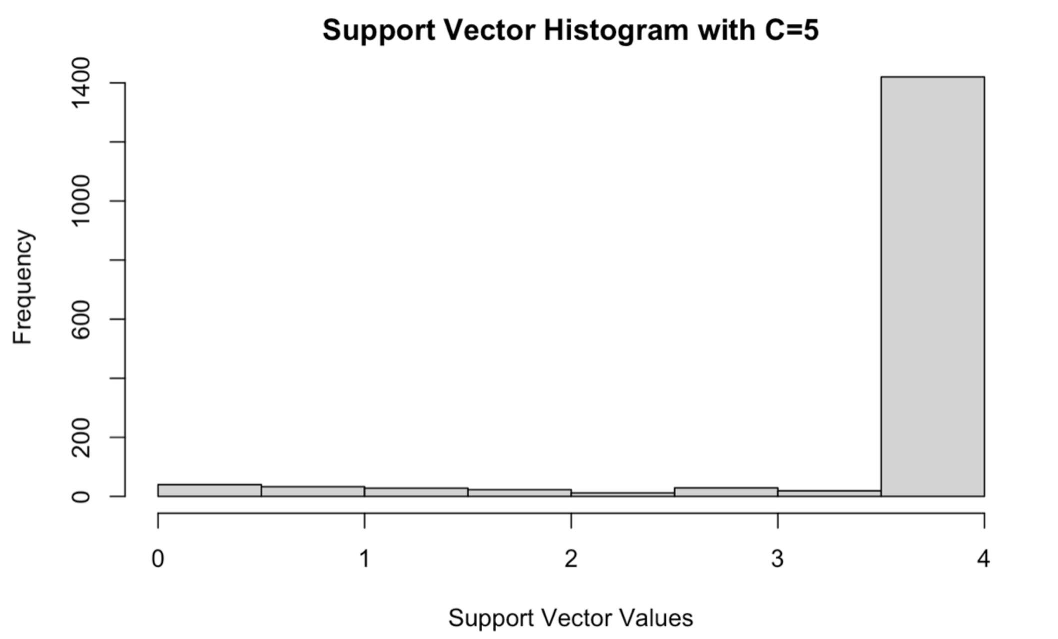


Fig. Supporting Histogram for SVM

6.3 Decision Tree

Regression Trees was the third model that we opted to use. Using the same significant predictors before.

```{r}
tree<-rpart(target ~ age+bmi+children+smoker+exercise+hypertension+yearly_physical, data = training, method =
"class",
            parms = list(prior = c(.65,.35), split = "information"))

rpart.plot(tree)
```

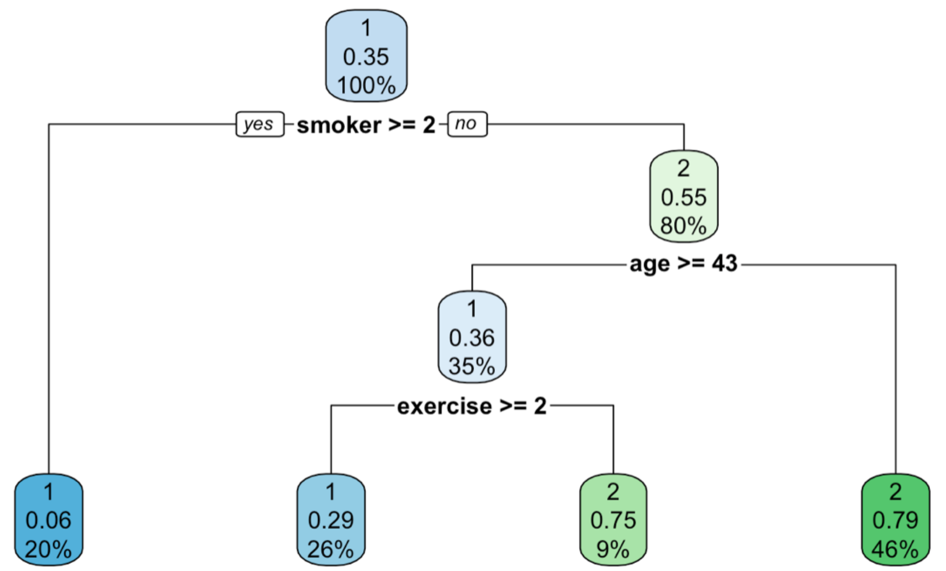


Fig. Decision Tree

```
pred_tree <-predict(tree,testing, type = "class")
confusionMatrix(pred_tree,testing$target)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 405 426
         2  59 965

               Accuracy : 0.7385
                 95% CI : (0.7179, 0.7584)
    No Information Rate : 0.7499
    P-Value [Acc > NIR] : 0.8752

                  Kappa : 0.4484

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8728
            Specificity : 0.6937
         Pos Pred Value : 0.4874
         Neg Pred Value : 0.9424
             Prevalence : 0.2501
         Detection Rate : 0.2183
   Detection Prevalence : 0.4480
      Balanced Accuracy : 0.7833

       'Positive' Class : 1
```

