A Report on

Health Management Organization

By

Collin Taylor (SUID: 563919675)

Jenil Sheth (SUID: 7754395981)

Shruti More (SUID: 528873433)

Vedant Patil (SUID: 235168671)

Table of Contents

1. PROJECT OVERVIEW:

- Multitude of factors must be taken into account when considering a person's overall health. In this project, we analyze age, location, location type, exercise, smoker, body mass index, yearly physical checkups, hypertension, gender, education level, marital status, number of children, and the cost of the people under the Health Management Organization's Plan to predict which customers.
- As a Health Management Organization, it is important to know which customers are considered expensive versus which customers are not. The monthly premiums that the organization charges are dependent on risk factors associated with each customer/patient.
- Recommendations in this report are established from analyzing the risk factors, and essentially suggesting the development of health programs and the offering of such to the expensive customers.

2. PROJECT TECHNICAL DETAILS:

- There are 14 variables and 7,582 observations available in the dataset.
- The dataset is not overwritten, and columns are not renamed for ease of understanding and coding.
- Null/missing values in the dataset removed for ease of calculation and code roadblocks.
- Identified the categorical and numerical data separately to ease conversion of categorical data to numerical data.
- Encoded the vectors as factors for columns having less distinct values.
- Displaying unique values of all the categorical data.

3. PROJECT GOAL:

The overall goal of the project is to provide actionable insight to the Health Management Organization based on the customer data available. Recommendations are provided in an attempt to lower healthcare costs for the customers under the HMO plan. Predictive models were developed to foresee healthcare costs based upon certain risk factors to general health.

4. PROJECT OBJECTIVES:

1. Load the data with appropriate libraries
2. Explore the data
3. Clean the data, which includes removing null data and incorrectly formatted values
4. Provide an overview of important variables
5. Provide data modeling and develop trend analysis and insight
6. Perform analysis compared to cost such as: age, hypertension, BMI (body mass index), smoker, exercise, offspring, location, among others. Also included are counts of these factors in the dataset.

Questions:

Based on the given data, we can analyze and find answers to the following questions:

1. Does a customer's age impact their health care costs?
2. Does a customer’s BMI impact their health care costs?
3. Does hypertension lead to increased health care costs?
4. Does smoking lead to higher health care costs?
5. Does exercise help lower a customer’s health care costs?
6. What are the variables that can be considered significant risk factors to a customer’s health?
7. From the data collection and analysis, can we make recommendations in an attempt to help a customer lower their health care costs?

5. PACKAGES REQUIRED:

- readr- to provide a fast and friendly way to read rectangular data from delimited files, such as comma-separated values (CSV).
- ggplot2- allows building of almost any type of chart. Greatly improves the quality and aesthetics of graphics.
- kernlab- Support Vector Machines, Spectral Clustering, Kernel PCA, Gaussian Processes and a QP solver.
- caret- a set of functions that attempt to streamline the process for creating predictive models.
- tidyverse- system for declaratively creating graphics, based on The Grammar of Graphics
- dplyr- provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges.
- rio- Streamlined data import and export by making assumptions that the user is probably willing to make
- rpart- Recursive partitioning for classification, regression and survival trees.
- rpart.plot- Plot 'rpart' models. Extends plot.rpart() and text.rpart() in the 'rpart' package.
- e1071- Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, generalized k-nearest neighbour ...
- rpart- Rpart is a powerful machine learning library in R that is used for building classification and regression trees
- randomForest- Random Forest in R used for classification and regression
- arules- The arules package for R provides the infrastructure for representing,manipulating and analyzing transaction data and patterns
- arulesviz- Extends package 'arules' with various visualization techniques for association rules and itemsets.
- mapproj- Converts latitude/longitude into projected coordinates.
- rsample- Classes and functions to create and summarize different types of resampling objects.

6. Data Importing and Cleaning:

```
library(tidyverse)
data <- read_csv("HMO_data.csv")
proj_df <- data.frame(data)
str(proj_df)
summary(proj_df)
```

	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
20	23	19	NA	0	no	PENNSYLVANIA	Urban	No College Degree	No	Active	Not_Married	0	male	146
21	24	32	31.920	1	yes	NEW JERSEY	Urban	No College Degree	Yes	Not-Active	Not_Married	0	female	18100
22	25	37	28.025	2	no	PENNSYLVANIA	Urban	PhD	No	Active	Married	0	male	1496
23	26	58	27.720	3	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Not_Married	0	female	2876
24	27	62	23.085	0	no	PENNSYLVANIA	Country	No College Degree	No	Active	Not_Married	1	female	3541
25	28	56	32.775	2	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	0	female	3962

```
Rows: 7582 Columns: 14── Column specification
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr (8): smoker, location, location_type, education_level, yearly_physical, exercise, married, gender
dbl (6): X, age, bmi, children, hypertension, cost
ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.'data.frame':   7582 obs. of  14 variables:
 $ X              : num  1 2 3 4 5 7 9 10 11 12 ...
 $ age            : num  18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
 $ children       : num  0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : chr  "yes" "no" "no" "no" ...
 $ location       : chr  "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...
 $ location_type  : chr  "Urban" "Urban" "Urban" "Country" ...
 $ education_level: chr  "Bachelor" "Bachelor" "Master" "Master" ...
 $ yearly_physical: chr  "No" "No" "No" "No" ...
 $ exercise       : chr  "Active" "Not-Active" "Active" "Not-Active" ...
 $ married        : chr  "Married" "Married" "Married" "Married" ...
 $ hypertension   : num  0 0 0 1 0 0 0 1 0 0 ...
 $ gender         : chr  "female" "male" "male" "male" ...
 $ cost           : num  1746 602 576 5562 836 ...
```

```
       X                  age              bmi            children       smoker            location          location_type      education_level
 Min.   :       1   Min.   :18.00   Min.   :15.96   Min.   :0.000   Length:7582        Length:7582        Length:7582        Length:7582
 1st Qu.:    5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000   Class :character   Class :character   Class :character   Class :character
 Median :   24916   Median :39.00   Median :30.50   Median :1.000   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :  712602   Mean   :38.89   Mean   :30.80   Mean   :1.109
 3rd Qu.:  118486   3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.:2.000
 Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000
                                     NA's   :78
 yearly_physical      exercise           married         hypertension       gender              cost
 Length:7582        Length:7582        Length:7582        Min.   :0.0000   Length:7582        Min.   :    2
 Class :character   Class :character   Class :character   1st Qu.:0.0000   Class :character   1st Qu.:  970
 Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Median : 2500
                                                          Mean   :0.2005                      Mean   : 4043
                                                          3rd Qu.:0.0000                      3rd Qu.: 4775
                                                          Max.   :1.0000                      Max.   :55715
                                                          NA's   :80
```

Observations:

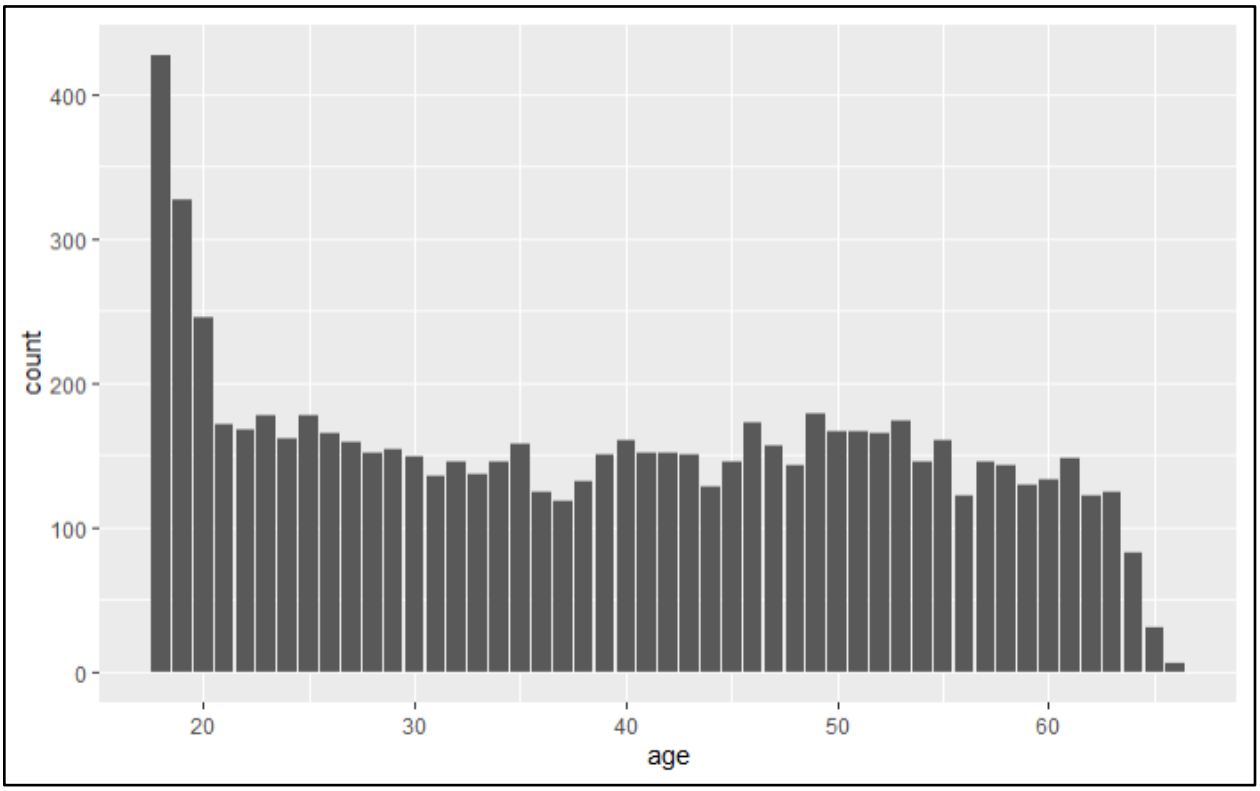
- The HMO data set has 14 columns and 40,060 records.
- All of the dataset for the HMO consists of variables such as Smoker, yearly_physical, exercise, hypertension, BMI, age, gender
- Most of the records from the HMO Data have a location located in the North-Eastern region of the United States of America.
- The age group for the members in the HMO_data is between 10-65 years old.
- The data set consists of Smoker variable for which the table contain yes/no answers based on if the particular member smokes or not. Similarly for hypertension it has 1/0, exercise is has active/non-active and yearly_physical it has yes/no
- BMI is NULL in 78 records and Hypertension is NULL in 80 records.

7. Exploratory Analysis:

For exploration of the data set, we have created few basic plots of certain key columns which gives us an overview of the data and columns in a factored way. Below are the plotsalong with a few lines for explanation of each.

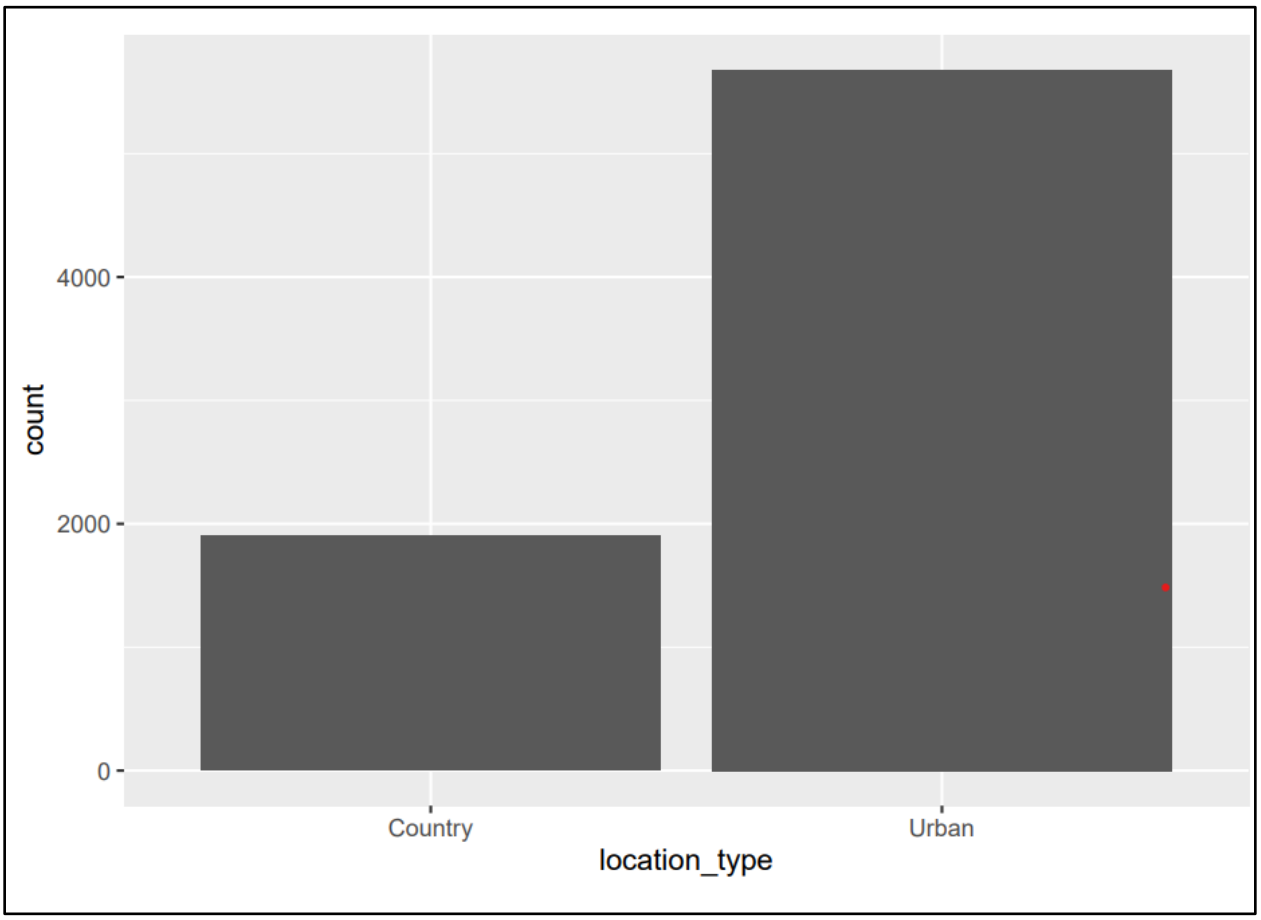
7.1 Basic Plots

7.1.1 Age



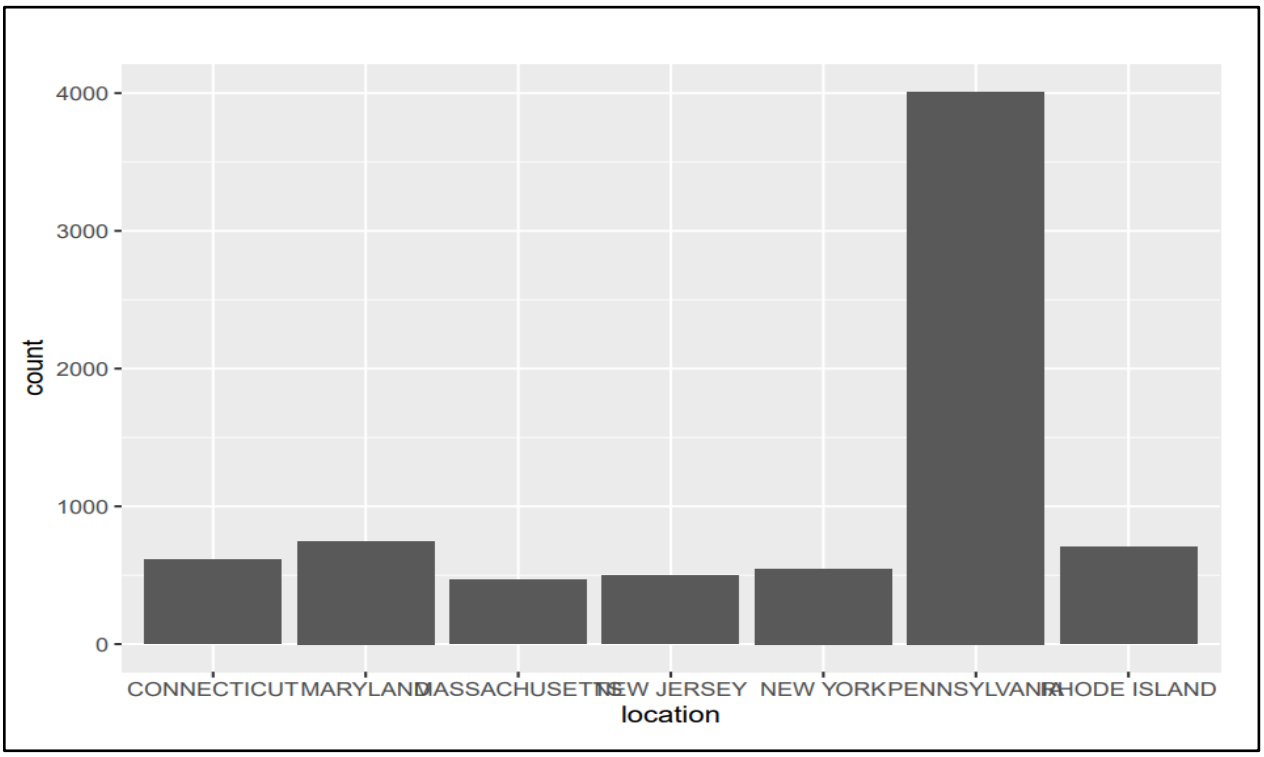
From the above plot, it clearly outlines the age distribution from the Healthcare Data

