Data Analysis on healthcare cost

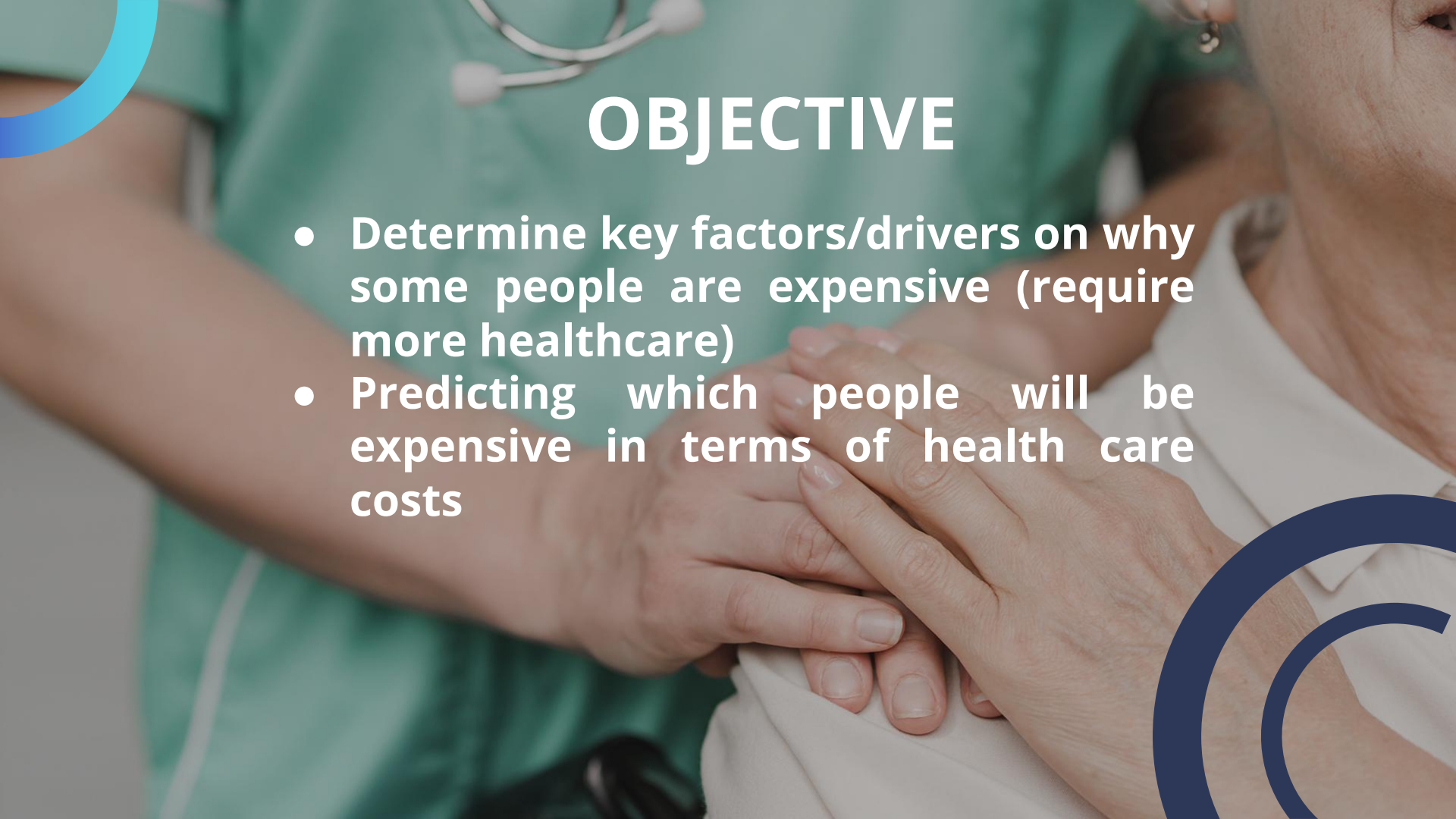
> Hiral Paghadal
> Kabir Thakur
> Sagnik Das
> Sowmeya M

Contents

Problem statement

Everybody's life is centered around health. Our lives are moving so quickly that we are forming bad habits which affects the health. There are many factors that results in some people paying more in the hospital.

So, a predictive model is used to understand the factors affecting the health and creating more medical bills and cost, further helping to identify the key drivers affecting cost.



OBJECTIVE

- Determine key factors/drivers on why some people are expensive (require more healthcare)
- Predicting which people will be expensive in terms of health care costs

Dataset and variables

Data Preprocessing and cleaning

```
sum(is.na(data$X))
sum(is.na(data$age))
sum(is.na(data$bmi))
sum(is.na(data$children))
sum(is.na(data$smoker))
sum(is.na(data$location))
sum(is.na(data$location_type))
sum(is.na(data$education_level))
sum(is.na(data$yearly_physical))
sum(is.na(data$exercise))
sum(is.na(data$hypertension))
```

```
[1] 0
[1] 0
[1] 78
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 0
[1] 80
```