Fig. Decision Tree Accuracy and Sensitivity

7. Findings

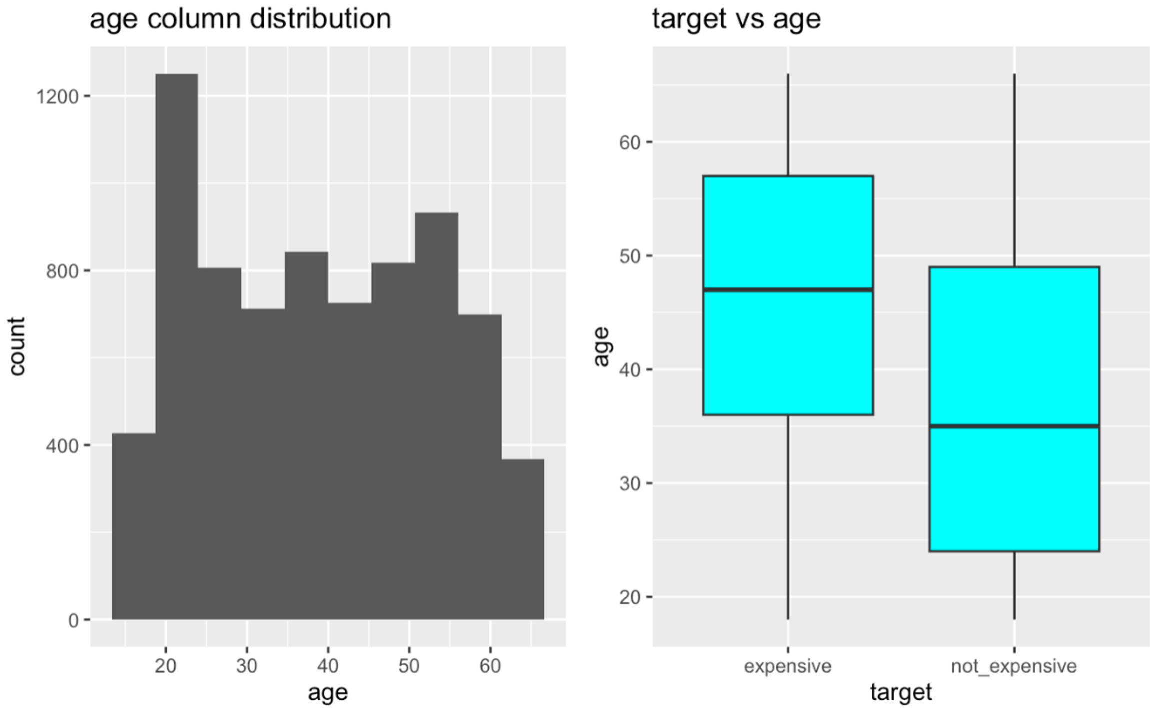


Fig. Rough analysis of entire dataset based on age and cost

People under 20 are more in count, followed by 20-25 and 45-55, least data is from people of age 65-70.
From a rough analysis as seen above, we can see that aged people have more medical expense than young ones. Mean age of expensive is around 57 whereas non_exp is 36.

To investigating further we first created 3 categories based on age, namely: Aged, Senior adults and Young_adult.

```
df = df%>%mutate(age_cat=case_when(
                                    age>=18 & age<35 ~'Young_adult',
                                    age>=35 & age<50 ~'Seniors_adult',
                                    age>=50  ~'Aged',
))
```