7.1.2 Location Type



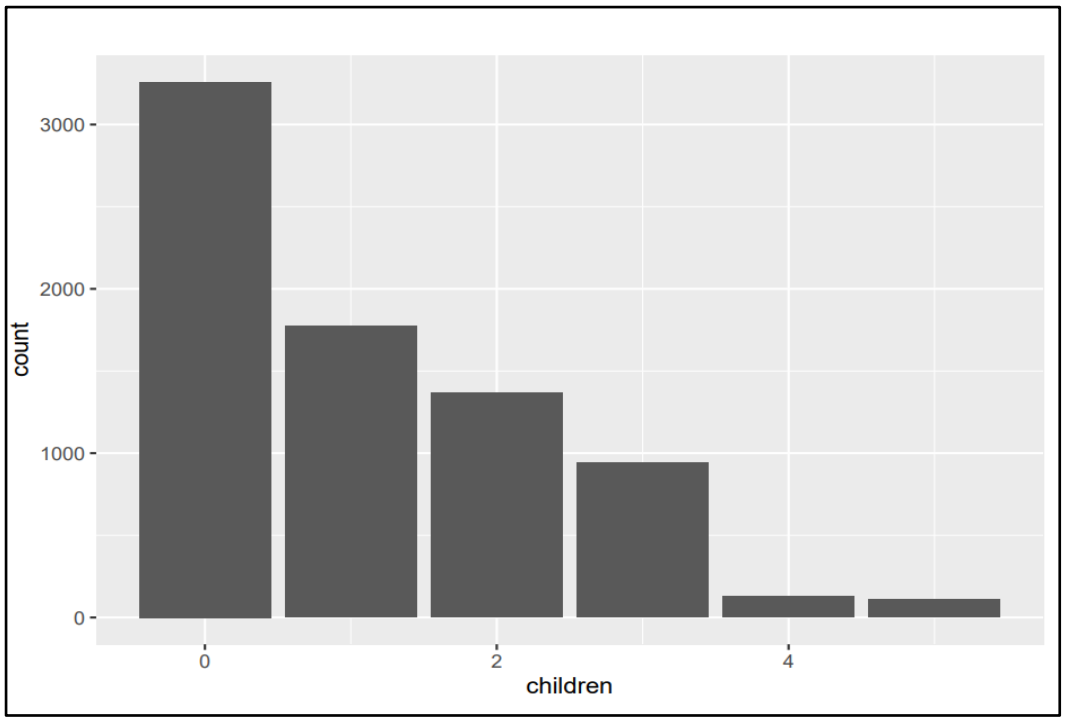
From the above plot, it is clearly visible that the location type is densely populated by urban people compared to the countryside.

7.1.3 Location



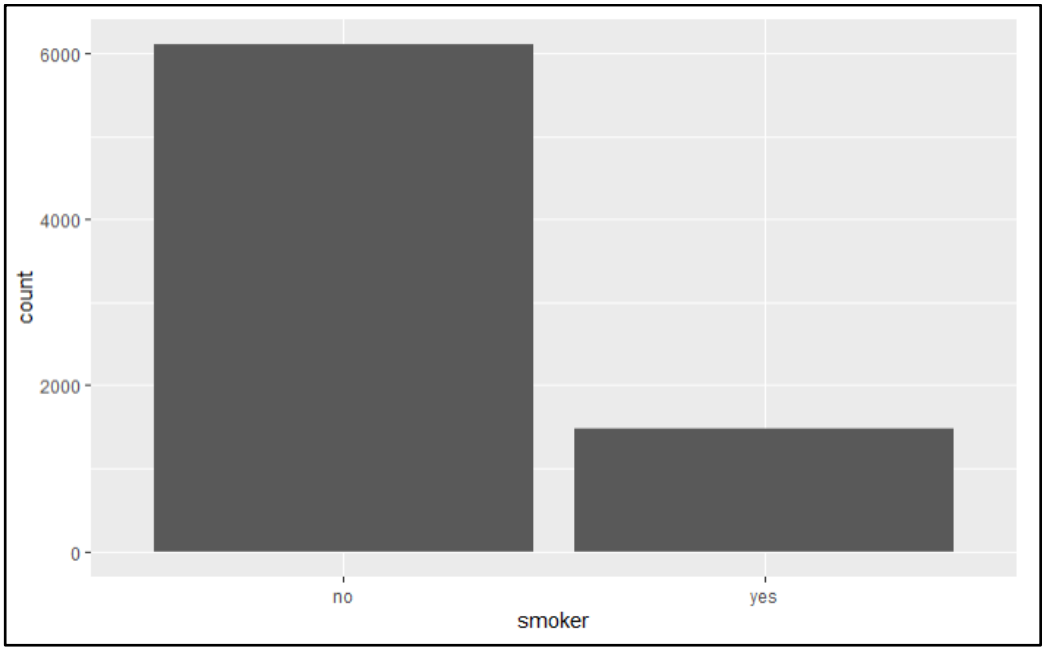
The above plot displays the distribution of HMO data by location in north east region of United States of America showing that it is highly populated in Pennsylvania, New york, Maryland and Massachusetts

7.1.4 Children



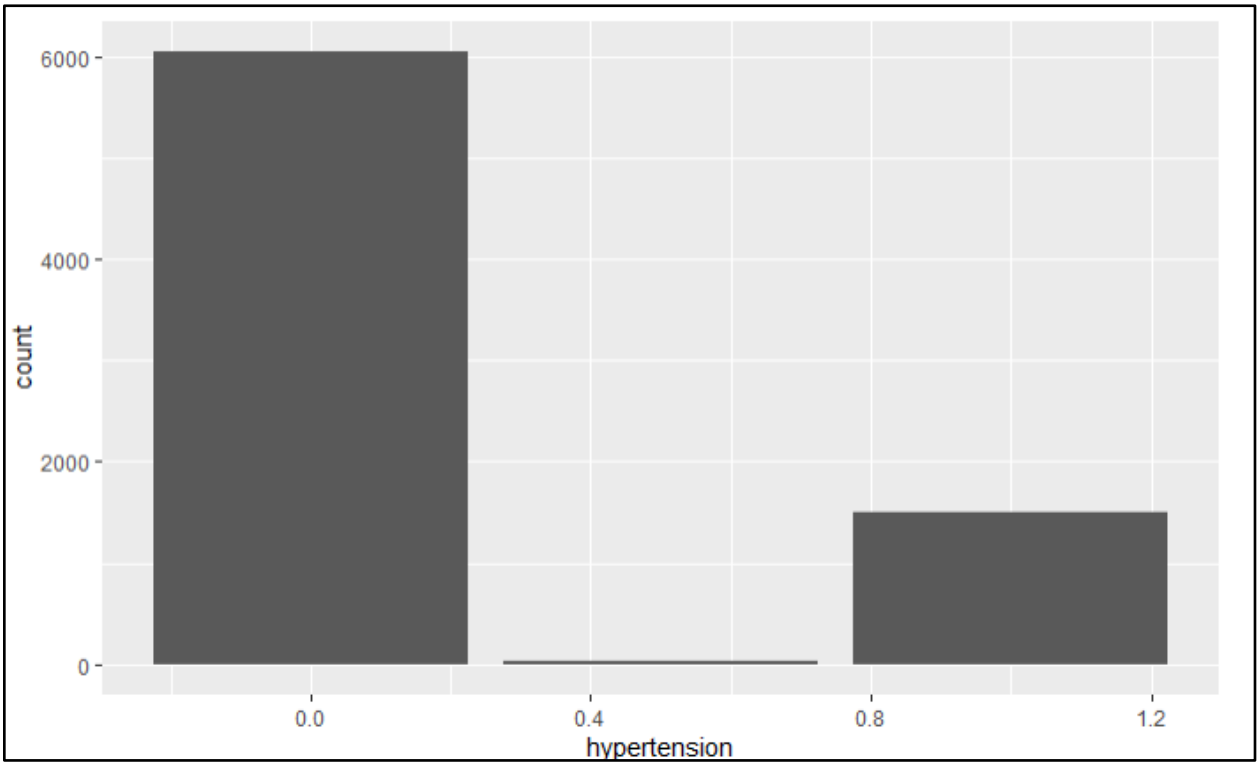
The above graph shows that in the HMO data provided, the customers having 0 children is much higher compared to the customers having one to three children. Whereas, The customers having 4 or more children are the least.

7.1.5 Smoker



From the above analysis we get the count of the number of customers who smoke and those who do not smoke.

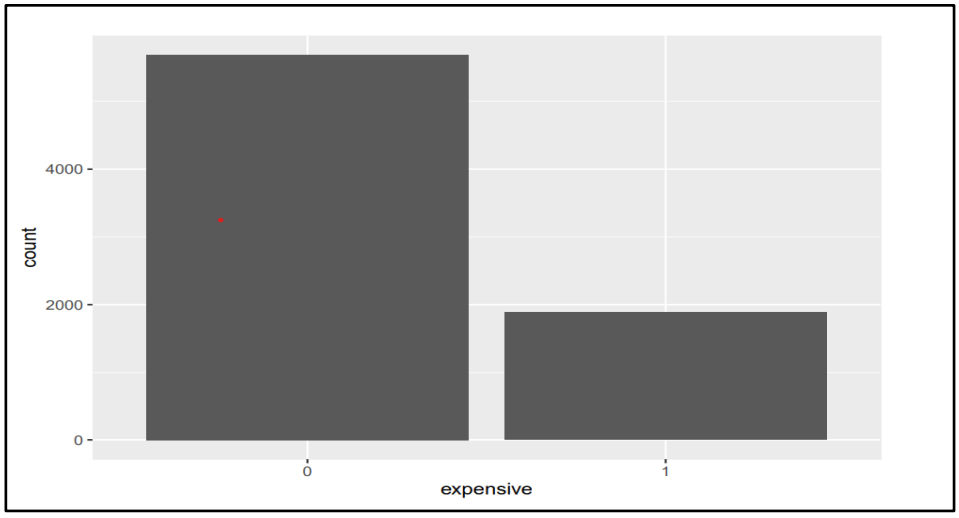
7.1.6 Hypertension



The above plot displays that if the customer's have hypertension it is shown as 1 and those who do not have it is 0.

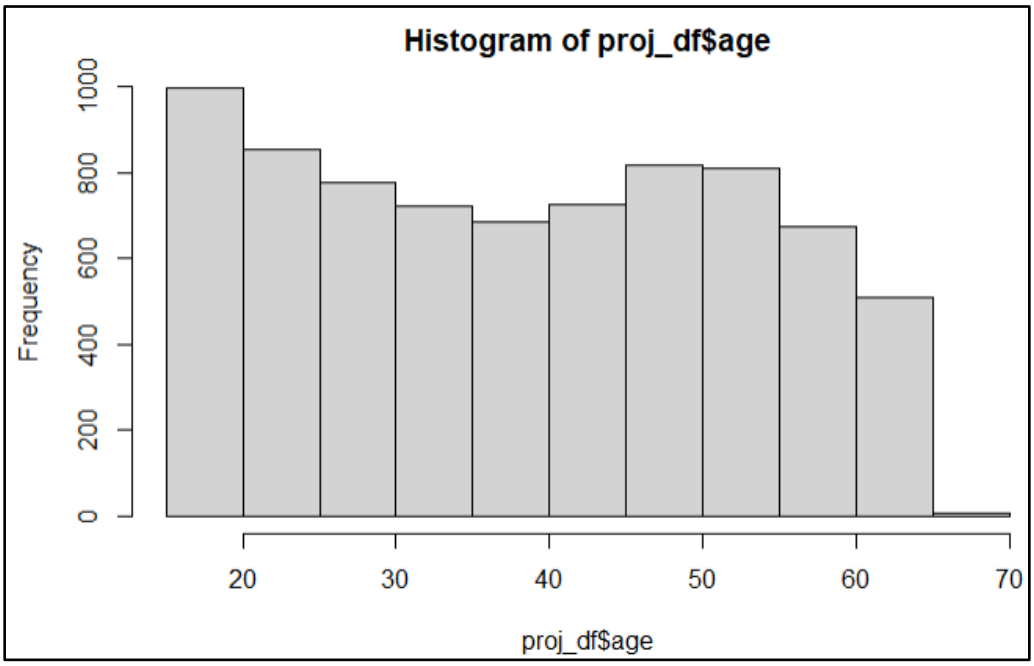
7.1.7 Expensive

The below graph displays the distribution of customers who are expensive. Our analysis of the cost metrics using quartile function revealed that 75% of customers have costs below $4775. Therefore, the members who are above $4775 are categorized as expensive customers.



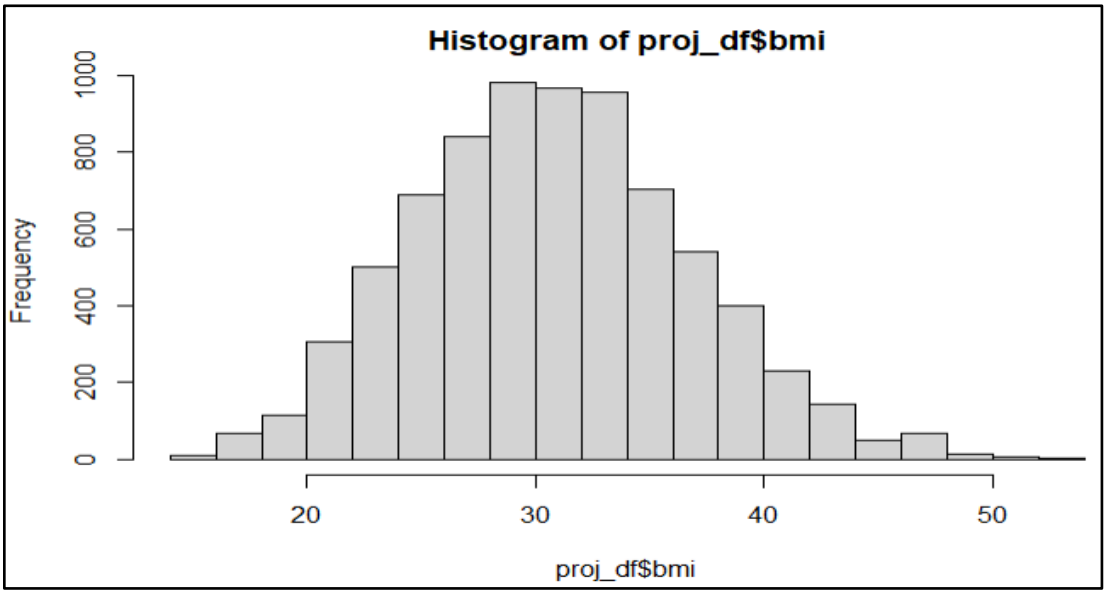
7.2 Histograms .