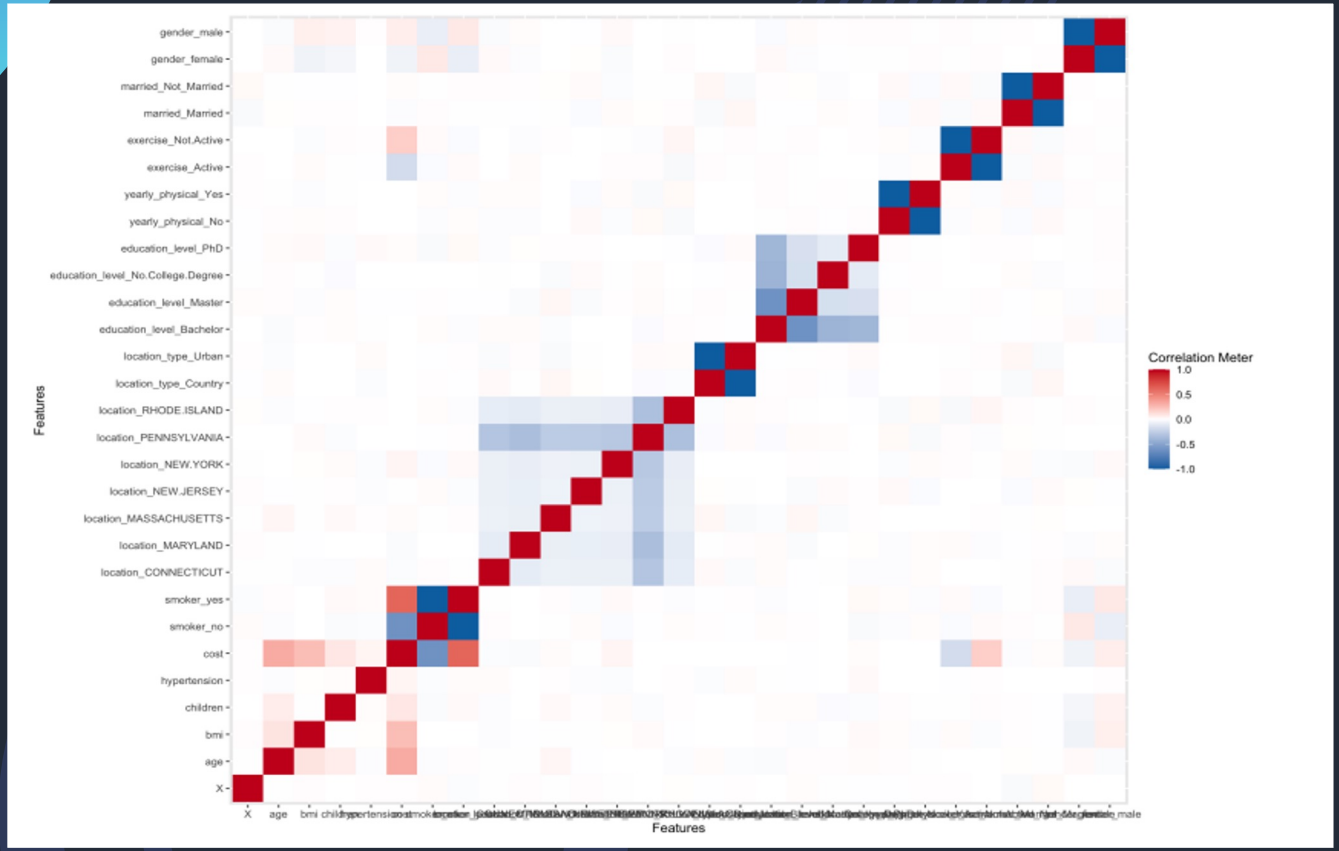
Finding NA values

```
```{r}
sum(is.na(data))
```
```

```
[1] 158
```

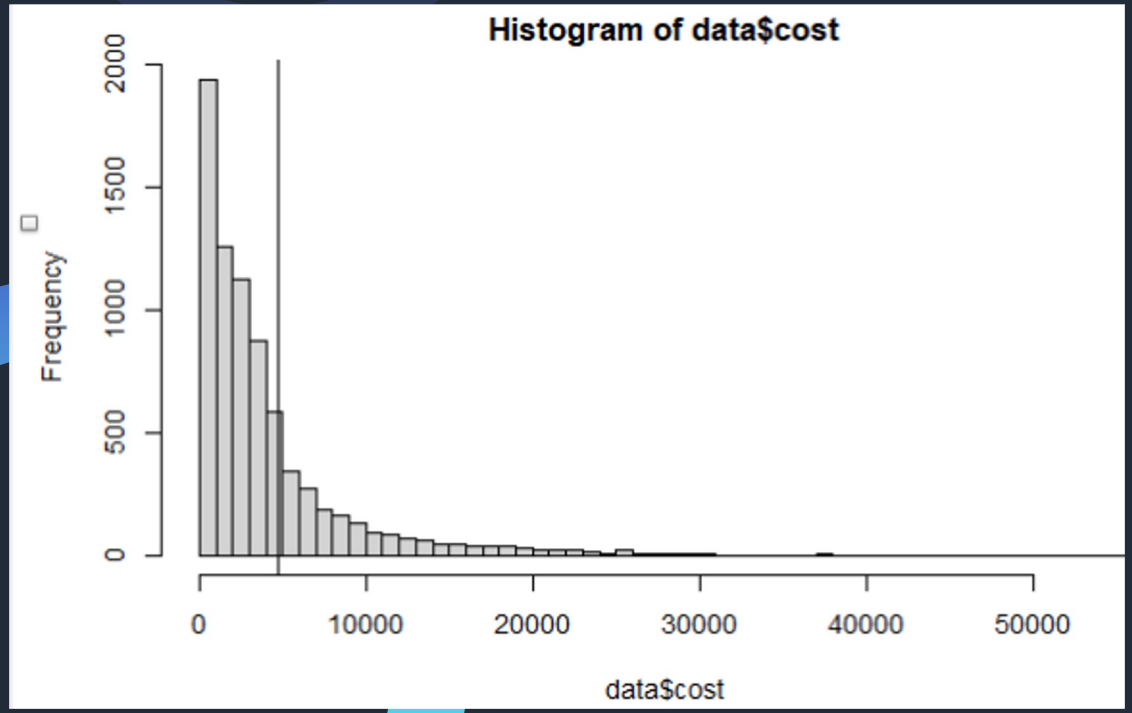
- Bmi and hypertension were the only variables that had NA
- Using na_interpolation to remove the NA from above numerical variables
- Converting categorical variables to numerical variables

Correlation Analysis



Univariate Analysis

Histogram of cost distribution



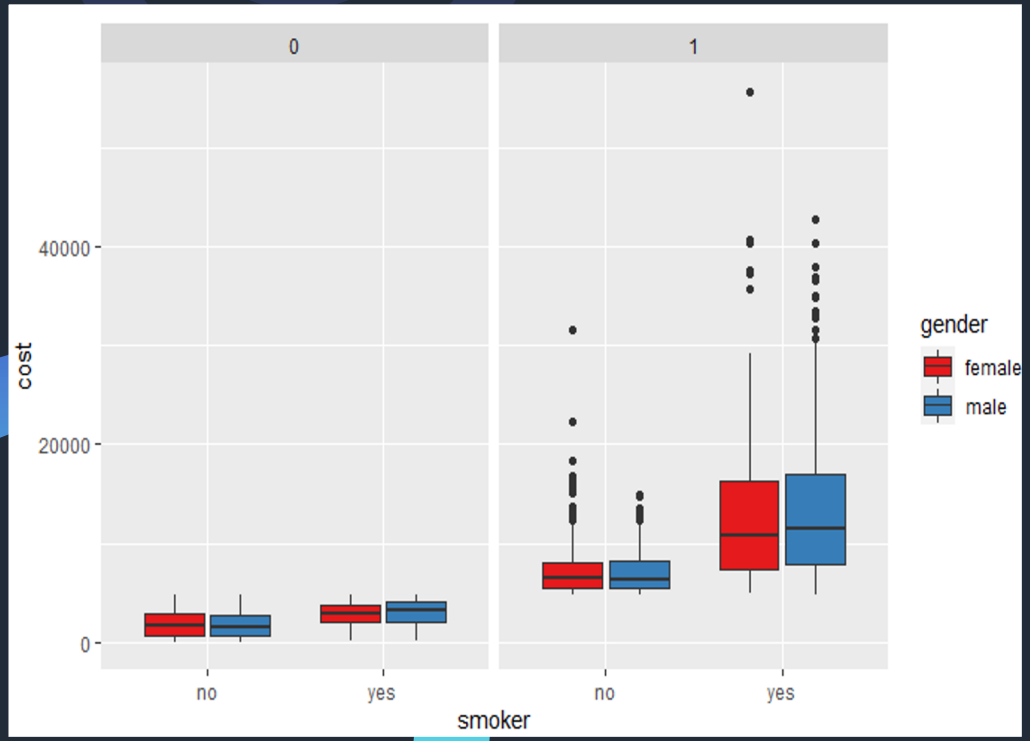
We decided to have the splitting line at 75 % quantile of cost