Fig. Segregating based on age

We then found out that indeed people over the age of 50 have significantly higher medical costs compared to other age category. They spent an average of $4022 per year on medical expenses. Senior adults those within the age of 50 and 35 spent on average $3028 and with young adults below the age of 35 spent just $861.

````
p <-df%>%group_by(age_cat)%>%summarise(Median =median(cost))
p1 <-ggplot(as.data.frame(p),aes(x=age_cat,y=Median))+geom_bar(stat="identity", position =
"dodge",fill='cyan')+geom_text(aes(label = Median), vjust = 0)
p1
```
````

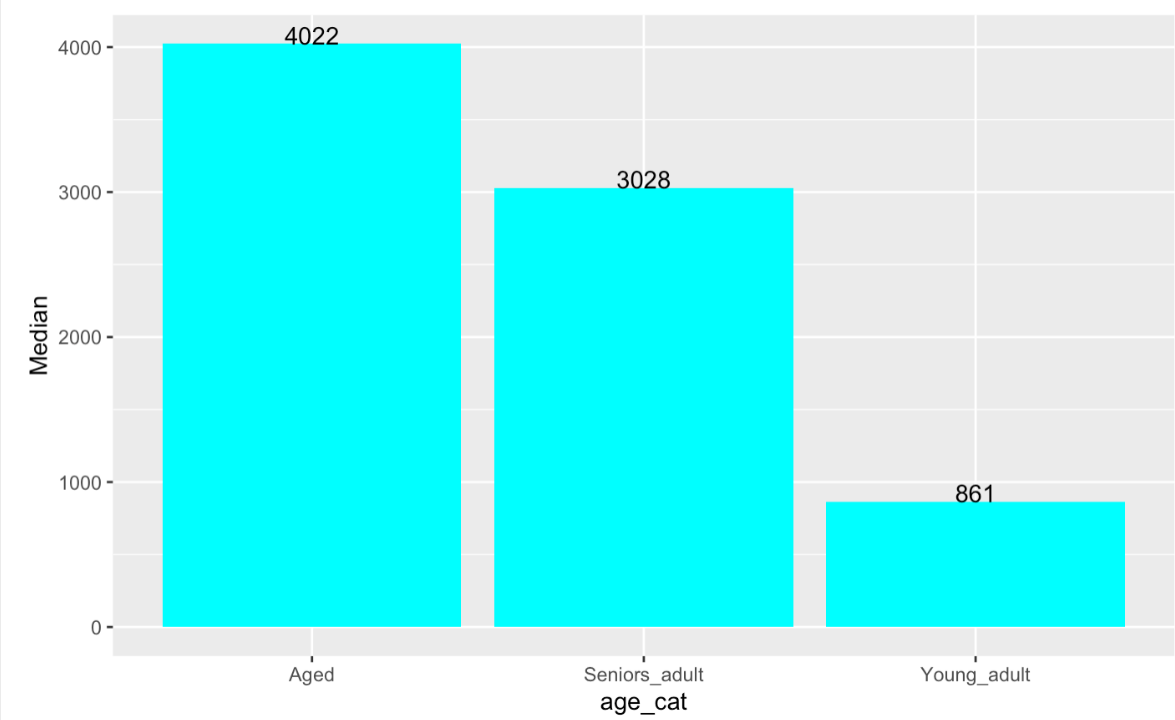


Fig. Analysis based on age category

We then decided to look at smokers and non-smokers present in our dataset.

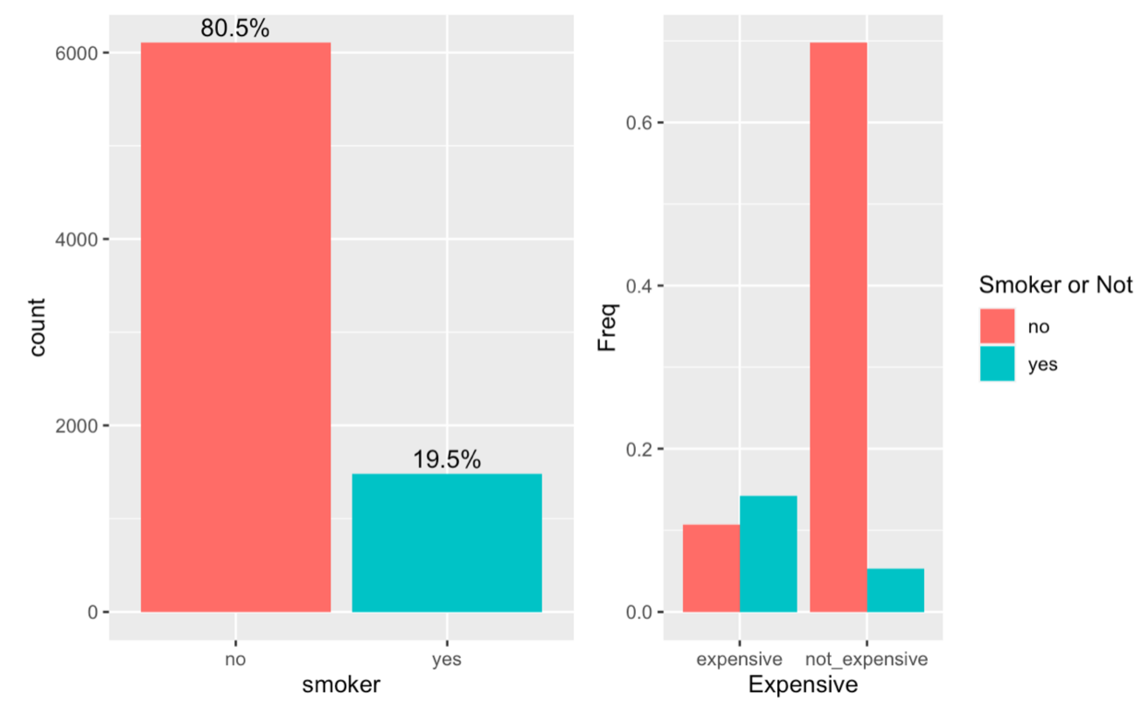


Fig. Smoker and Non-Smokers
````