7.2.1 Histogram of count of age



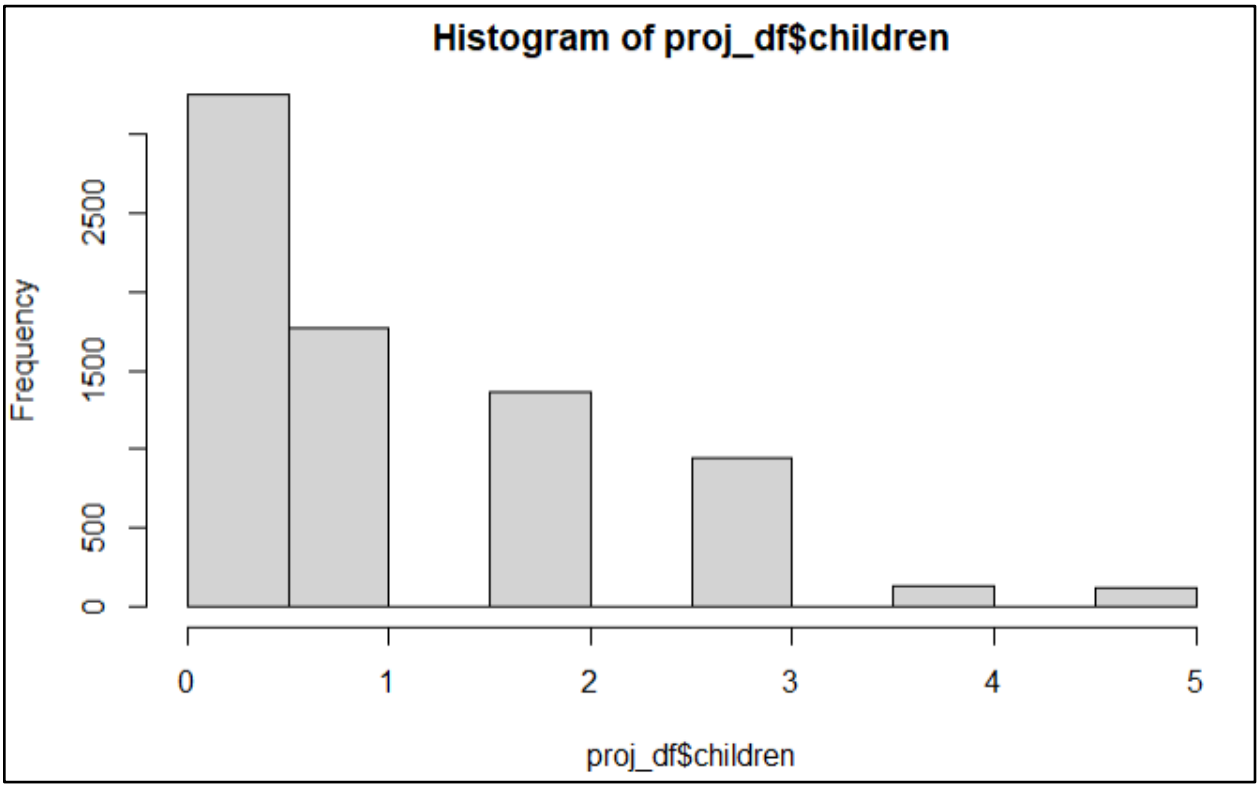
The above graph displays the age distribution from the healthcare data.

7.2.2 Histogram of count of BMI



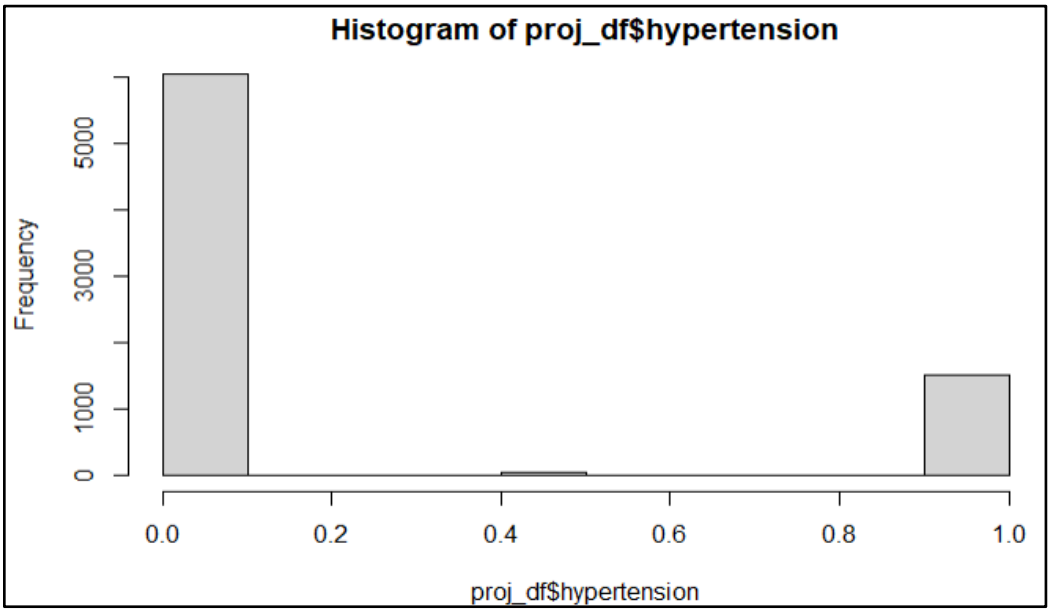
The above histogram displays the BMI distribution of customers from the healthcare data.

7.2.3 Histogram of count of Children



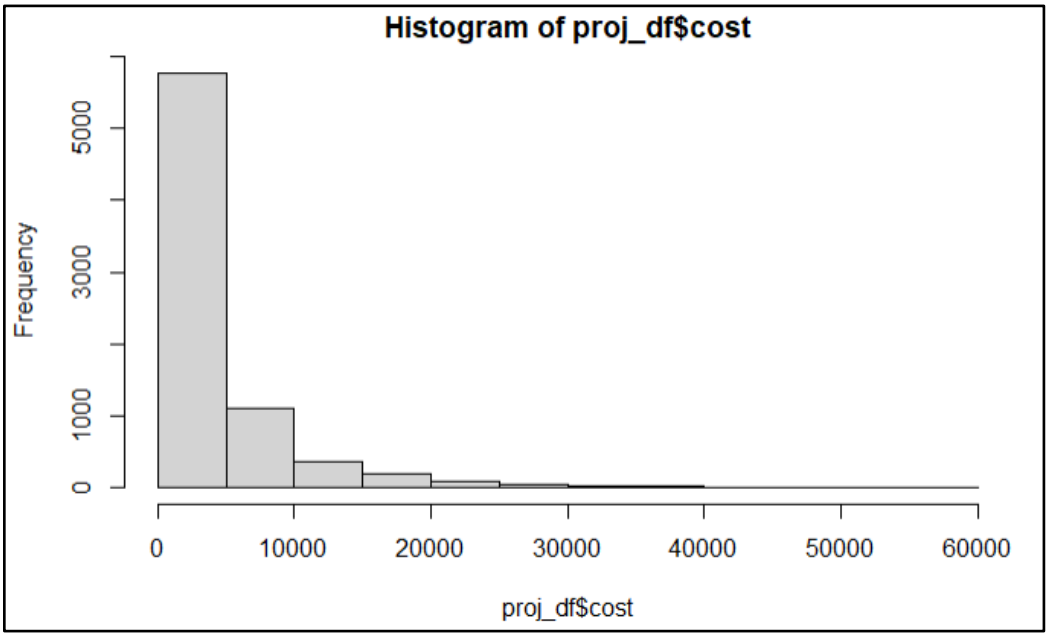
According to the histogram, there are much more customers with 0 children than there are with one to three children in the HMO statistics. The least number of customers have four or more children.

7.2.4 Histogram of count of Hypertension



According to the histogram plot, customers who have hypertension are indicated as 1 and those who do not as 0.

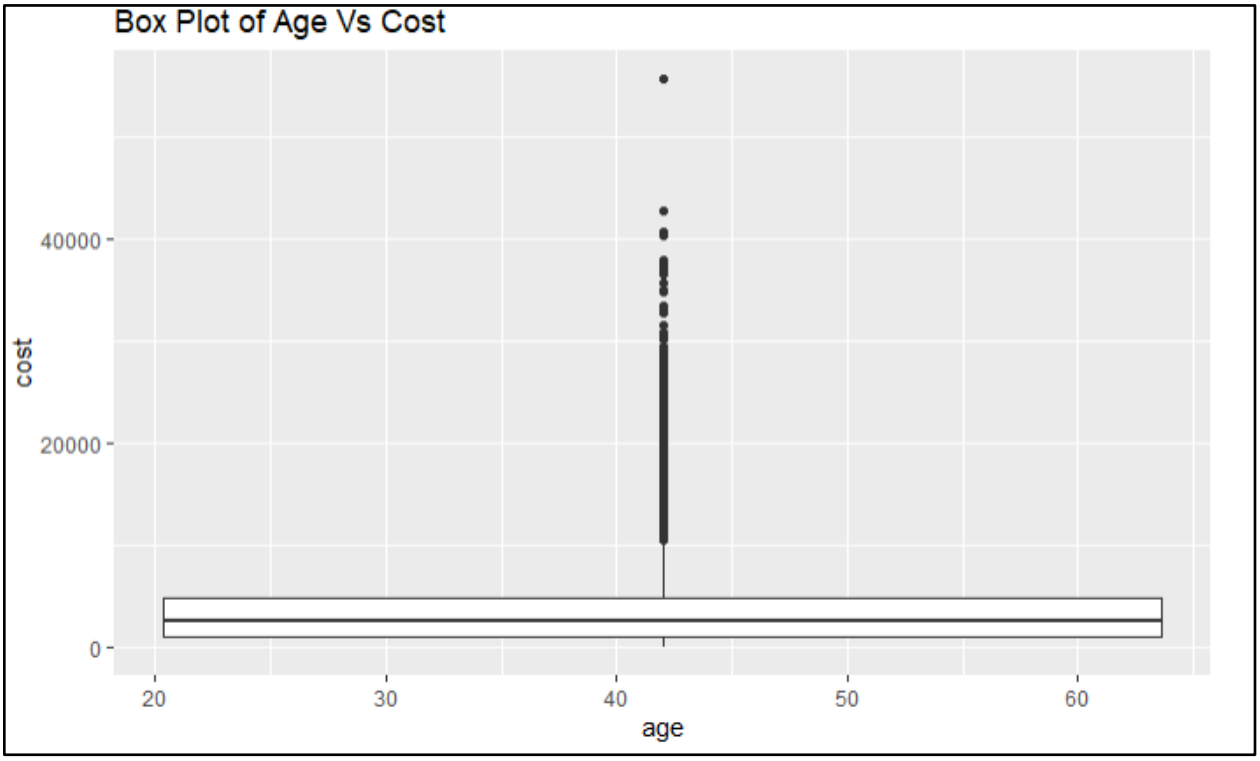
7.2.5 Histogram of count of Cost



The above histogram displays the cost distribution from the healthcare data. Majority of the cost lies between 0 and 10000. There are a few costs up to 40000.

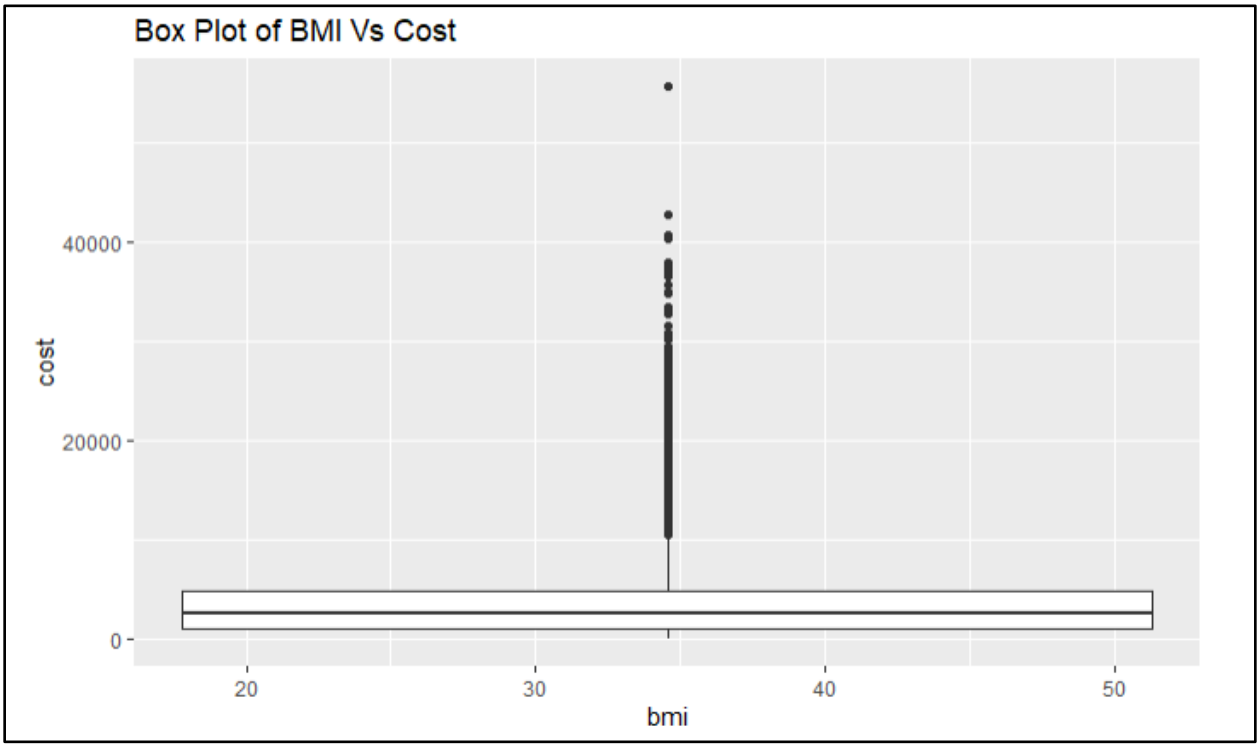
7.3 Box Plots

7.3.1 Boxplot of count of Age



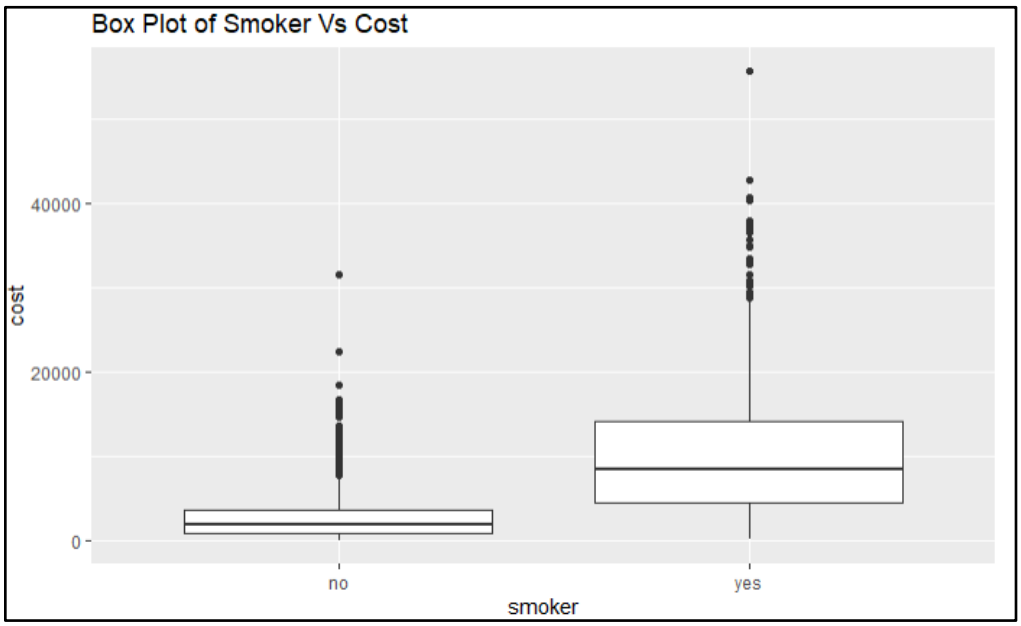
We can see from the above box plot that the average cost for the age distribution is around 4000-5000. And it shows that after the age of 40, there is a steep rise in costs.

7.3.2 Boxplot of count of BMI



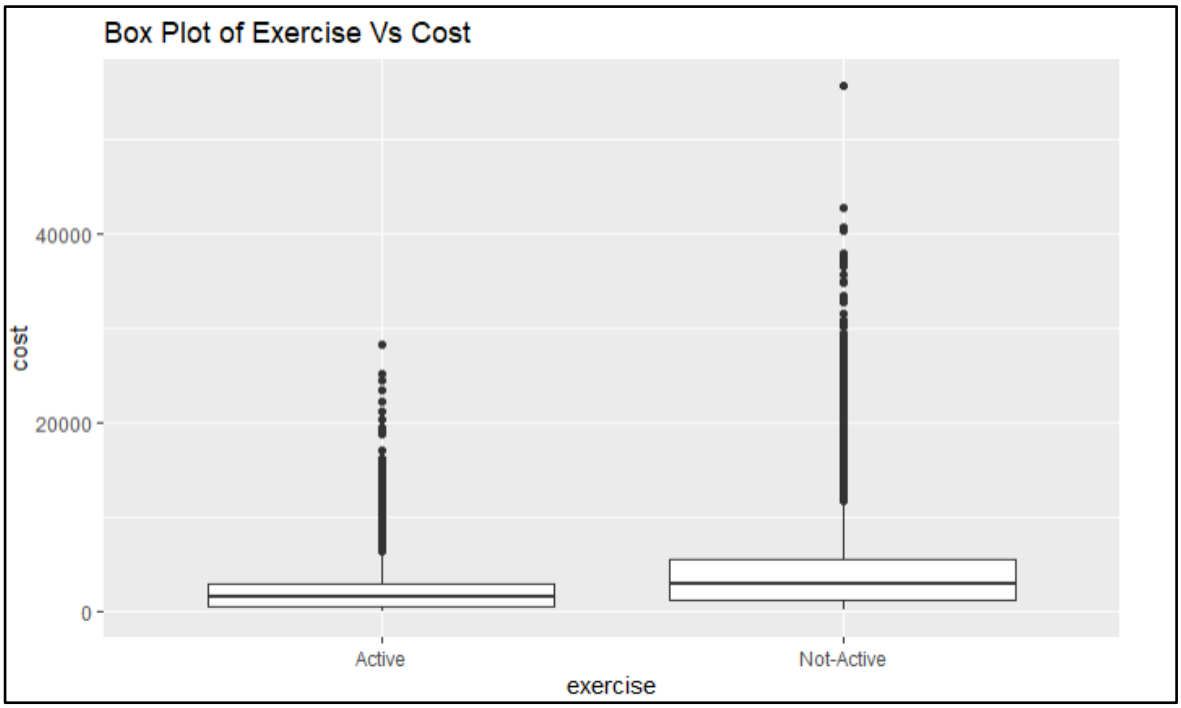
The above boxplot shows the relation of bmi and cost. The bmi ranging from 10 to 50 and the mean cost for them between 5000-10000.