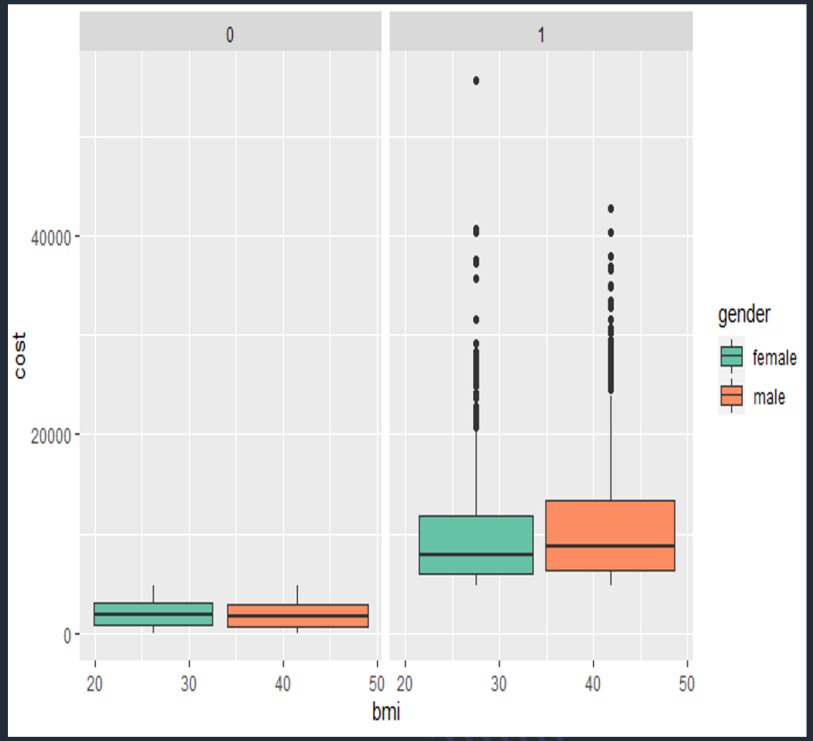
75%
4775

Individual is termed “expensive” when the cost is greater than 4775

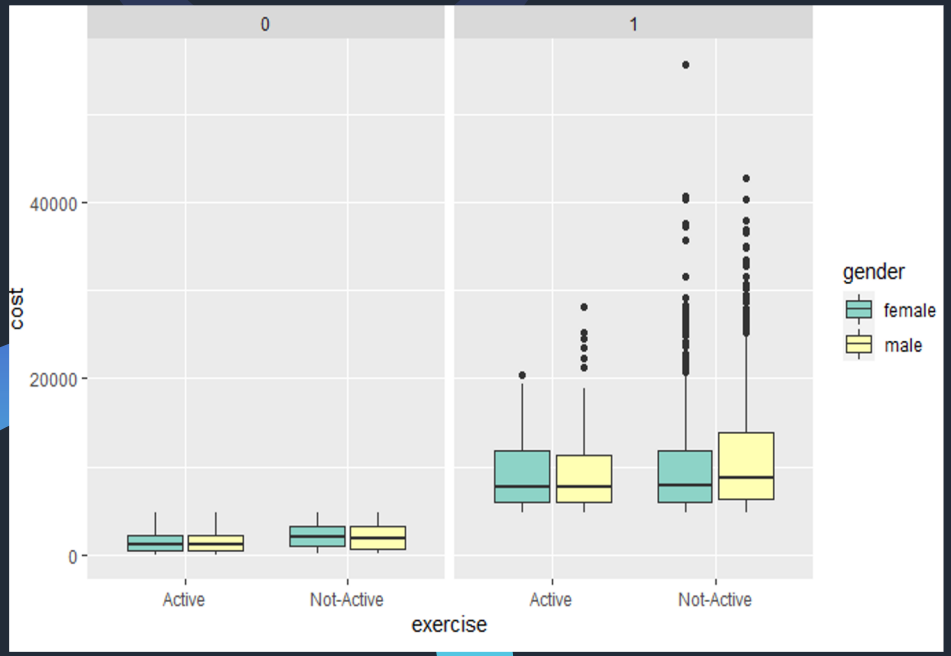
Termed “non-expensive” when the cost is lower than 4775