We can see that in our dataset 80.5% are non-smokers and other 19.5% are smokers. On comparing with cost, we found out that out of the 19.5% smokers, 15% have expensive costs. As smoking was also a significant predictor, we looked at it further.

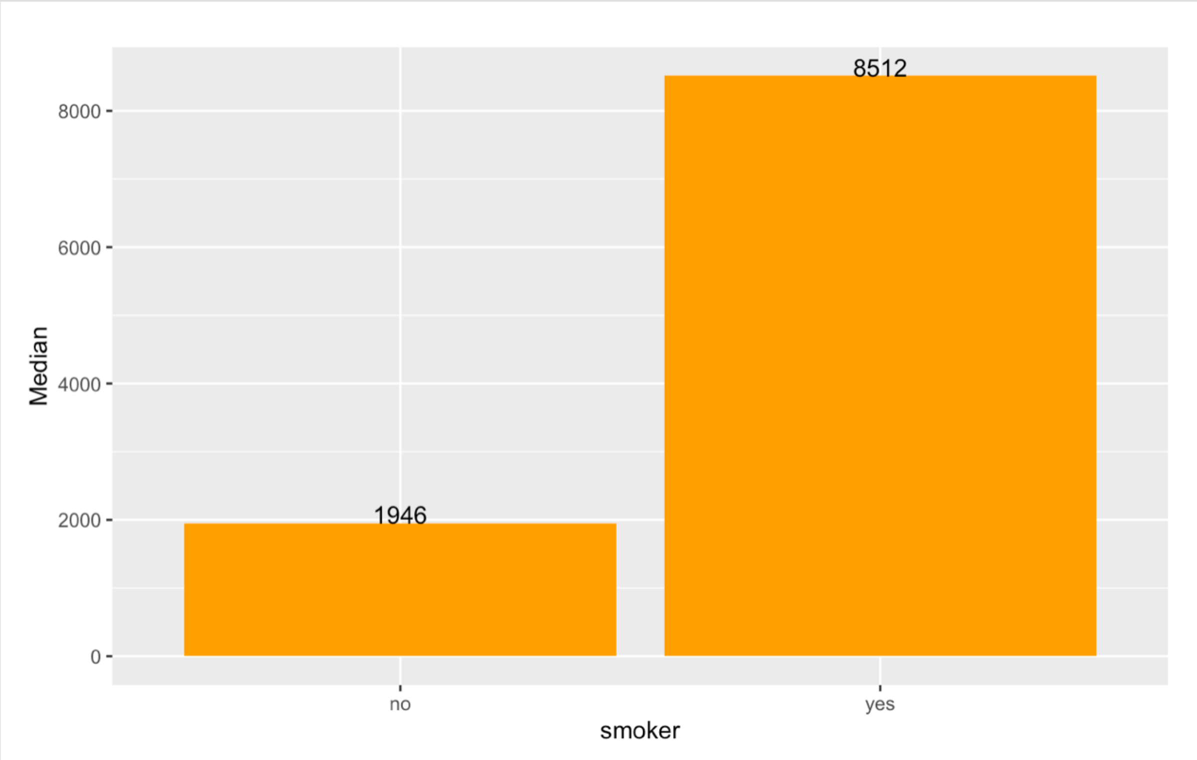


Fig. Analysis based on smoking habits

As expected, people who smoke have significantly higher medical costs. They spend around $8500 and more than 4x what non-smokers spend on medical related costs.