7.3.3 Boxplot of distributions of smokers



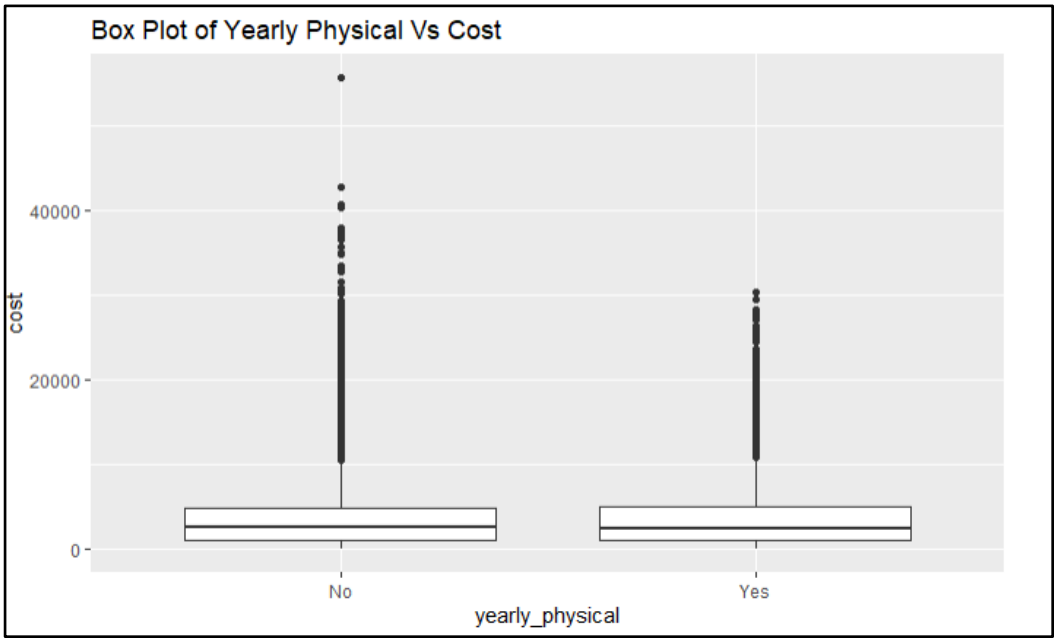
The above box plot shows us the comparison between smokers and non smokers and also compares their costs. We can clearly see that the average cost for smokers is much higher than that of non-smokers. Also the smokers range is wider than non smokers.

7.3.4 Boxplot for distributions of customers that exercise



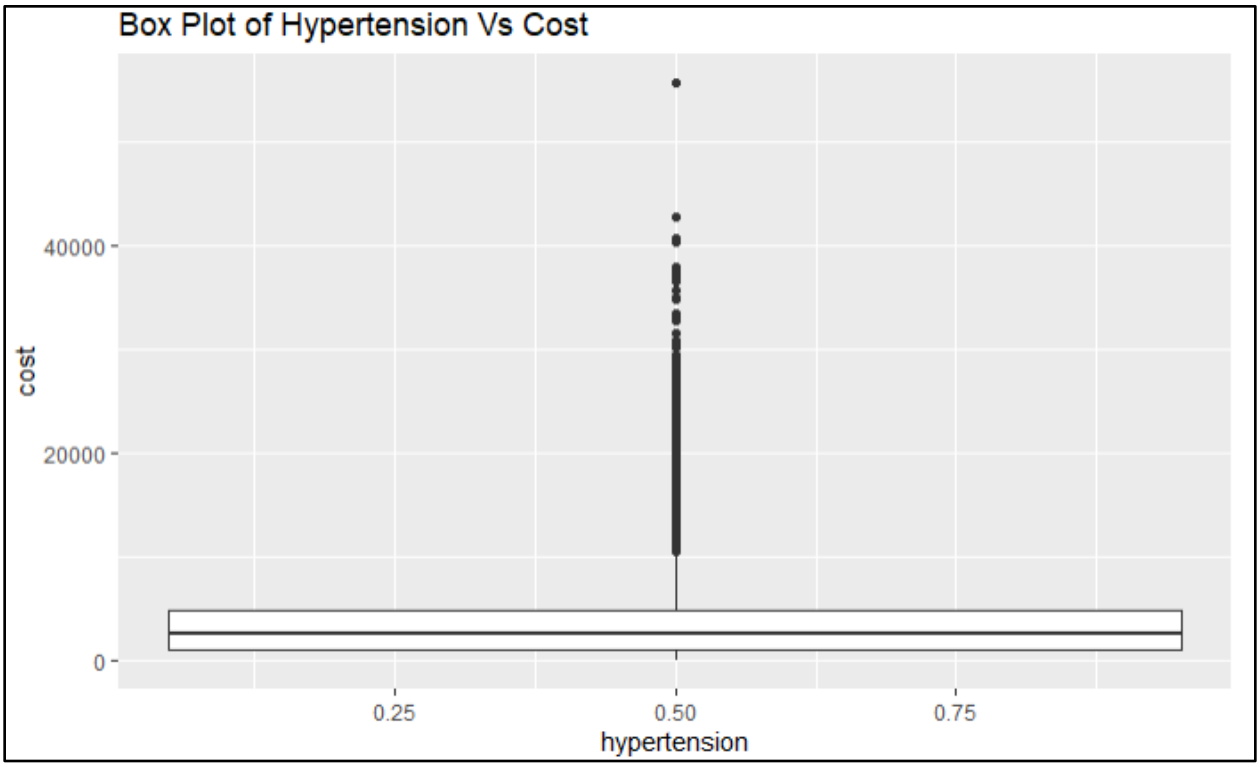
Here we can see that the cost for non active people us higher that that for the active people

7.3.5 Boxplot for Yearly Physical on weekends



From the above boxplot we can tell the cost for people who are physically active yearly and those who are not physically active yearly.

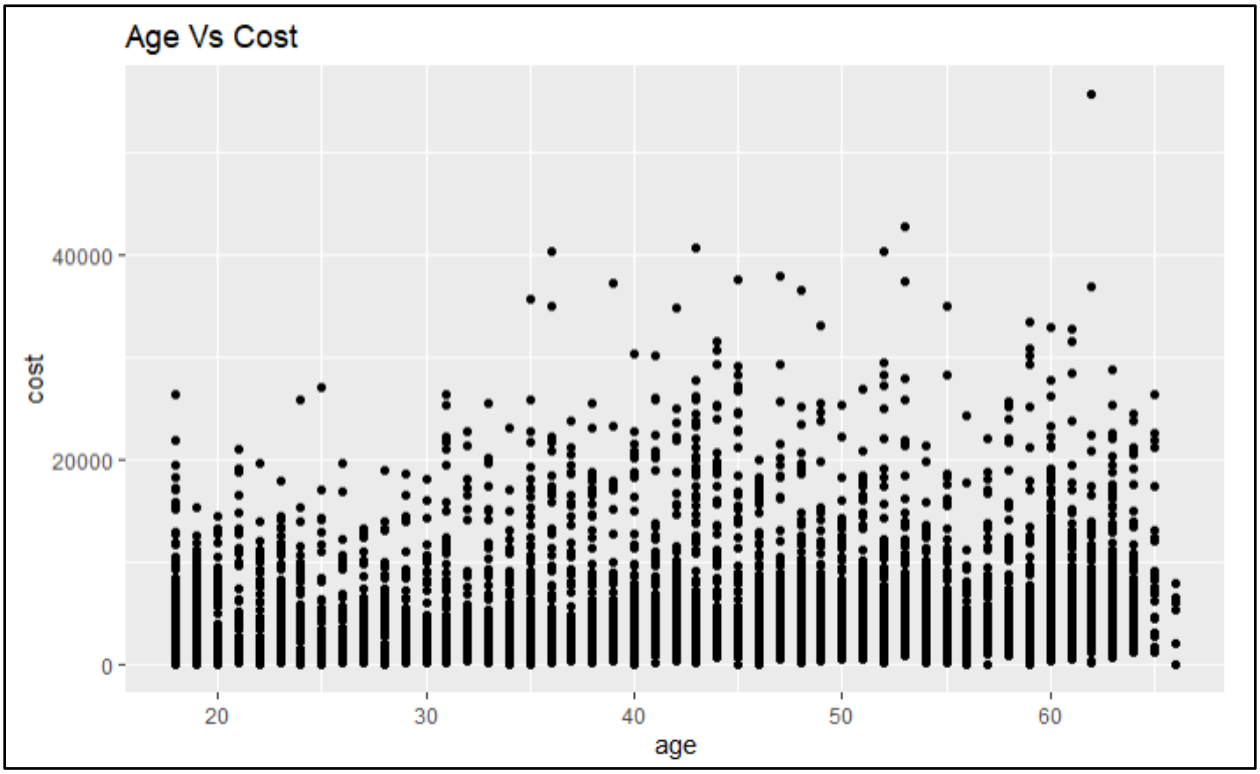
7.3.6 Boxplot for customers with Hypertension



Here we see the comparison between hypertension and cost. If the people have hypertension or not (0 or 1), the average cost is around 3000-5000.

7.4 Scatter Plots

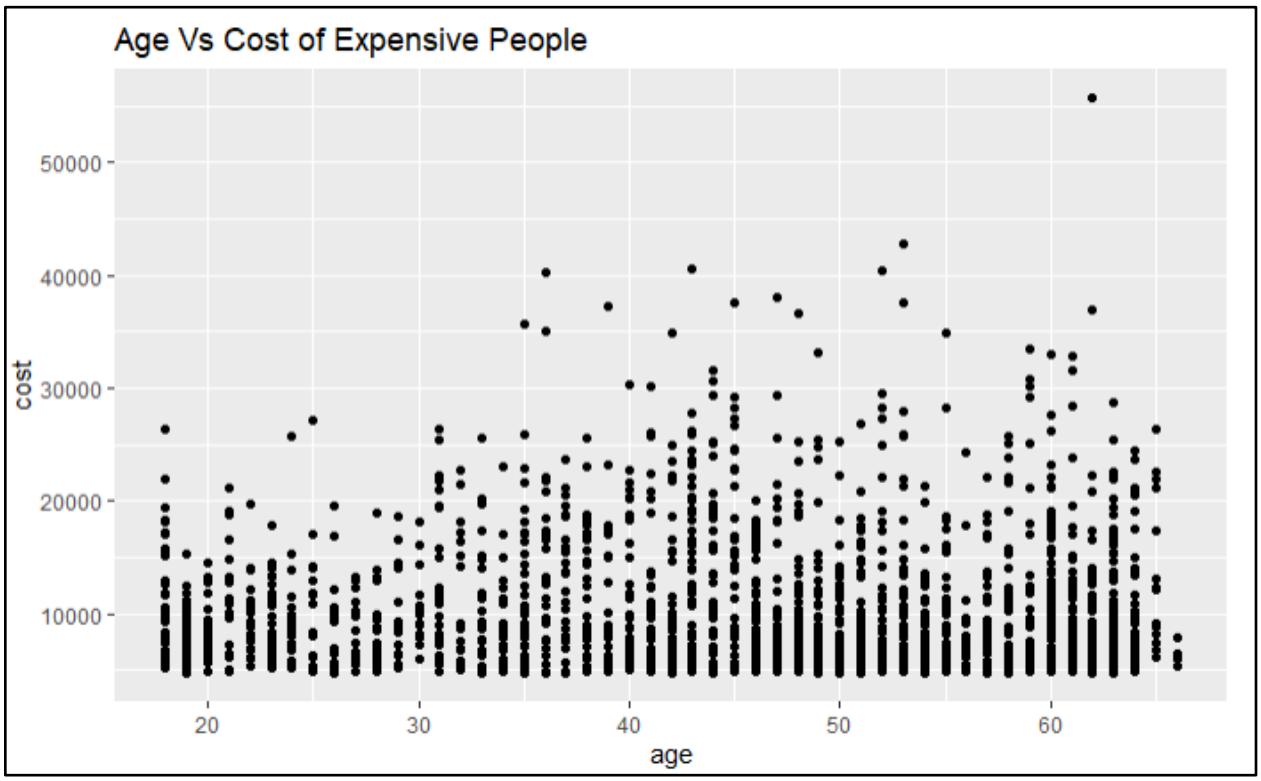
7.4.1 Scatter Plot for Age vs Cost



The above scatter plot depicts the distribution of age vs cost.

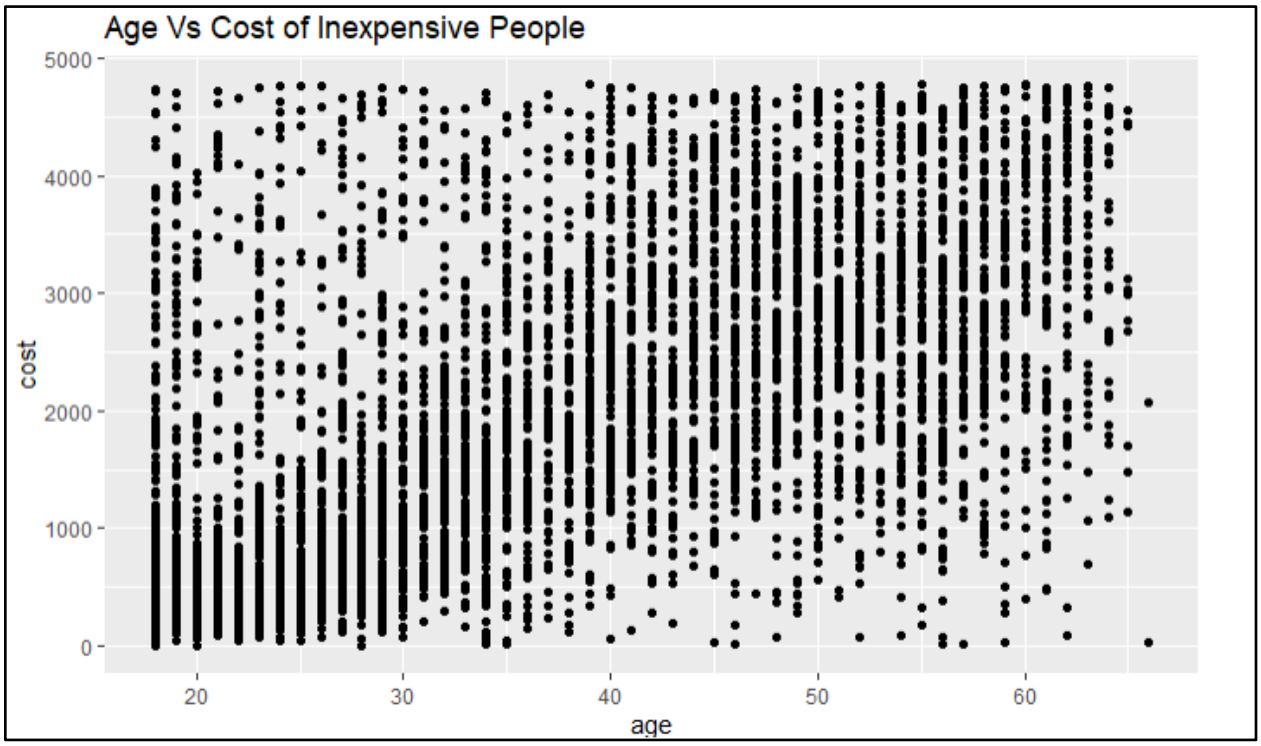
The above scatter plot displays the distribution of members who are smokers with respect to their age and expenses and we can find out that people above the mean age of 35 years have higher cost of healthcare and are smokers as well. This plot gives us the relationship between age and cost.