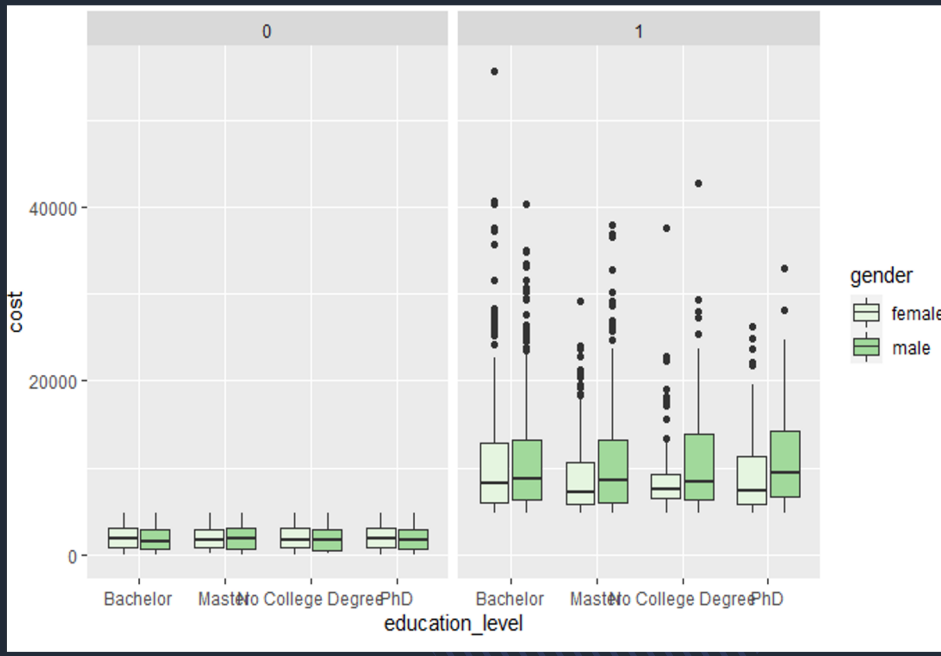
Univariate Analysis



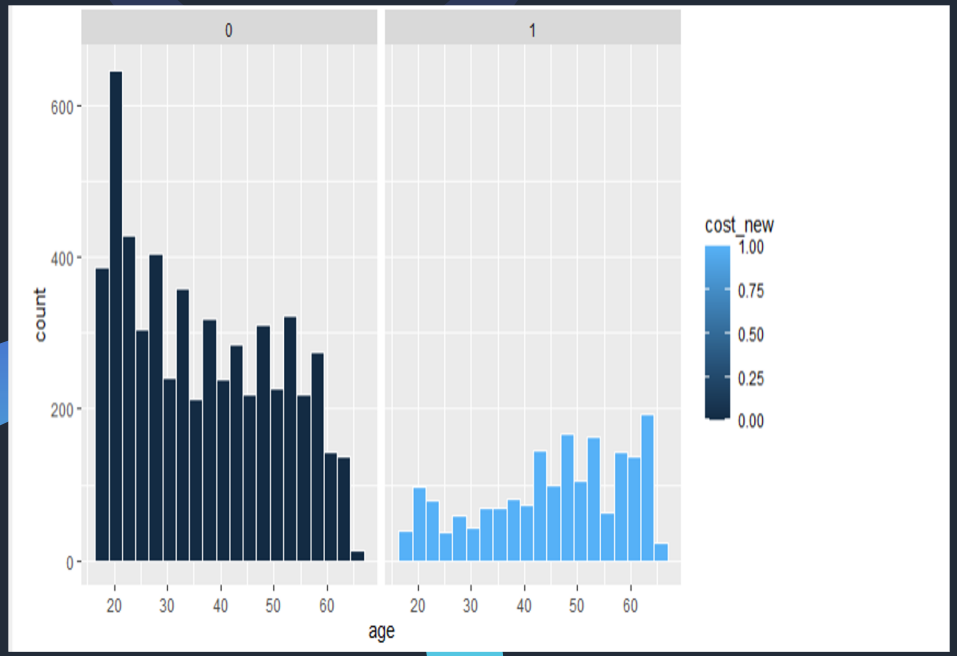


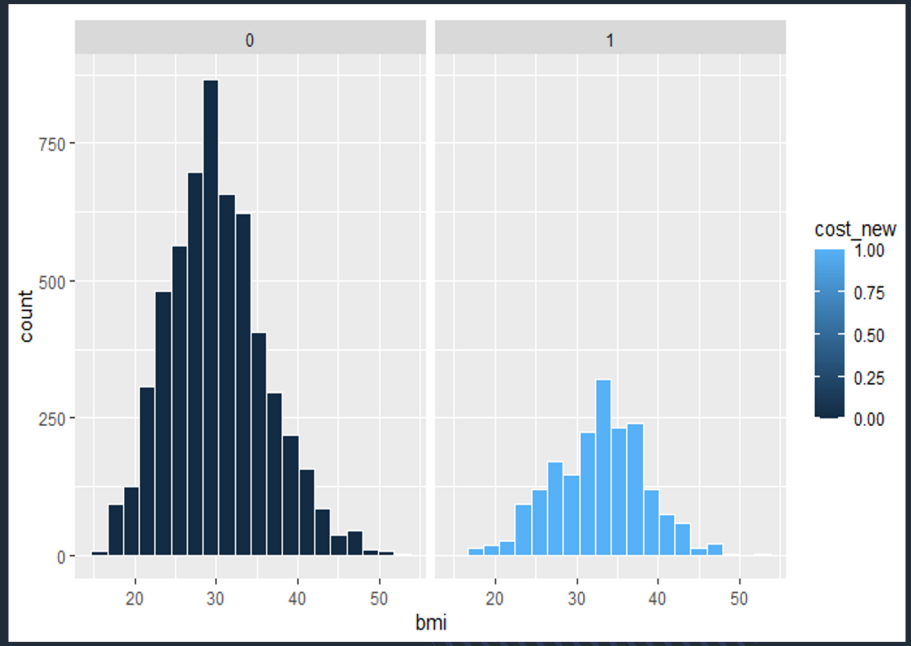
Univariate Analysis





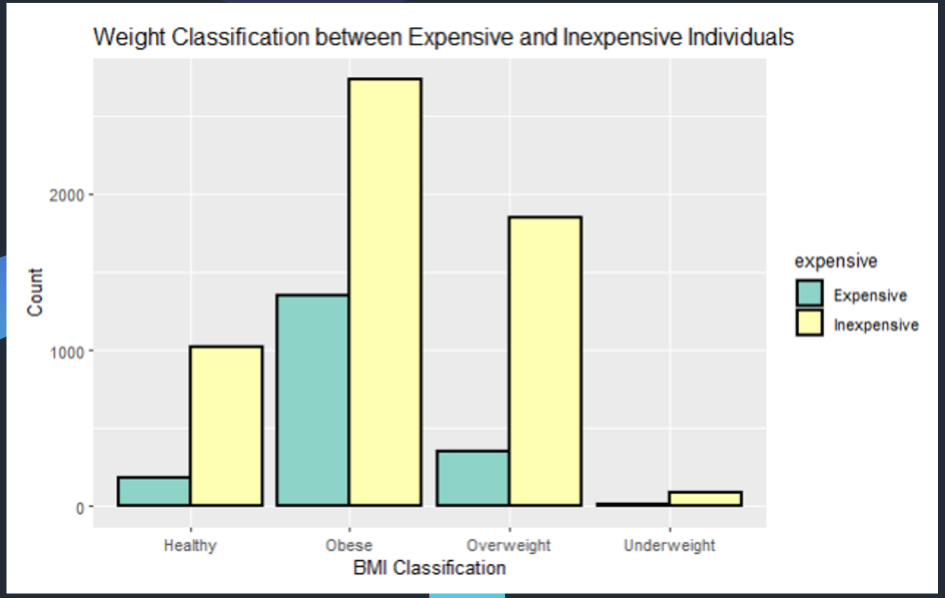
Univariate Analysis



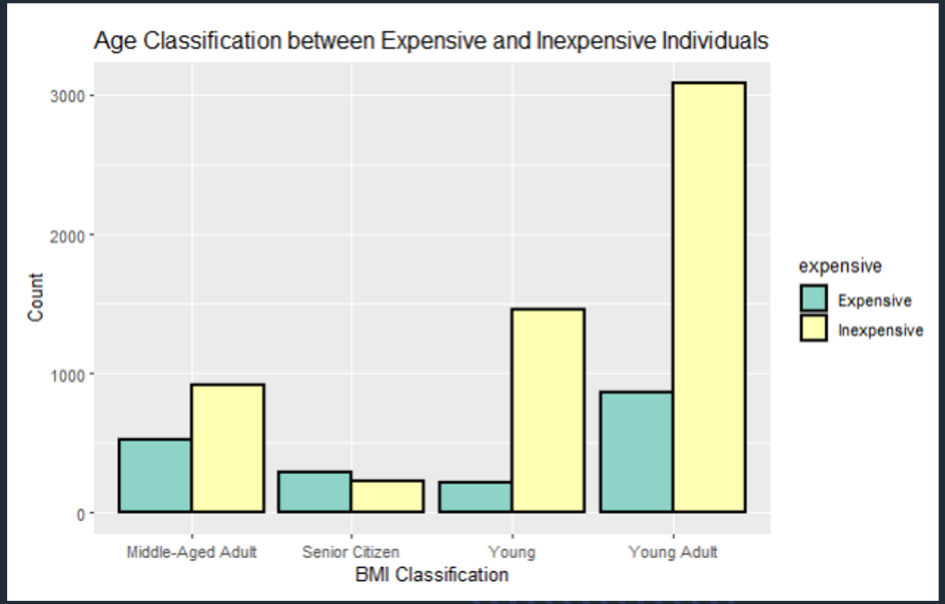