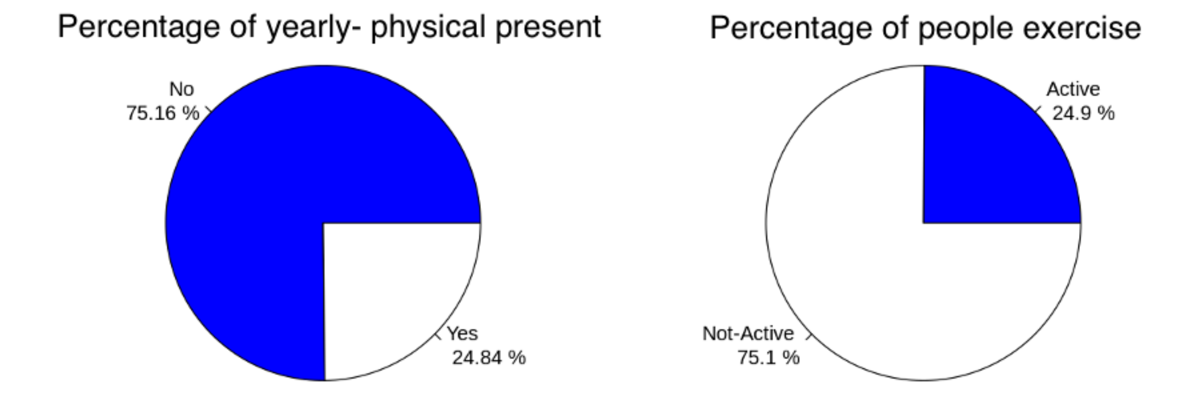


Fig. Co-relation between yearly physical visit and % people exercising

There was interesting co-relation between yearly physical visits and the percentage of people exercising.

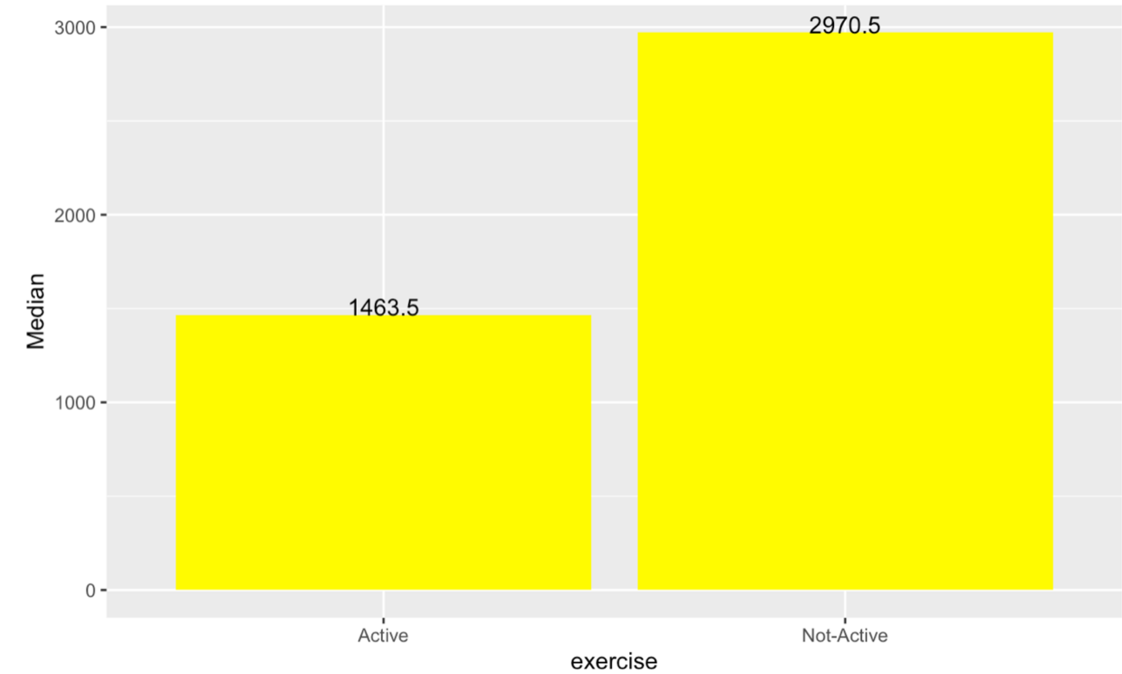


Fig. Analysis based on exercise

Those who exercise and are active spend less around $1463.5 whereas those who are not active spend more. This also has an interesting relation with yearly_visits which denotes whether they had a well visit with their doctor in that year.