7.4.2 Scatter Plot for Age vs Expensive



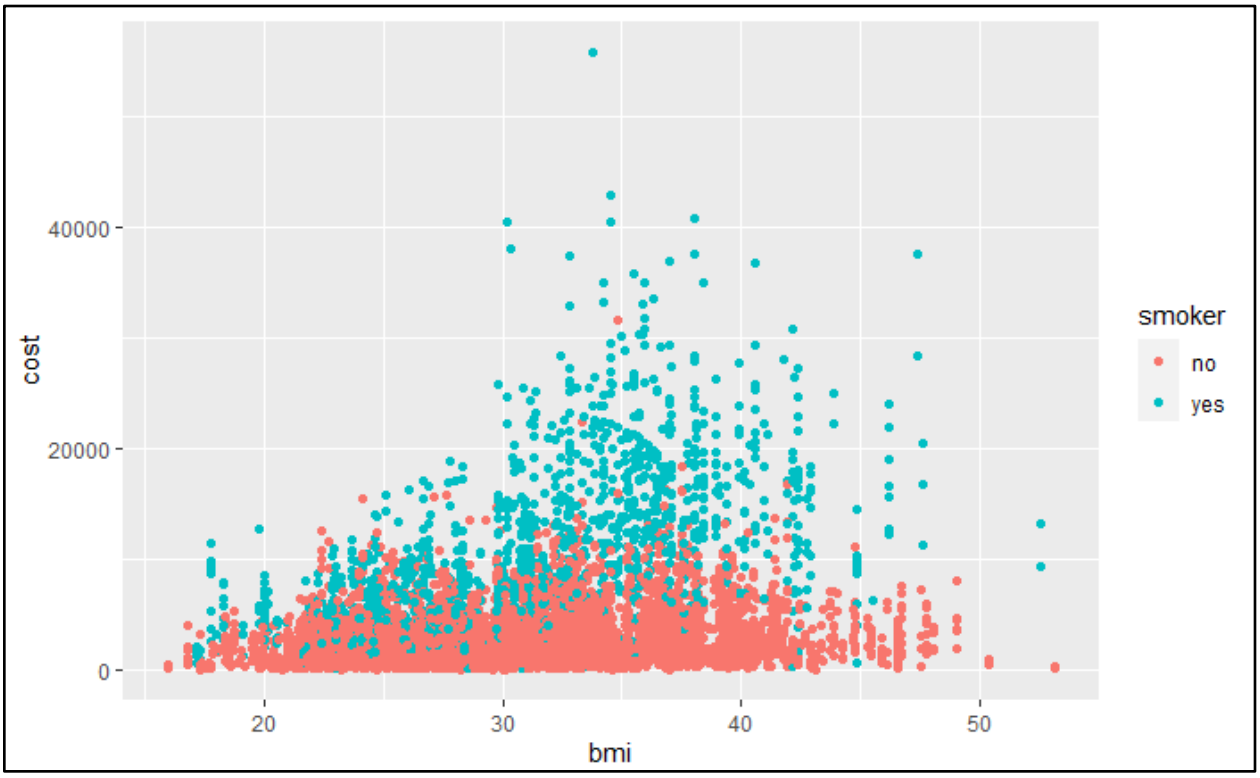
The above scatter plot displays that the data is concentrated more as the age increases the mean value of 35 for expensive customers.

7.4.3 Scatter Plot for Age vs Inexpensive

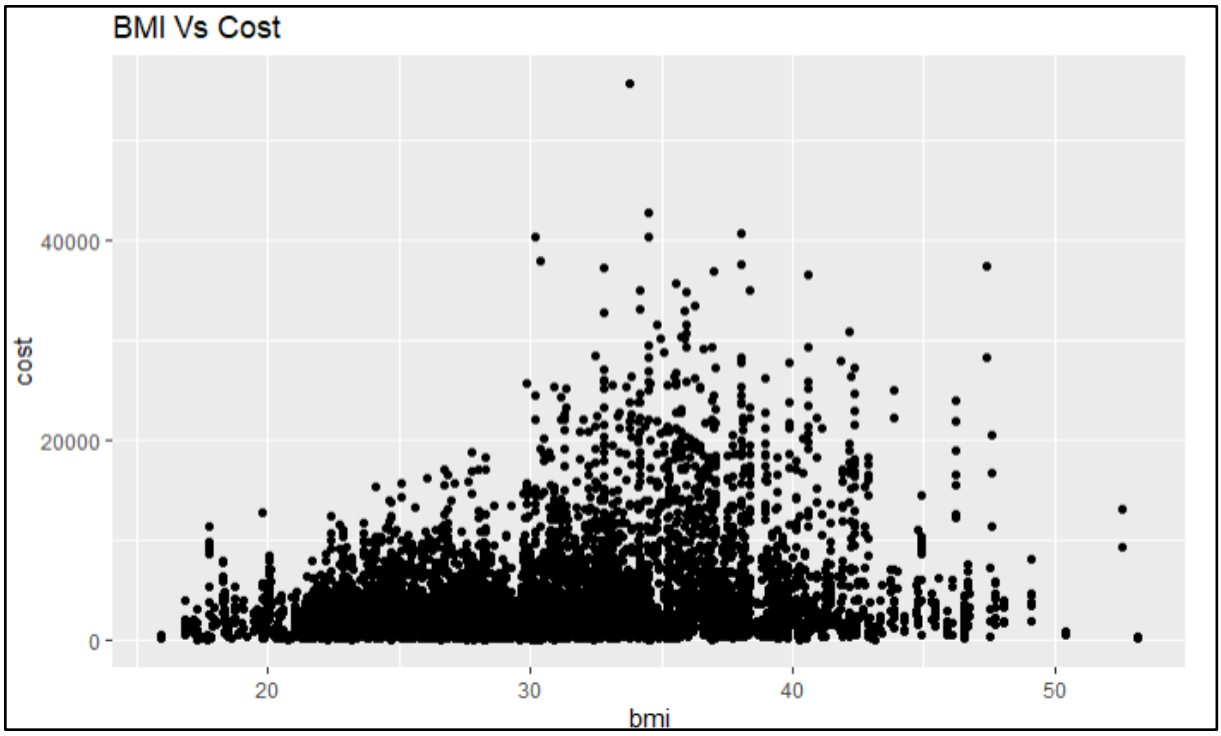


The above scatter plot displays that the data is less concentrated as the age increases the mean value below 35 for inexpensive customers and is between 0$ - 5000$

7.4.4 Scatter Plot for BMI vs Cost

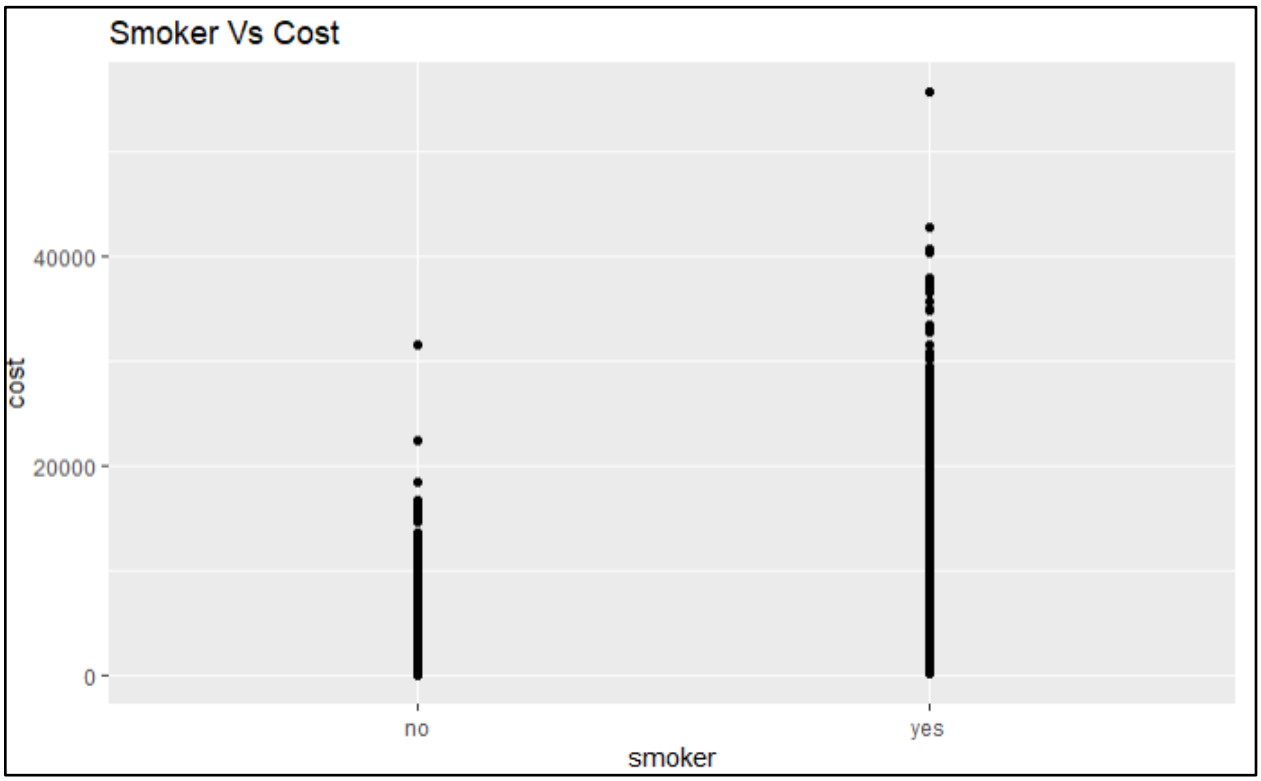


The above scatter plot displays the distribution of members who are smokers with respect to their BMI and expenses and we can find out that people above the mean age of 32 have higher cost of healthcare and are smokers as well. This plot gives us the relationship between BMI and cost.



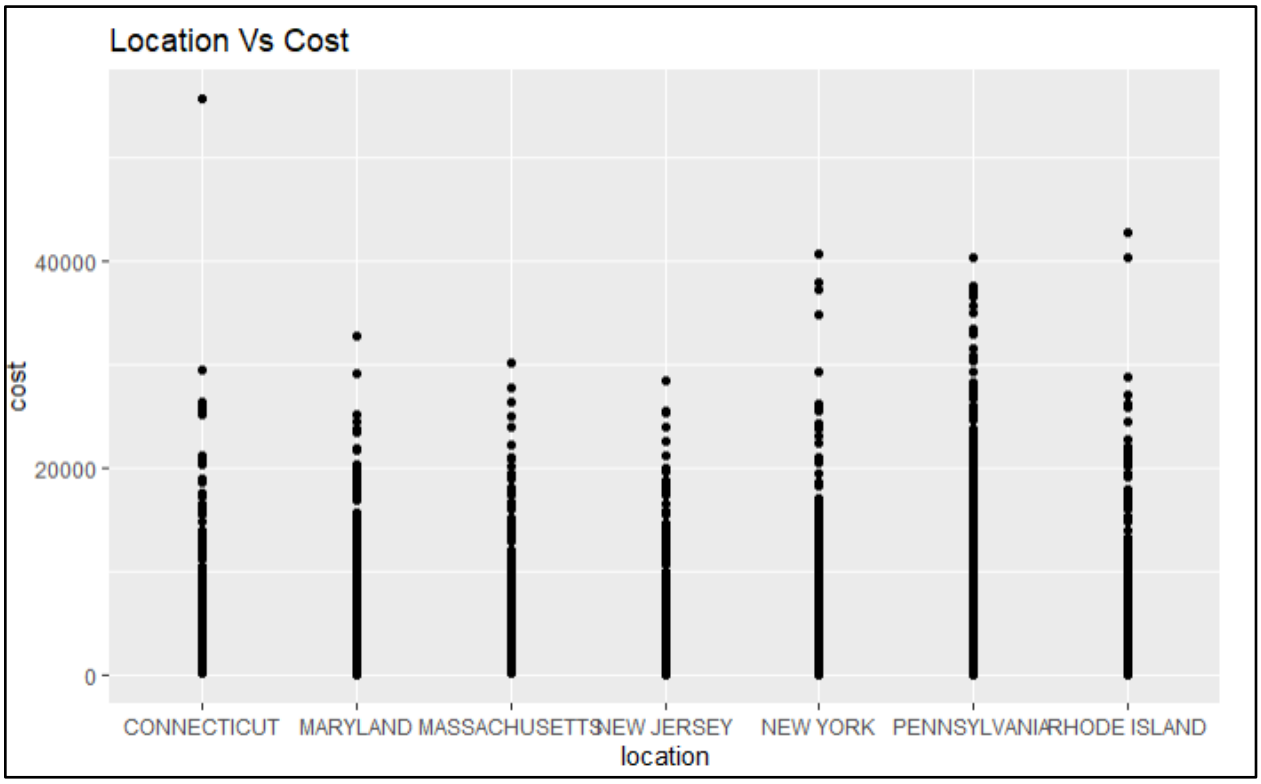
The above scatter plot displays that the health care cost is more expensive for customers as the age increases the mean value above 30.

7.4.5 Scatter Plot for Smoker vs Cost



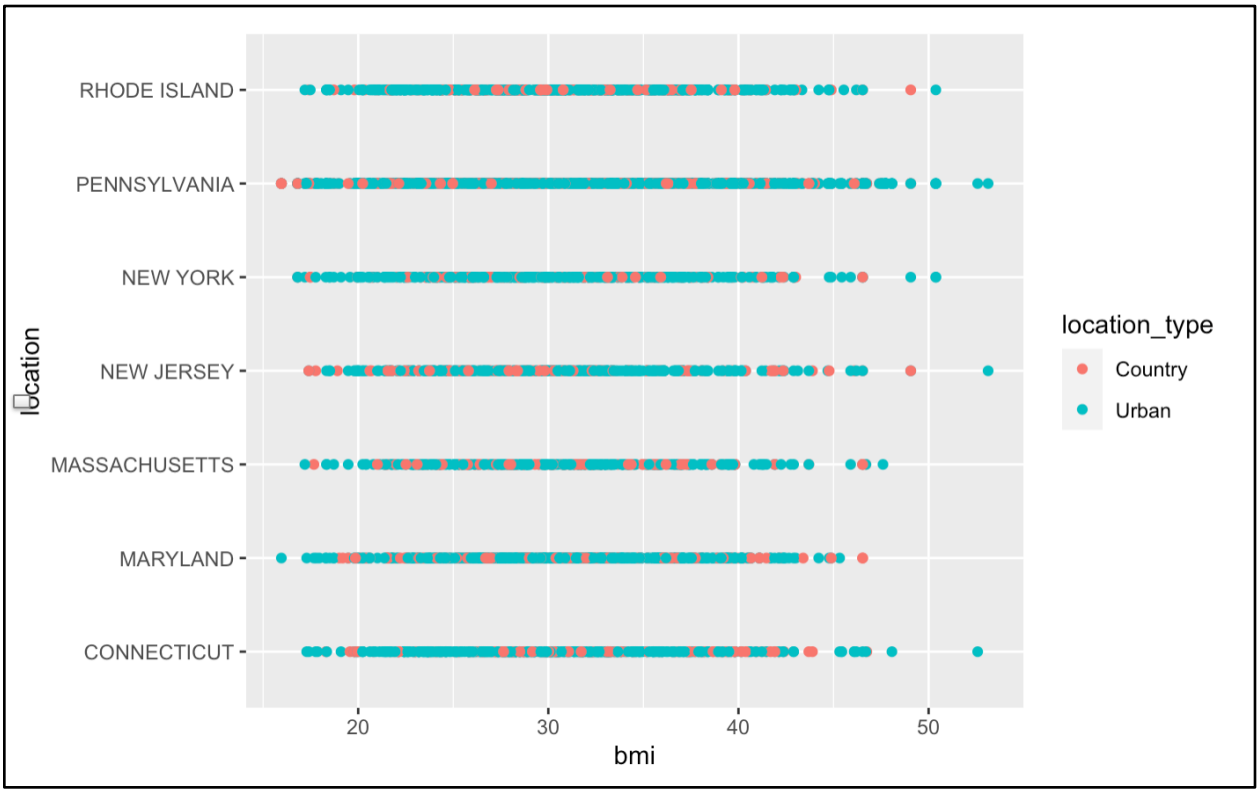
The above scatter plot displays that the health care cost is much higher for smokers compared to the customers who do not smoke.