Univariate Analysis

Weight Classification between Expensive and inexpensive individual

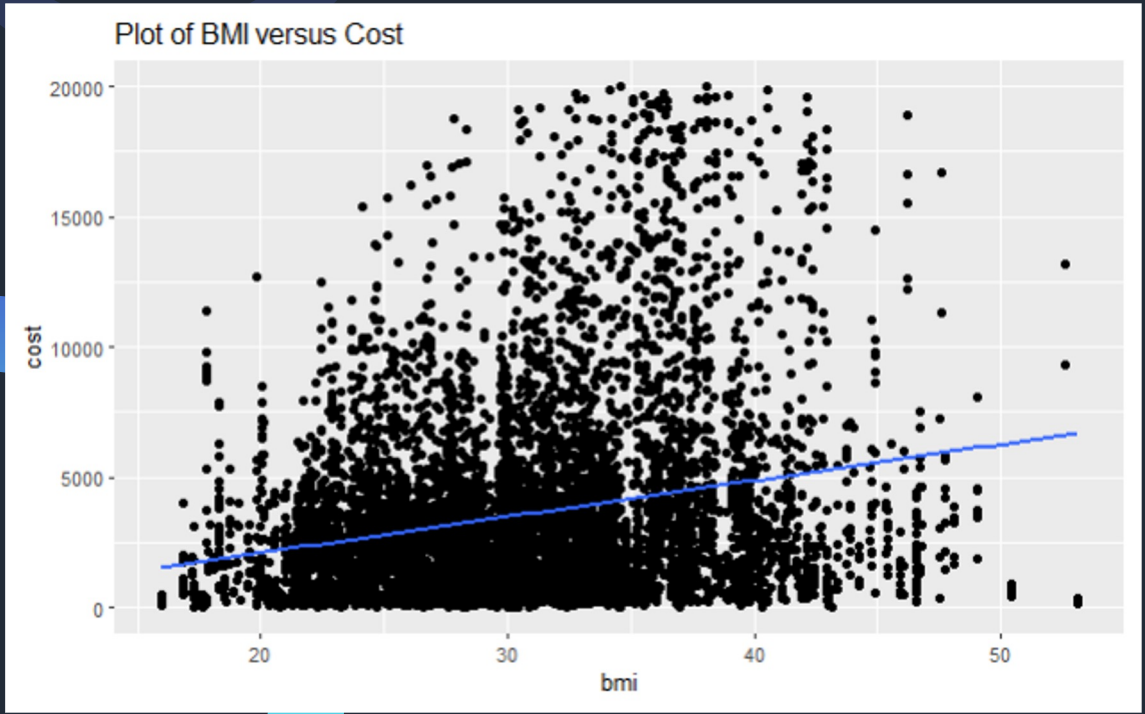


Age Classification between Expensive and Inexpensive Individuals



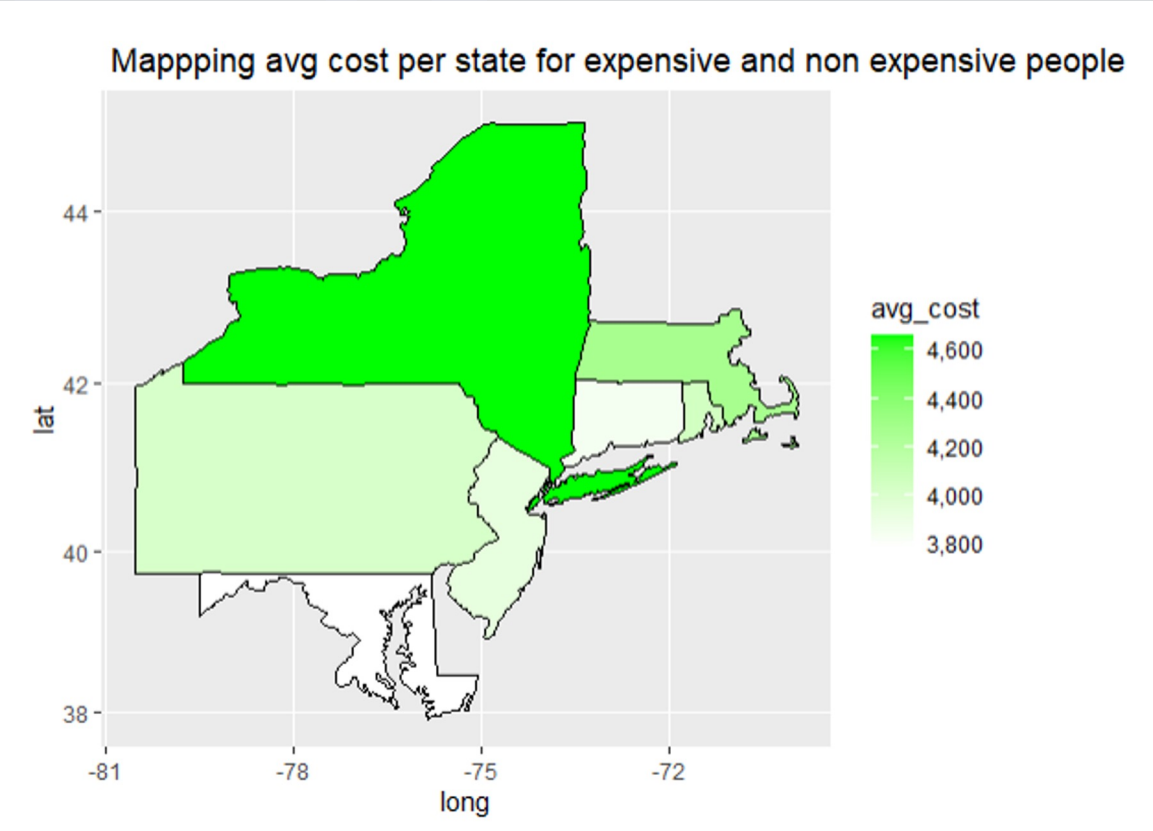
Bivariate Analysis

Scatterplots



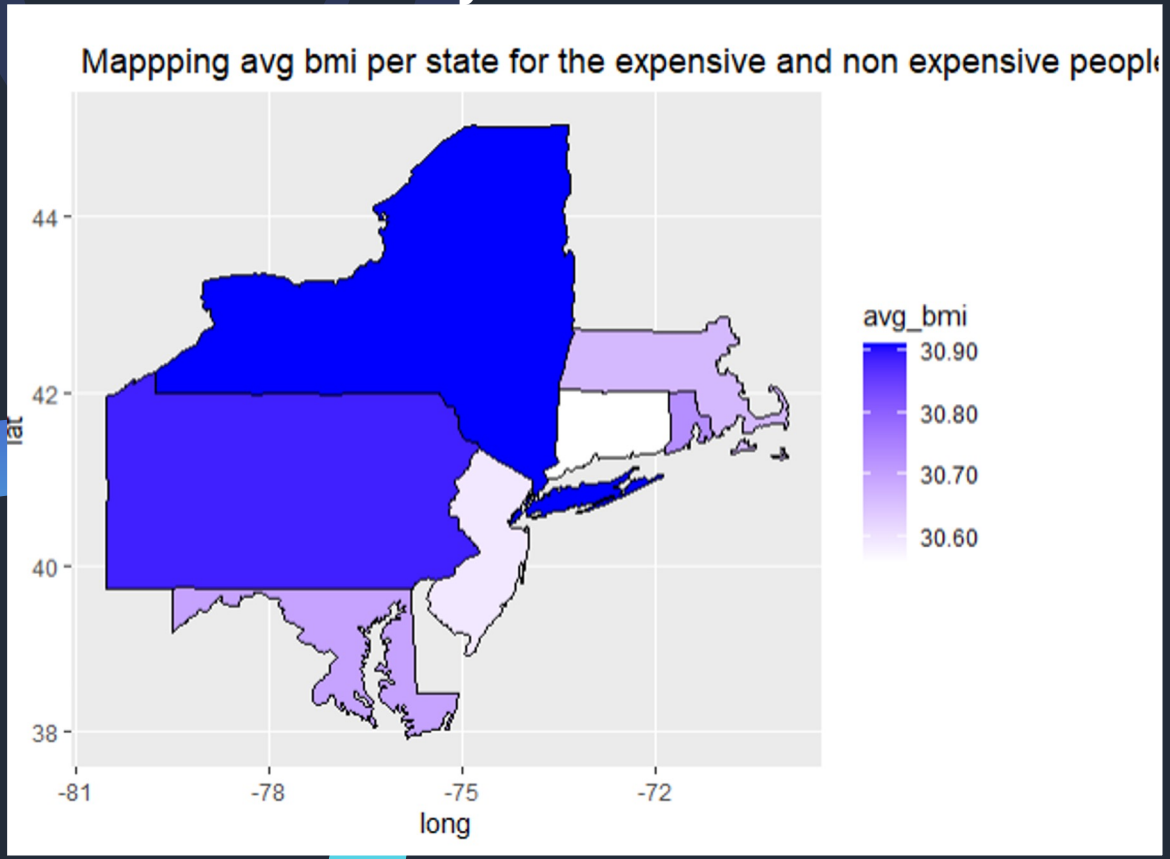