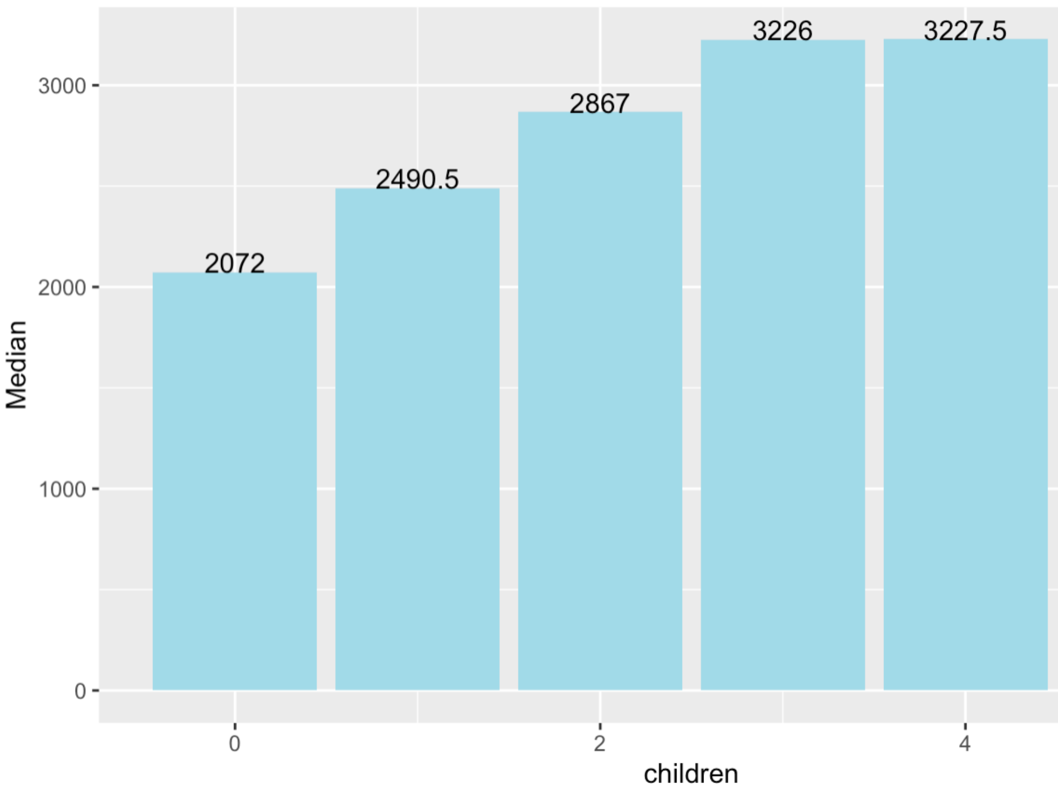


Fig. Analysis of cost based on no.of childrens

We also found during our analysis that people with more children spend more on healthcare. This is expected as the chances of someone experiencing medical problems increases which indirectly increases medical expenses.