7.4.6 Scatter Plot for Location vs Cost

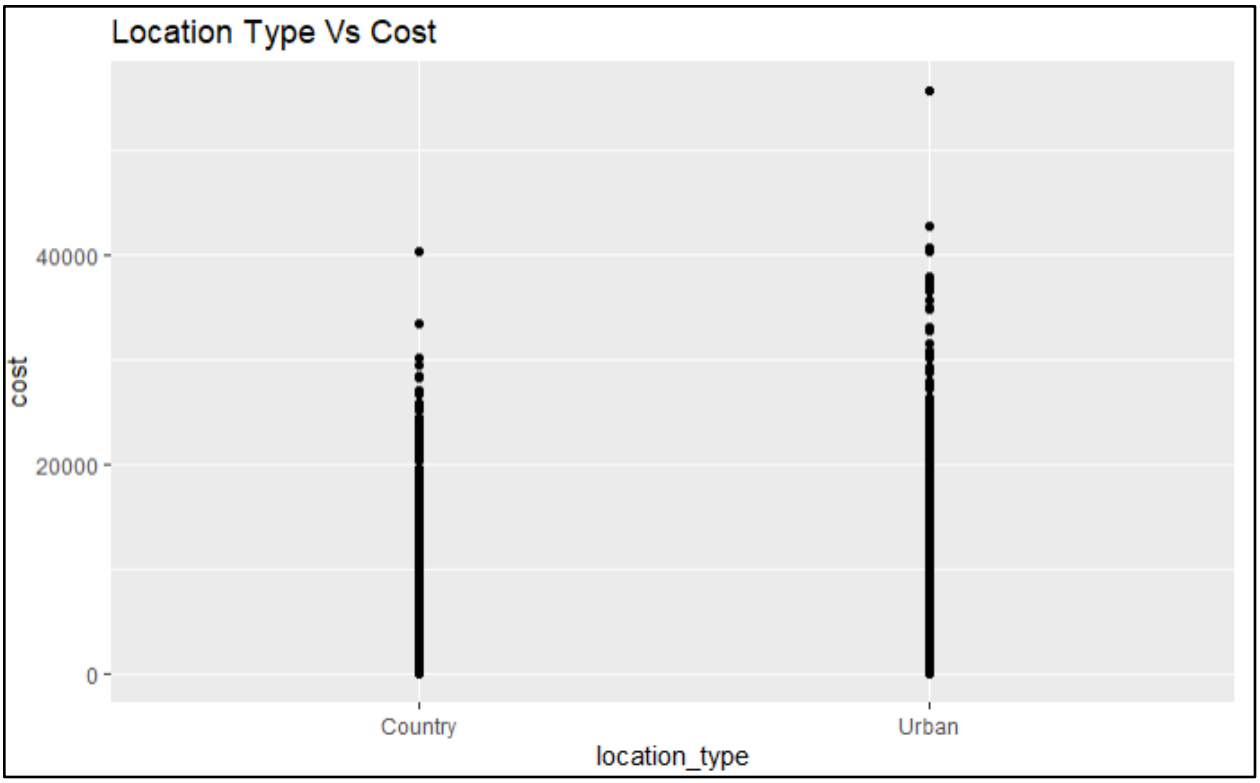


The above scatter plot displays that the health care cost in Connecticut, New York, Pennsylvania and Rhode Island is much higher compared to the other remaining states.

7.4.7 Scatter Plot for Location Type vs Cost

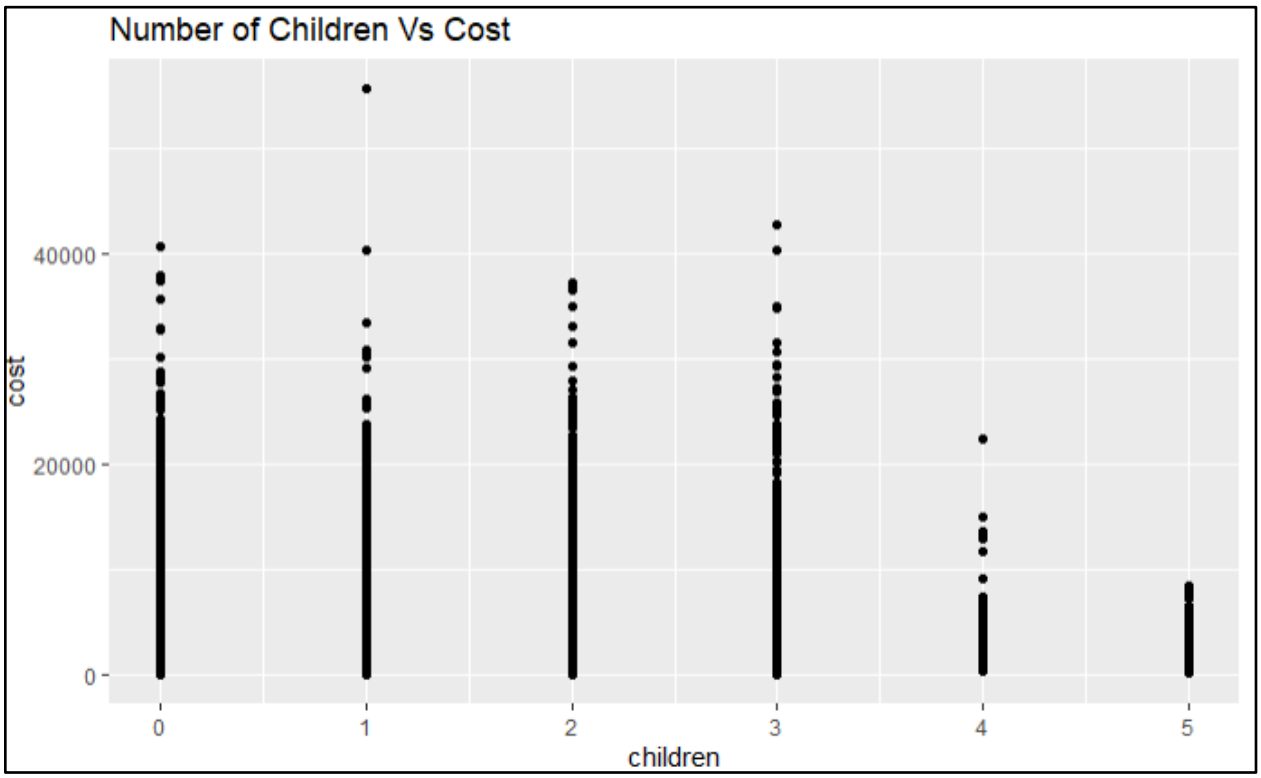


The above scatter plot displays the distribution of members with respect to their location_type and BMI. Much cannot be anticipated based on location_type because of inconsistencies.



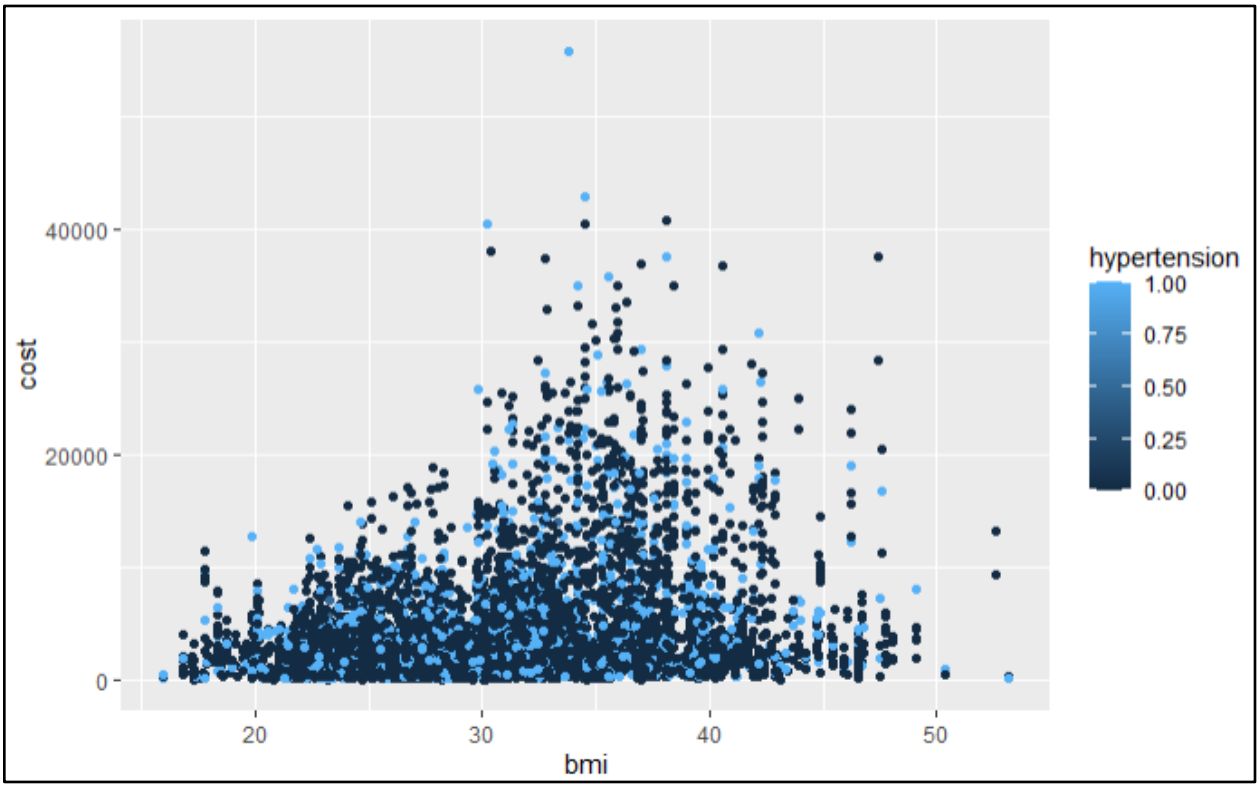
In the above scatter plot, it is clearly visible that the location type is densely populated by urban people compared to the countryside.

7.4.8 Scatter Plot for Number of Children vs Cost

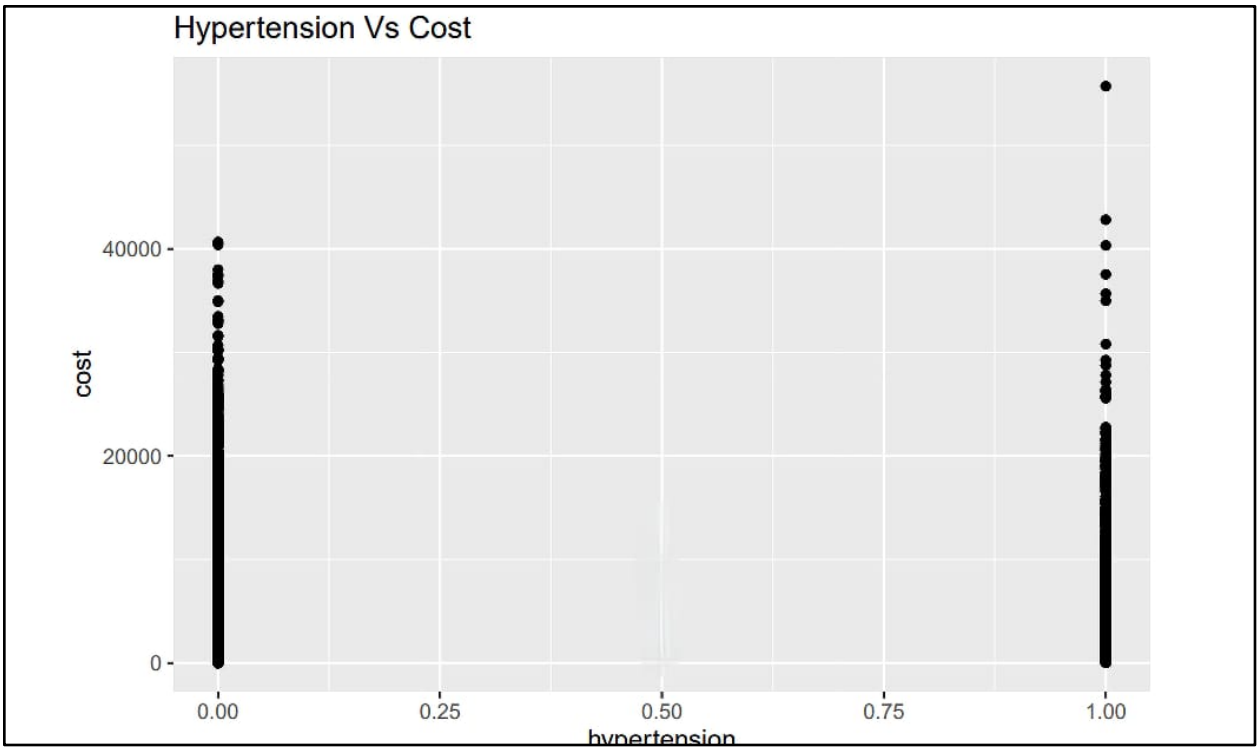


The above scatter plot suggests that the healthcare cost of customers having 0 to 3 children is higher compared to the customers having 4 or more children.

7.4.9 Scatter Plot for Hypertension vs Cost

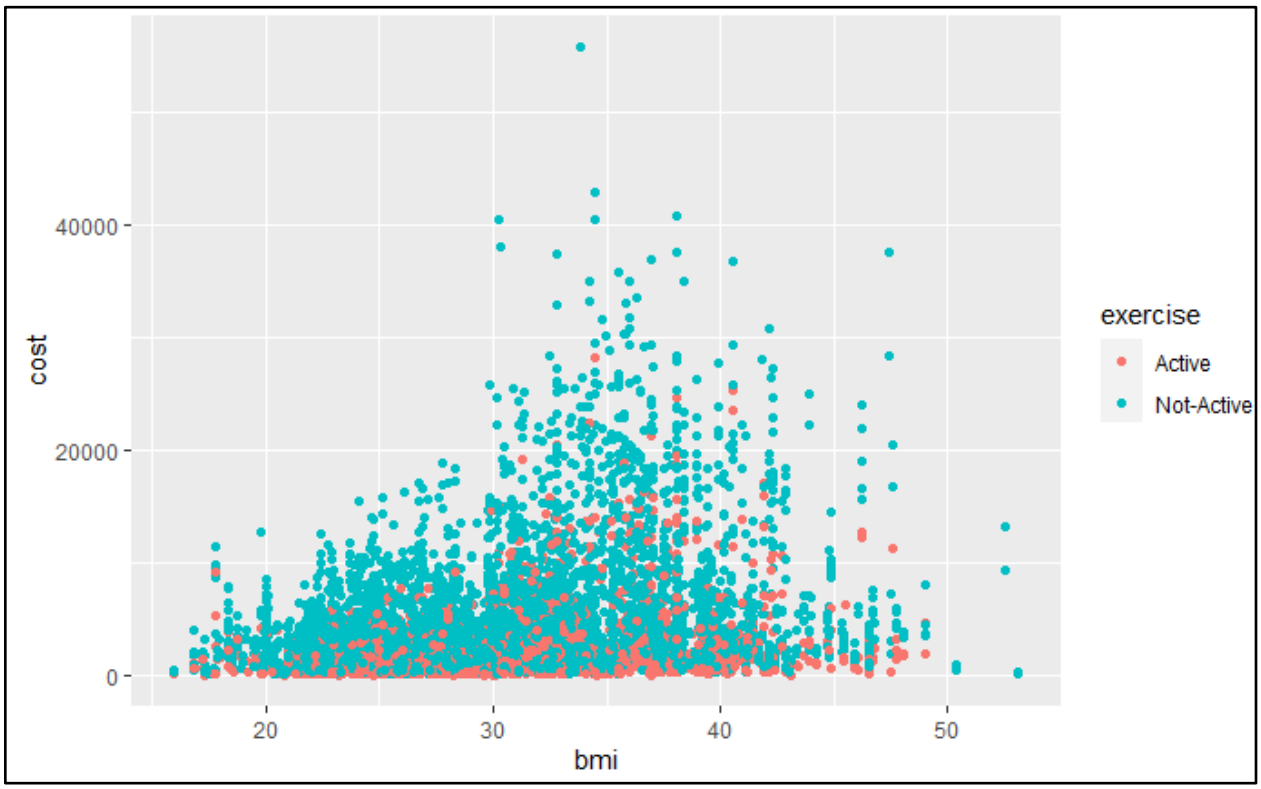


The above scatter plot displays the distribution of members who have hypertension with respect to BMI and Healthcare Costs and we can find out that people above the mean BMI age of 32 have higher healthcare costs and have hypertension as well. This plot gives us the relationship between BMI and cost.