- As the BMI increases, the cost of the patient increases too.

Multivariate Analysis



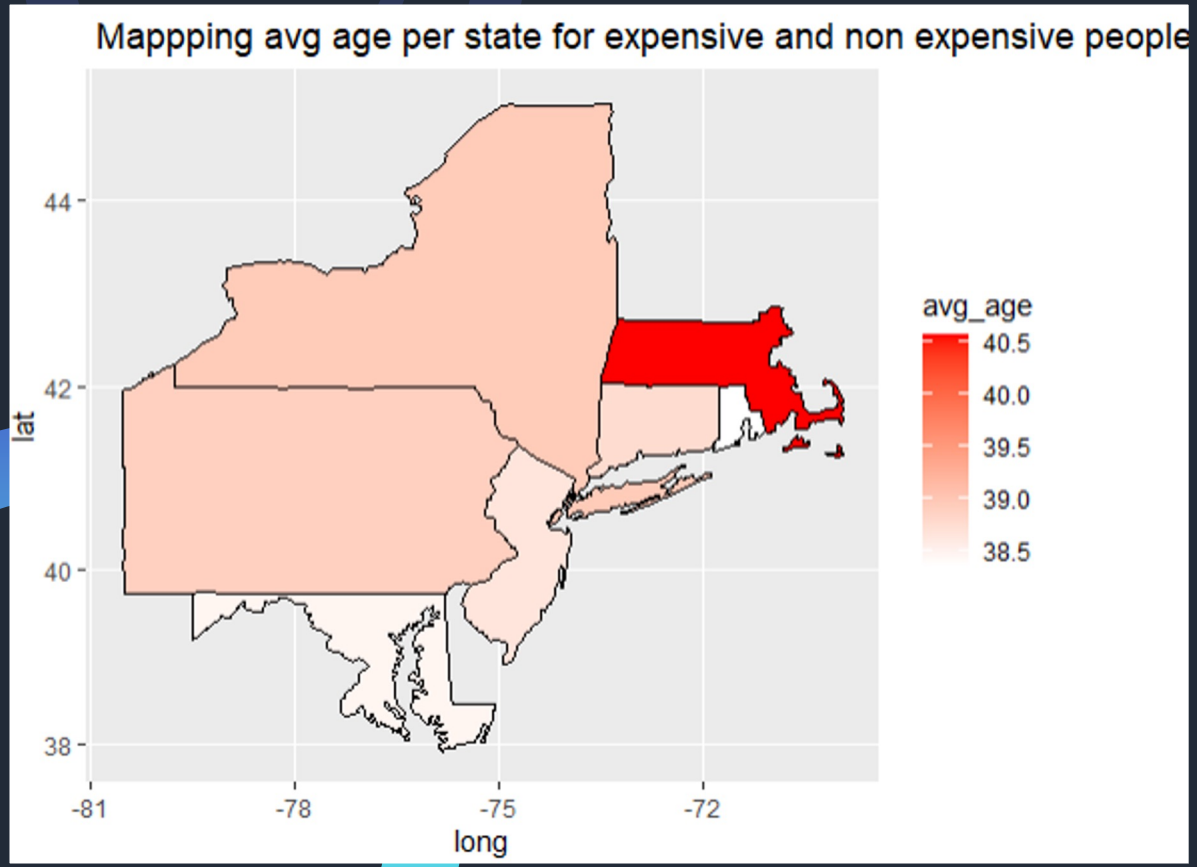
- The average cost of healthcare for an individual is the highest in New York and Massachusetts
- The average cost is lowest in Connecticut