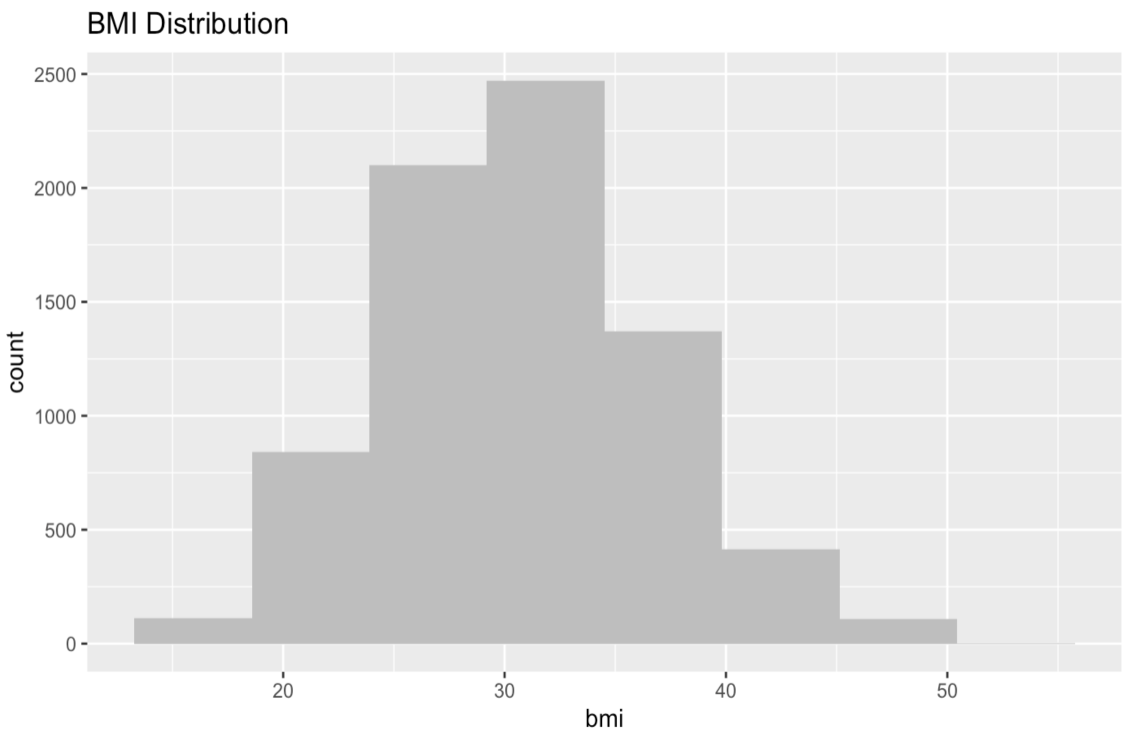


Fig. BMI distribution of the dataset

It is a normal distribution. Mean BMI is 30, which according to study is obesity zone. Indicating more people in data are obese.
Probably obese people have higher medical expense. Mean is 33 for expensive, whereas non_exp it is almost 30.

Before generating final analysis, we divided the dataset into categories like we did for age to make it easy to present our analysis to stakeholders.

```
df = df%>%mutate(bmi_cat=case_when(
                                   bmi<18.5 ~'Underweight',
                                   bmi>=18.5 & bmi<25 ~'Healthy',
                                   bmi>=25 & bmi<30 ~'Overweight',
                                   bmi>=30 ~'Obesity'
))
```