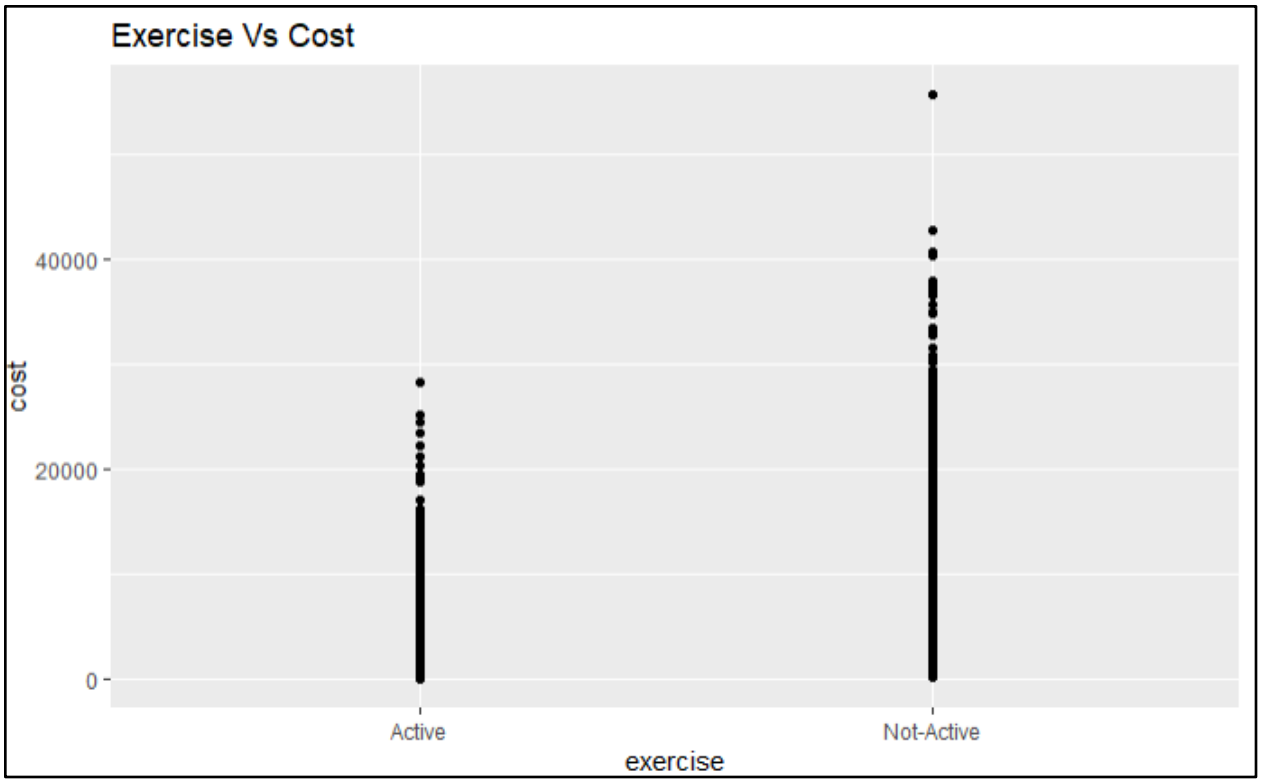


The above scatter plot suggests that the healthcare cost of customers who have hypertension is expensive.

7.4.10 Scatter Plot for Exercise vs Cost

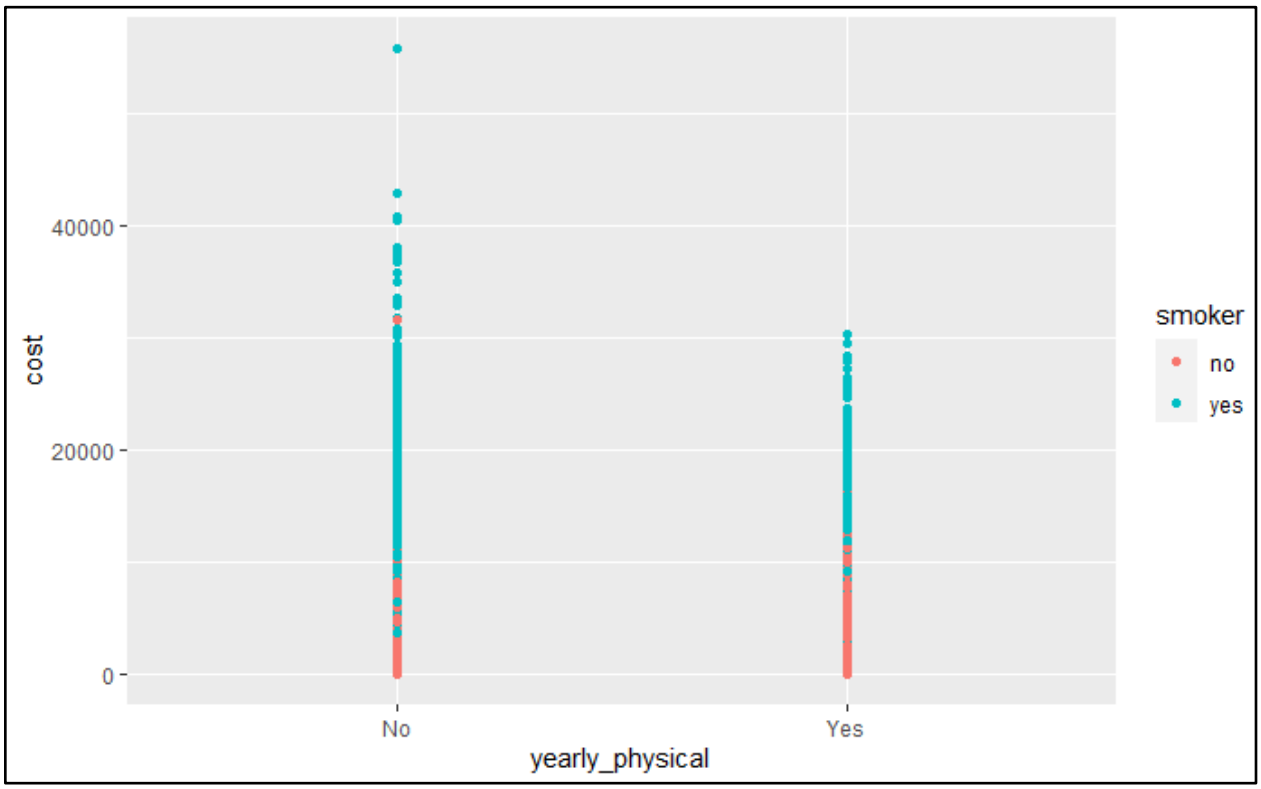


The above scatter plot displays the distribution of members who exercise with respect to BMI and Healthcare Costs and we can find out that people above the mean BMI value of 32 have higher healthcare costs and do not exercise as well. This plot gives us the relationship between BMI and cost.



The above plot suggests that the healthcare costs of customers who are not active when it comes to exercising are expensive compared to the customers who do not exercise.

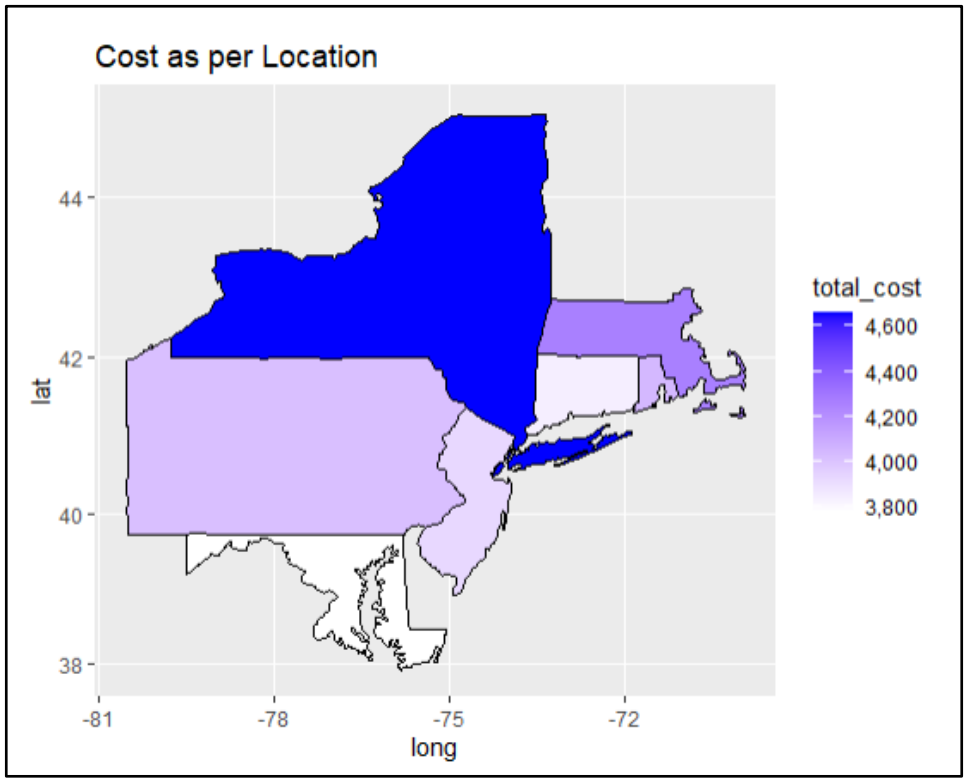
7.4.11 Scatter Plot for Yearly_physical vs Cost



The above scatter plot displays the distribution of members who are smokers with respect to their yearly physical activity and healthcare costs and we can find out that the members who are smokers and do not have any physical activity tend to have the higher healthcare costs.

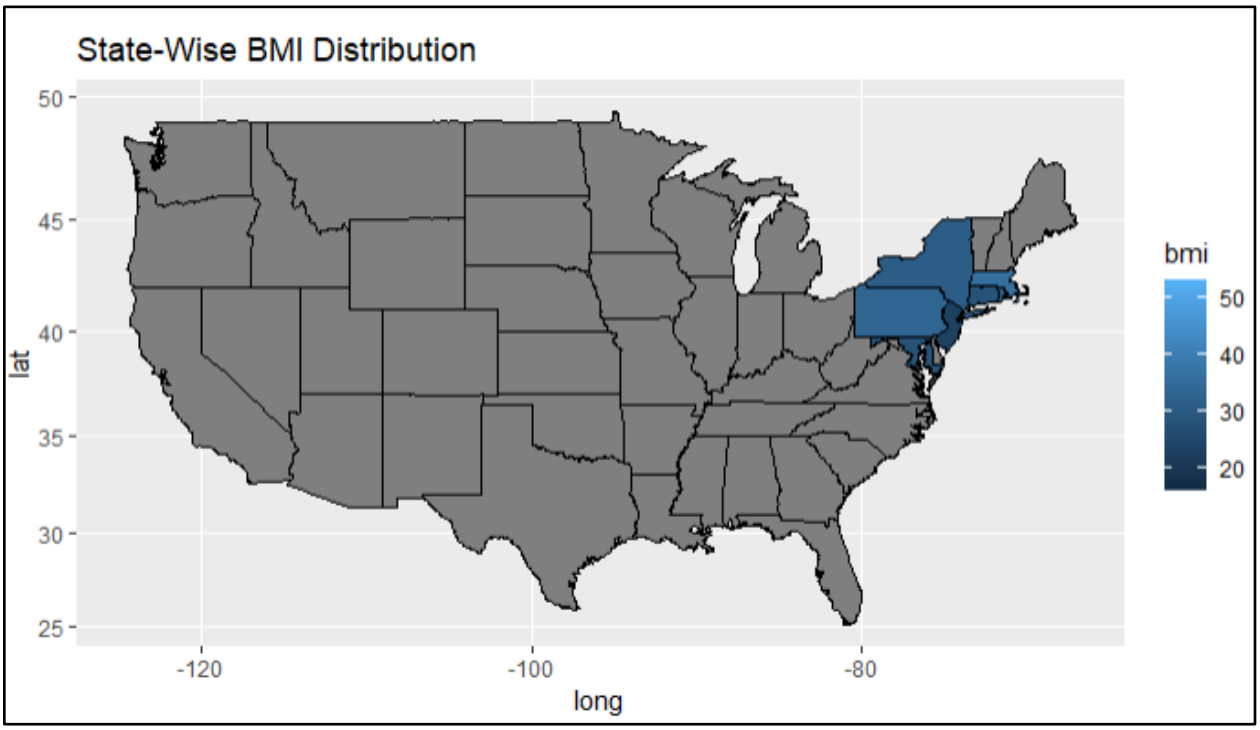
8.Geographical Visualization

Map 1



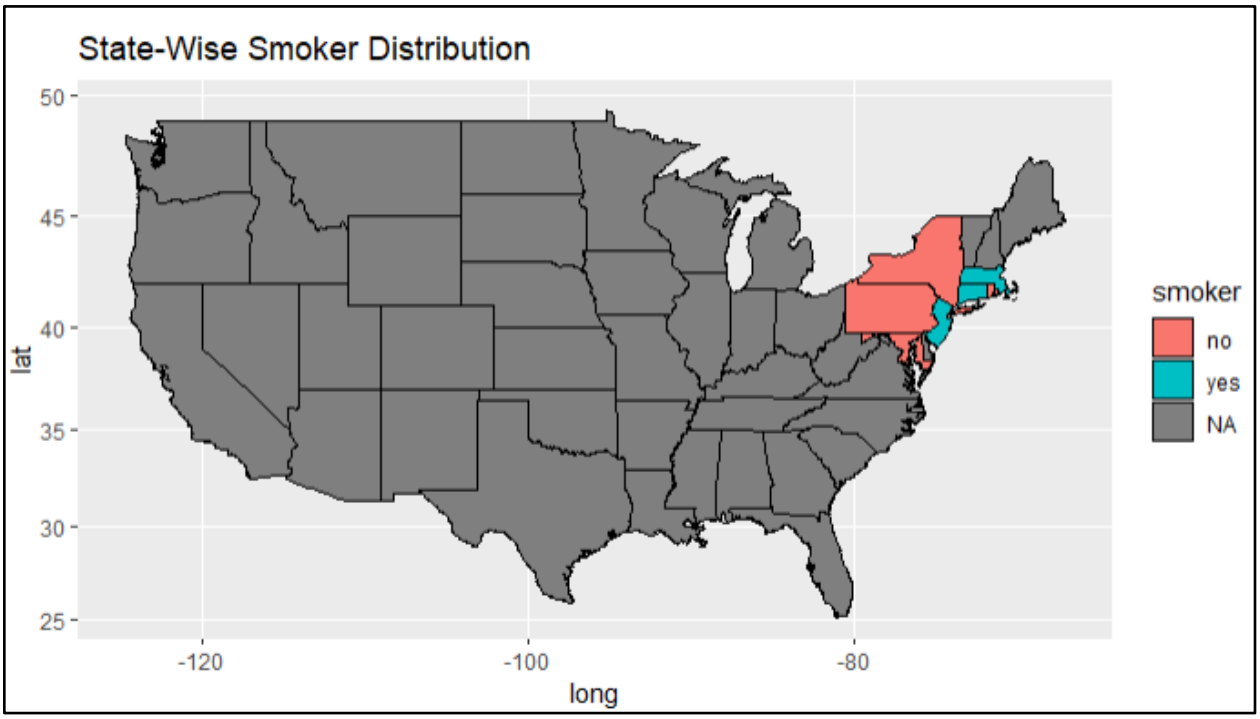
The distribution of patient total healthcare costs in the Northeastern United States region is shown on the map above.

Map 2



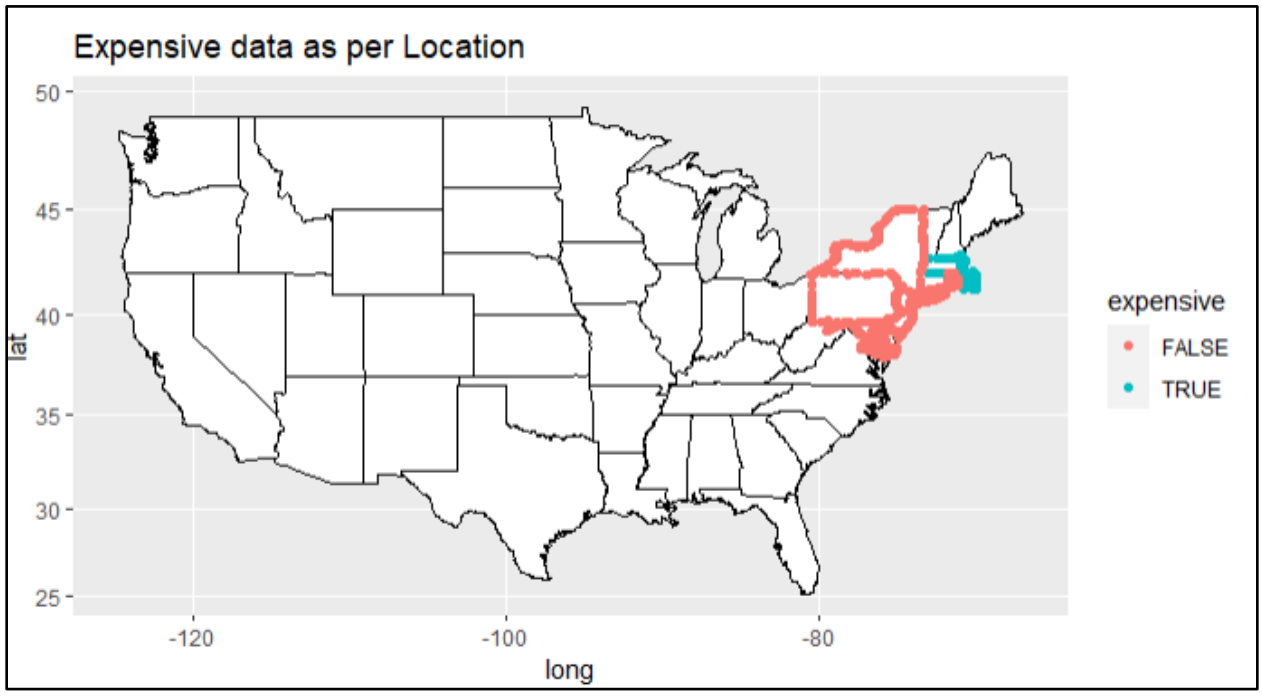
This particular map displays the Distribution of HMO Members based on their BMI value.