Multivariate Analysis



- The average bmi of people is the highest in New York and Pennsylvania
- The average bmi is lowest in Connecticut

Multivariate Analysis



- The average age of people requiring healthcare is more in Massachusetts
- The average age is lowest in Rhode island

Linear and multiple linear regression model

```
Call:
lm(formula = cost ~ ., data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93560 -0.20358 -0.05617  0.12732  1.15998

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.510255   0.045653 -33.081  < 2e-16 ***
age              0.007504   0.000308  24.359  < 2e-16 ***
bmi              0.013038   0.000728  17.910  < 2e-16 ***
children         0.008794   0.003560   2.470  0.01352 *
smoker           0.591129   0.011025  53.618  < 2e-16 ***
location         0.001699   0.002357   0.721  0.47101
location_type   -0.018316   0.010053  -1.822  0.06850 .
education_level  0.002229   0.004391   0.508  0.61169
yearly_physical  0.024928   0.010043   2.482  0.01308 *
exercise         0.167360   0.010044  16.662  < 2e-16 ***
married          0.006766   0.009182   0.737  0.46127
hypertension     0.030518   0.010870   2.808  0.00501 **
gender           0.018559   0.008722   2.128  0.03338 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3264 on 5674 degrees of freedom
Multiple R-squared:  0.4256,    Adjusted R-squared:  0.4244
F-statistic: 350.4 on 12 and 5674 DF,  p-value: < 2.2e-16
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1321  212
         1   73  289

               Accuracy : 0.8496
                 95% CI : (0.8327, 0.8654)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5756

 Mcnemar's Test P-Value : 2.973e-16

            Sensitivity : 0.9476
            Specificity : 0.5768
         Pos Pred Value : 0.8617
         Neg Pred Value : 0.7983
             Prevalence : 0.7356
         Detection Rate : 0.6971
   Detection Prevalence : 0.8090
      Balanced Accuracy : 0.7622

       'Positive' Class : 0
```

Tree Bag Model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1324  142
         1   70  359

               Accuracy : 0.8881
                 95% CI : (0.8731, 0.902)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6985

 Mcnemar's Test P-Value : 1.081e-06

            Sensitivity : 0.9498
            Specificity : 0.7166
         Pos Pred Value : 0.9031
         Neg Pred Value : 0.8368
             Prevalence : 0.7356
         Detection Rate : 0.6987
   Detection Prevalence : 0.7736
      Balanced Accuracy : 0.8332

       'Positive' Class : 0
```

	Overall <dbl>
bmi	100.0000000
age	93.4208907
smoker	54.9893379
exercise	25.2835446
location	22.8062861
children	20.6775534
education_level	13.2378651
gender	2.3296319
married	1.1912997
yearly_physical	0.4075212

This model clearly indicates that that the bmi, age and smoker are in the top in determining the healthcare as they are the most significant

SVM Model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1362  232
         1   32  269

               Accuracy : 0.8607
                 95% CI : (0.8443, 0.876)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5893

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

            Sensitivity : 0.9770
            Specificity : 0.5369
         Pos Pred Value : 0.8545
         Neg Pred Value : 0.8937
             Prevalence : 0.7356
         Detection Rate : 0.7187
   Detection Prevalence : 0.8412
      Balanced Accuracy : 0.7570

       'Positive' Class : 0
```

	Overall <dbl>
smoker	100.00000000
age	23.57877920
bmi	14.56006341
exercise	7.80266838
gender	1.78182397
children	0.85696299
hypertension	0.38021220
location_type	0.11558424
education_level	0.08910050
yearly_physical	0.07339697

Significant predictors by Models

	treebag	Linear model (cost~.)	SVM
1	bmi	smoker	smoker
2	age	age	age
3	smoker	bmi	bmi
4	exercise	exercise	exercise

Model Confusion

	treebag	Linear model (cost~.)	Linear model (cost~significant)	SVM
Accuracy	88.81%	84.96%	84.96%	86.07%
Sensitivity	94.98%	94.76%	94.76%	97.70%

Numbers and insights

27%

Men are expensive

21%

Women are expensive

Cost

$4042.961

Average spend per patient

BMI

30.11 **Non expensive**

32.83 **Expensive**

Age

Non expensive **expensive**

ANALYSIS AND RECOMMENDATIONS

Smoker

As smoking affects the health and the health care cost we recommend the joining therapy sessions and rehab centers