Fig. BMI Categories

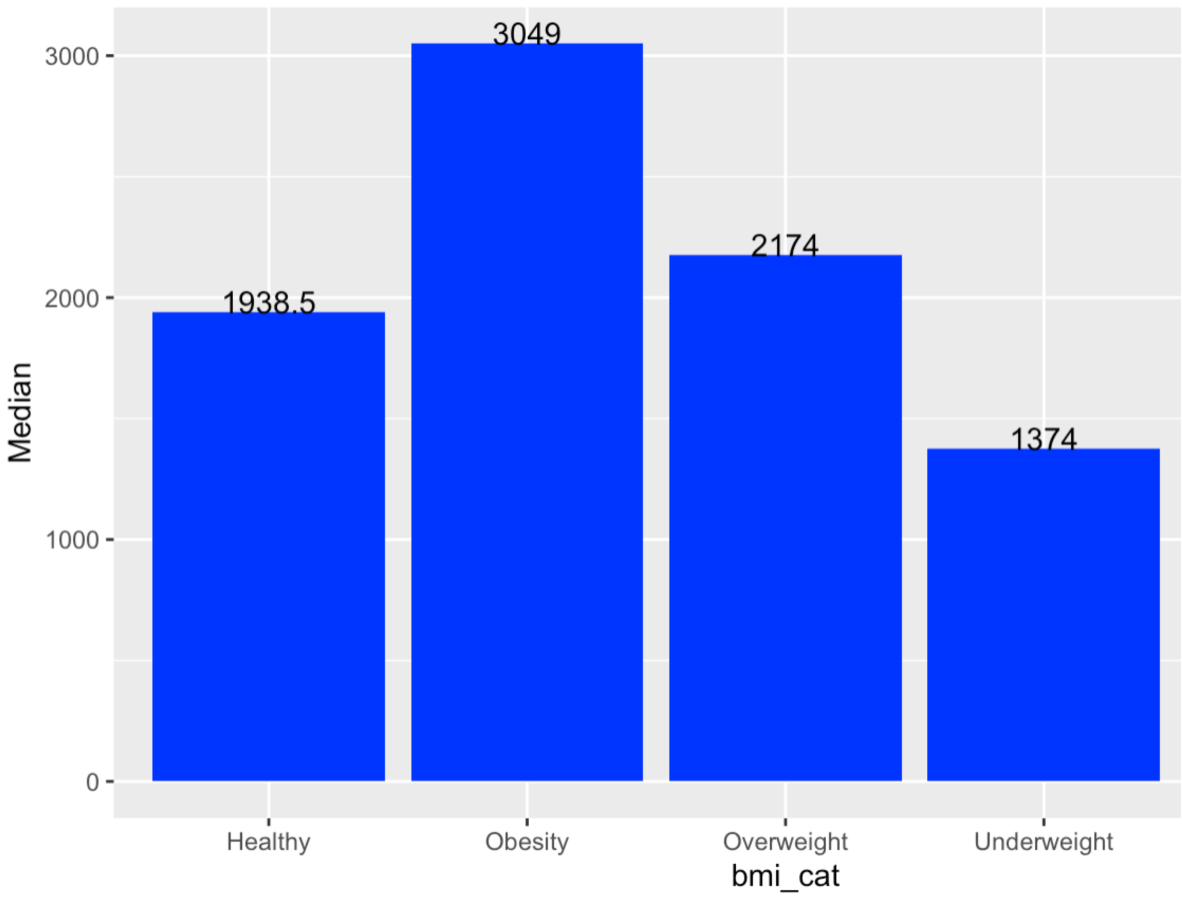


Fig. Analysis based on BMI

On further analysis, we can clearly see that Obese people spend more on healthcare.

Finally, we looked at how location is playing a role in influencing health care costs.

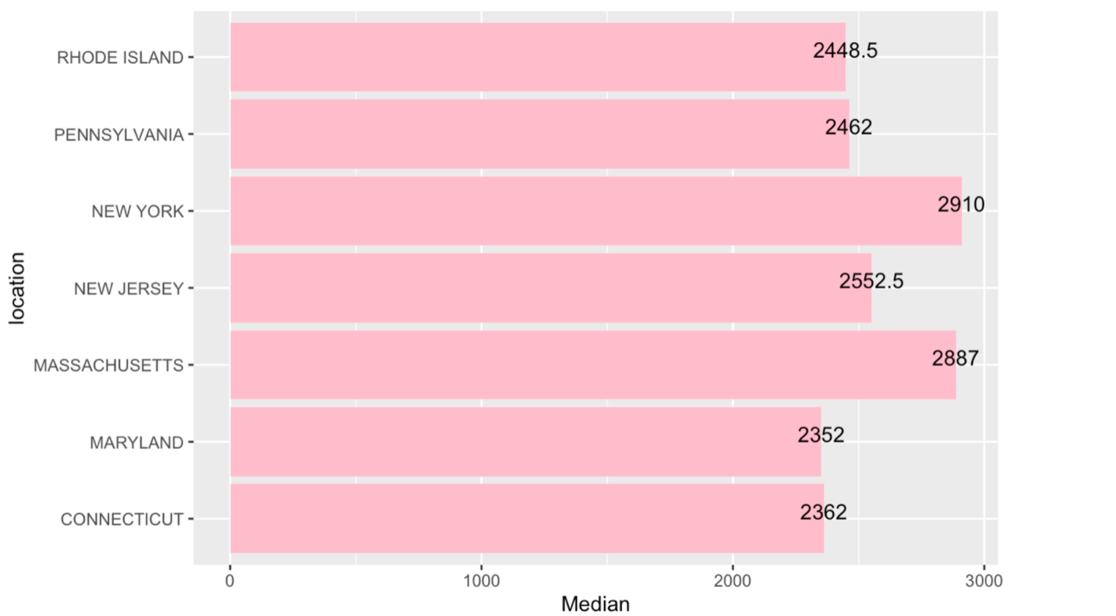


Fig. Breakdown of cost by Location

New York and Massachusetts have the highest medical costs.

8. Recommendation

We would recommend HMO should introduce insurance policies based on various factors such as smoking habits, location, BMI.
HMO should look into opening small health clinics in big cities so that people have more affordable healthcare options.

9. Project Link (Shinnyapp)

https://dsproject-ist687-group3.shinyapps.io/ShinyApp_Group3_IST687/