Map 3



The distribution of states according to their population of smokers is shown on this map.

Map 4



The region's distribution in terms of expensive and inexpensive members is shown on this map. Our analysis of the cost metrics using quartile function revealed that 75% of customers have costs below $4775. Therefore, the members who are above $ 4775 are categorized as expensive customers.

9. Models

Multiple Regression Model

````
1. Multiple regression model
```{r}
mrLmOut <- lm(expensive ~ age+bmi+hypertension+smoker+exercise,proj_df)
summary(mrLmOut)
```
````

```
Call:
lm(formula = expensive ~ age + bmi + hypertension + smoker + 
    exercise, data = proj_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9337 -0.2020 -0.0588  0.1293  1.1509 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -0.6786414  0.0227758 -29.797  < 2e-16 ***
age                 0.0074486  0.0002674  27.853  < 2e-16 ***
bmi                 0.0126240  0.0006344  19.899  < 2e-16 ***
hypertension        0.0352105  0.0094574   3.723 0.000198 ***
smokeryes           0.5966752  0.0095265  62.633  < 2e-16 ***
exerciseNot-Active  0.1687156  0.0087288  19.329  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3286 on 7576 degrees of freedom
Multiple R-squared:  0.4245,    Adjusted R-squared:  0.4241 
F-statistic:  1118 on 5 and 7576 DF,  p-value: < 2.2e-16
```

In this Linear regression model the Adjusted R-squared value is low which translates to low accuracy which is around 42.41% in this case. Even though the Intercept and the variable itself is significant the accuracy is still low. This issue is caused mainly because of the huge amount of data set. For this model we have used the important variables which had an impact on cost based on our exploratory analysis the variables are age, BMI, Hypertension, smoker and exercise.

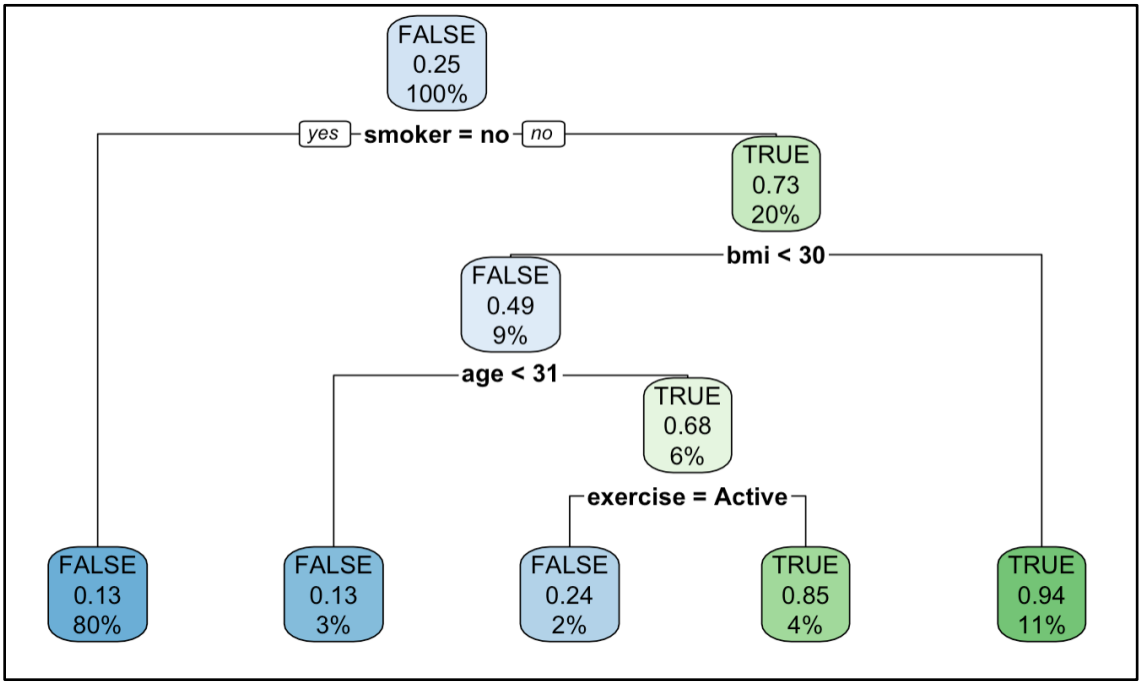
SVM Model

```
Length  Class   Mode
     1   ksvm     S4
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  2198  310
     TRUE     76  448

               Accuracy : 0.8727
                 95% CI : (0.8603, 0.8844)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6216

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

            Sensitivity : 0.9666
            Specificity : 0.5910
         Pos Pred Value : 0.8764
         Neg Pred Value : 0.8550
             Prevalence : 0.7500
         Detection Rate : 0.7249
   Detection Prevalence : 0.8272
      Balanced Accuracy : 0.7788

       'Positive' Class : FALSE
```

In the above SVM model 60% of the data in the data set was used for training and rest for testing. The accuracy of the model is 87.27% which is higher than the Multiple regression model. Also the sensitivity of the model is 96.66%. This makes the model very good for prediction.

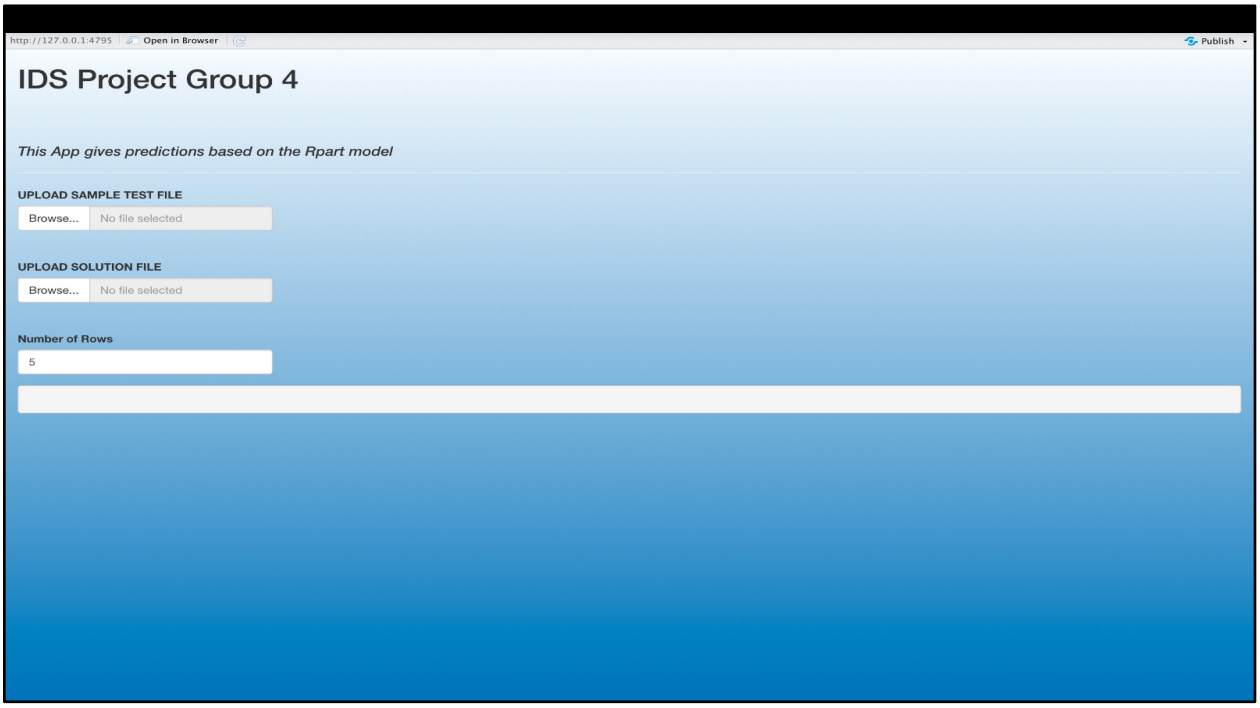
RPart / Tree Model

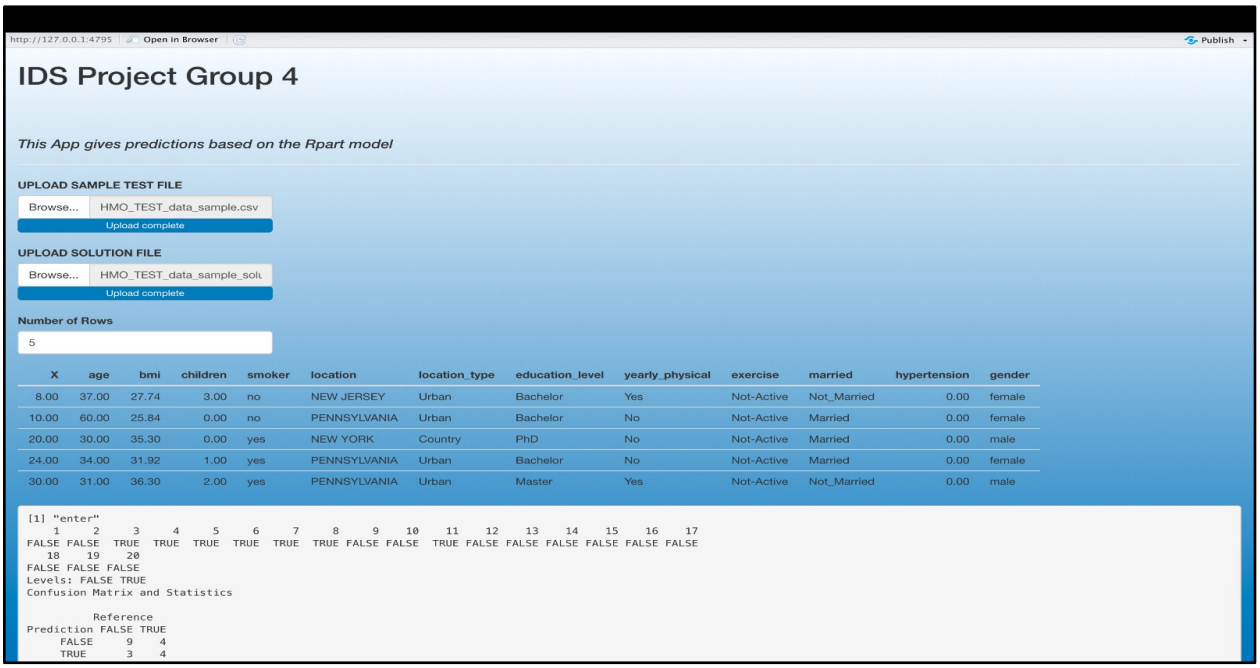
```
          Reference
Prediction FALSE TRUE
     FALSE  2233  361
     TRUE     41  397

               Accuracy : 0.8674
                 95% CI : (0.8548, 0.8793)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5885

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

            Sensitivity : 0.9820
            Specificity : 0.5237
         Pos Pred Value : 0.8608
         Neg Pred Value : 0.9064
             Prevalence : 0.7500
         Detection Rate : 0.7365
   Detection Prevalence : 0.8555
      Balanced Accuracy : 0.7529

       'Positive' Class : FALSE
```

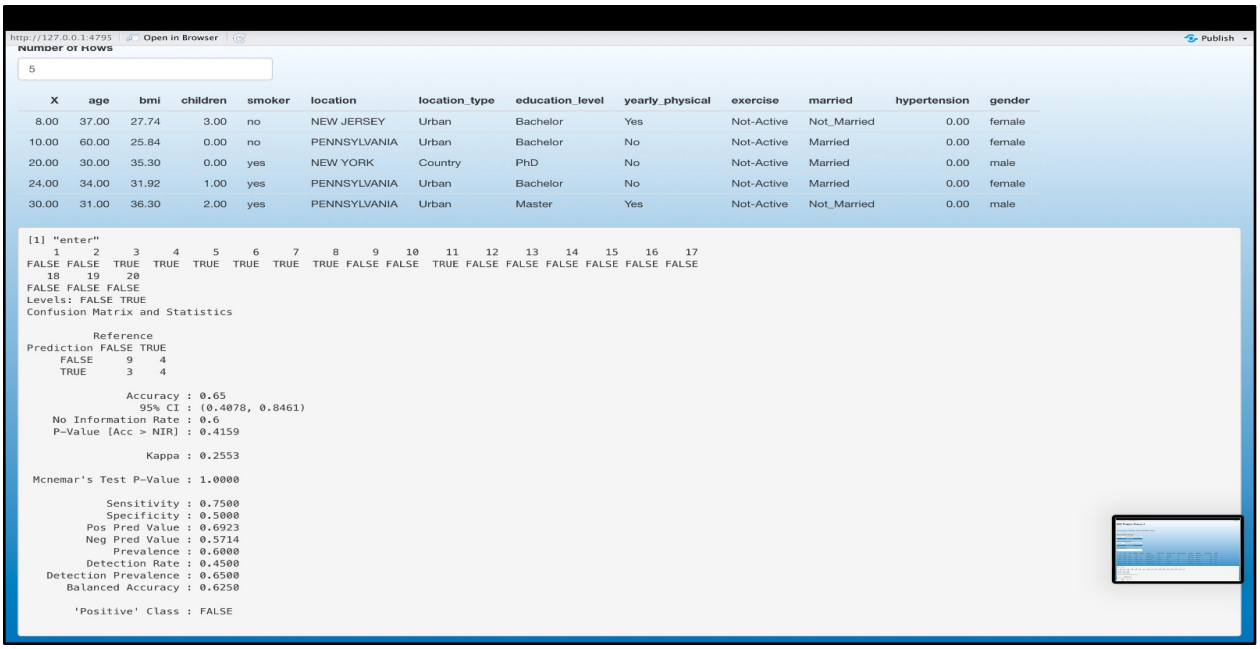


Here we can see that the accuracy of the model is 86.74% which is lower than the SVM model. But the sensitivity of the rpart model is 98.20% which is better than the SVM. This makes the rpart model the ideal choice for prediction and to be used in the shiny app.

10. Shiny Application







By running the shiny app using the rpart model we can see that it takes in the input of the sample file and the solution file. It then tries to predict the values using the rpart model that evaluates that prediction. Here the accuracy of the model came out to be 65% and the sensitivity is 75%. It also displays the confusion matrix of its prediction.

11. Conclusion

After analysis of the dataset, we are able to answer the questions from part 4 of this report:

1. Age has a strong positive correlation with health care cost. Generally, the older the customer is, the higher the health care costs.
2. BMI has a strong positive correlation with health care cost. Generally, the higher the BMI of a customer, the higher the health care costs.
3. Smoking has a strong positive correlation with health care costs. Generally, if the customer smokes, their health care costs will be more expensive.
4. Hypertension has a strong positive correlation with health care cost. Generally, if a patient has hypertension then their health care costs will be more expensive.
5. Exercise has a strong positive correlation with health care cost. Generally, a customer who is exercising will have lower health care costs.

Listed above are the variables that can be considered risk factors to an individual's health. It is evident that the customers under the HMO plan who either smoke, lack exercise, have hypertension, have a high BMI over 30, or are older than the age of 45 have higher health care costs.

12. Recommendations

1. Hire personal trainers to work with customers who exhibit high risk factors in an attempt to help achieve an overall healthy lifestyle. We need to be able to identify and categorize which customers fall under each risk factor in order to lower their health risks. Trainers will offer health surveys that will be accessible by their network of doctors under the HMO plan.
2. To incentivize customers using the health program, offer discounts on premiums for active participation with personal trainers.
3. Offer a health program that automatically enrolls customers in annual physicals and benefits such as a network of personal trainers.
4. Smokers and customers over the age of 45 are charged higher premiums to account for their potential increase of use in health care
5. To retain older aged customers, offer family HMO plans at reduced premiums. It is unlikely that all family members will require health care all at once or within a short period of time.