BMI

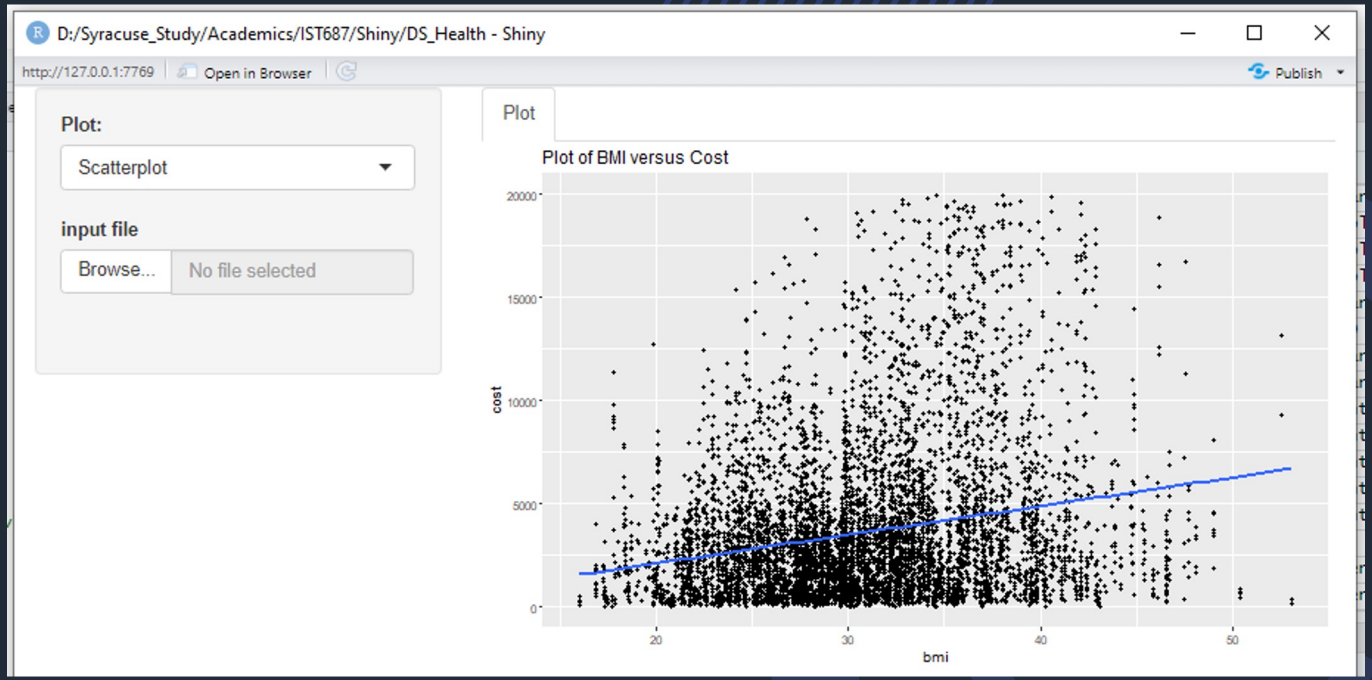
When BMI is high, healthcare cost increases, we therefore recommend, reducing bmi by participating in physical activities. Dieting also helps in reducing BMI, healthcare organizations can have profound dietitians.

Exercise

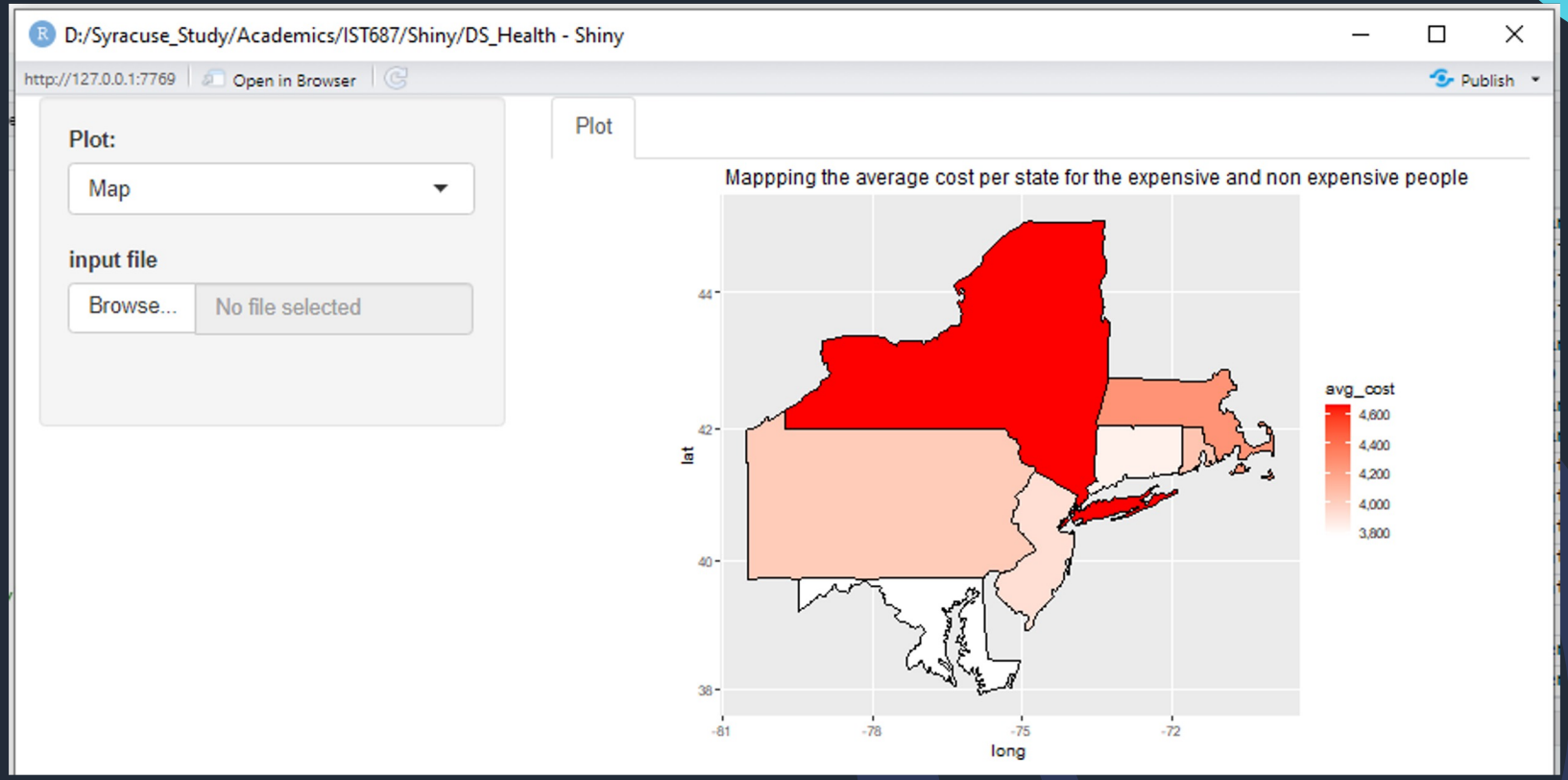
We found non active people in the data has more health care cost, and therefore engaging into daily exercise such as walking, jogging, yoga would significantly reduce the health care cost.

SHINY WEB APP

We have developed a shiny apps which displays some options to the user and based on the user input, it displays a data visualization. For example, if the user selects scatterplot from the dropdown menu, shiny app will generate the scatterplot and display the output as shown below.



SHINY WEB APP



THANK YOU !