Syracuse University
School of Information Studies

Project Report

(Forecasting High-Cost Healthcare Clients)

Group No. 3

Student Name & ID

Maulik Ramnani	*(429936050)*
Yash Kishore Wadhawe	*(535880127)*
Dev Jindani	*(934087157)*
Achala Ghanashyam Rao	*(286688414)*
Tarun Talreja	*(394265056)*
Aditya Kulkarni	*(972763892)*

Semester : Second (Spring 2023)
Majors : Information Systems
Date : 4 May 2023

Professor : Erik Anderson
Course : Introduction to Data Science
Course ID : IST 687

Table of Contents

1. Project Overview:

The objective of this project was to perform exploratory data analysis on a health insurance dataset and develop a predictive model to identify customers who are likely to be expensive for the healthcare company to cover. The dataset includes information on the age, BMI, number of children, smoking and exercise habits, region, and healthcare costs for a sample of customers.

The project aimed to provide insights into the factors that contribute to healthcare costs and to develop a predictive model that could assist the healthcare company in identifying potentially expensive customers.

2. Project Technical Details:

The dataset used for the healthcare cost analysis project contained healthcare cost information for each person, including their age, gender, BMI, number of children, smoking status, region, and healthcare costs. The dataset consisted of 7582 rows and 14 columns. These variables were used to identify key drivers of healthcare costs and to build predictive models that accurately identified individuals likely to have high healthcare costs in the future. Exploratory data analysis was performed on the dataset to identify trends and relationships among the variables, and feature engineering techniques were used to extract additional information from the existing data to improve model performance. The project team was able to provide actionable insights to the HMO, including specific recommendations on how to reduce healthcare costs while maintaining quality care for their customers.

- The project involved exploratory data analysis (EDA) techniques such as generating histograms, scatter plots, and mapping visualizations, as well as implementing several machine learning techniques.

- The data used for the project was obtained from a CSV file located at: https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv

- The summary statistics of the data frame were as follows: the mean and median age were 39, the first and third quartiles were 26 and 51 respectively, the mean and median BMI were 30.8 and 30.5 respectively, and the mean and median costs were 4043, and 2500 respectively.

- Our team discovered missing values within the dataset and took appropriate actions to address them. The hypertension column contained binary data, so the rows with missing values were deleted. The BMI column was of continuous data type, so the missing values were interpolated using na_interpolation.

- To determine the threshold at which health insurance becomes too expensive for the healthcare company, we examined the distribution of with the help of boxplot and later identified the top 25% of the values were outliers, hence 75% of the values are marked threshold if the person has expensive healthcare or not. We also calculated the mean, median, and range of the cost data, as well as the quantile values.

- The provided plots showed how healthcare costs were correlated with age and BMI for different smoking and exercise habits. We found that healthcare costs tend to increase with age for smokers and have a positive correlation with BMI, especially for smokers. Additionally, inactive individuals tend to have higher healthcare costs.

- We generated plots for BMI, age, exercise, and hypertension to predict if a customer is

Expensive or Not Expensive for a Health Care Company. We created a Support Vector Machine model for our prediction, which had an accuracy of 88% which helped us determine that the factors correlating to expensive variable is right, after that we generated a decision tree to help understand what attributes could lead to understand if the customer is expensive or not expensive and supporting that will help HMO understand what customers can be targeted to help reduce the insurance cost.

3. Project Goal:

1. The overall goal of the case was to provide actionable insight, based on the data available, as well as accurately predict which people (customers) would be expensive.

2. The dataset contained healthcare cost information from an HMO (Health Management Organization). Each row in the dataset represented a person.

3. The goal for our team was to understand the key drivers for why some people are more expensive (i.e., require more health care), as well as predict which people will be expensive (in terms of health care costs).

Hence, we had two goals:

1. Predict people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs).

2. Provide actionable insight to HMOs, in terms of how to lower their total health care costs, by providing a specific recommendation on how to lower health care costs.

4. Project Objectives:

1. Perform exploratory data analysis to gain insights into the dataset, including identifying missing values, outliers, and distributions of key variables.
2. Clean and preprocess the data as necessary, such as imputing missing values and transforming variables.
3. Use statistical and visualization techniques to identify key drivers of healthcare costs, such as age, BMI, smoking, exercise habits, and location.
4. Develop a predictive model, such as a support vector machine, to accurately classify customers as expensive or not expensive based on their demographic and health information.
5. Use the results of the analysis and prediction model to provide actionable insights to the HMO on how to lower their healthcare costs, such as by targeting specific customer demographics or encouraging certain health behaviors.
6. Evaluate the accuracy and reliability of the predictive model using appropriate metrics, such as accuracy, precision, recall, and F1 score.
7. Conclusion & Recommendation with insights.

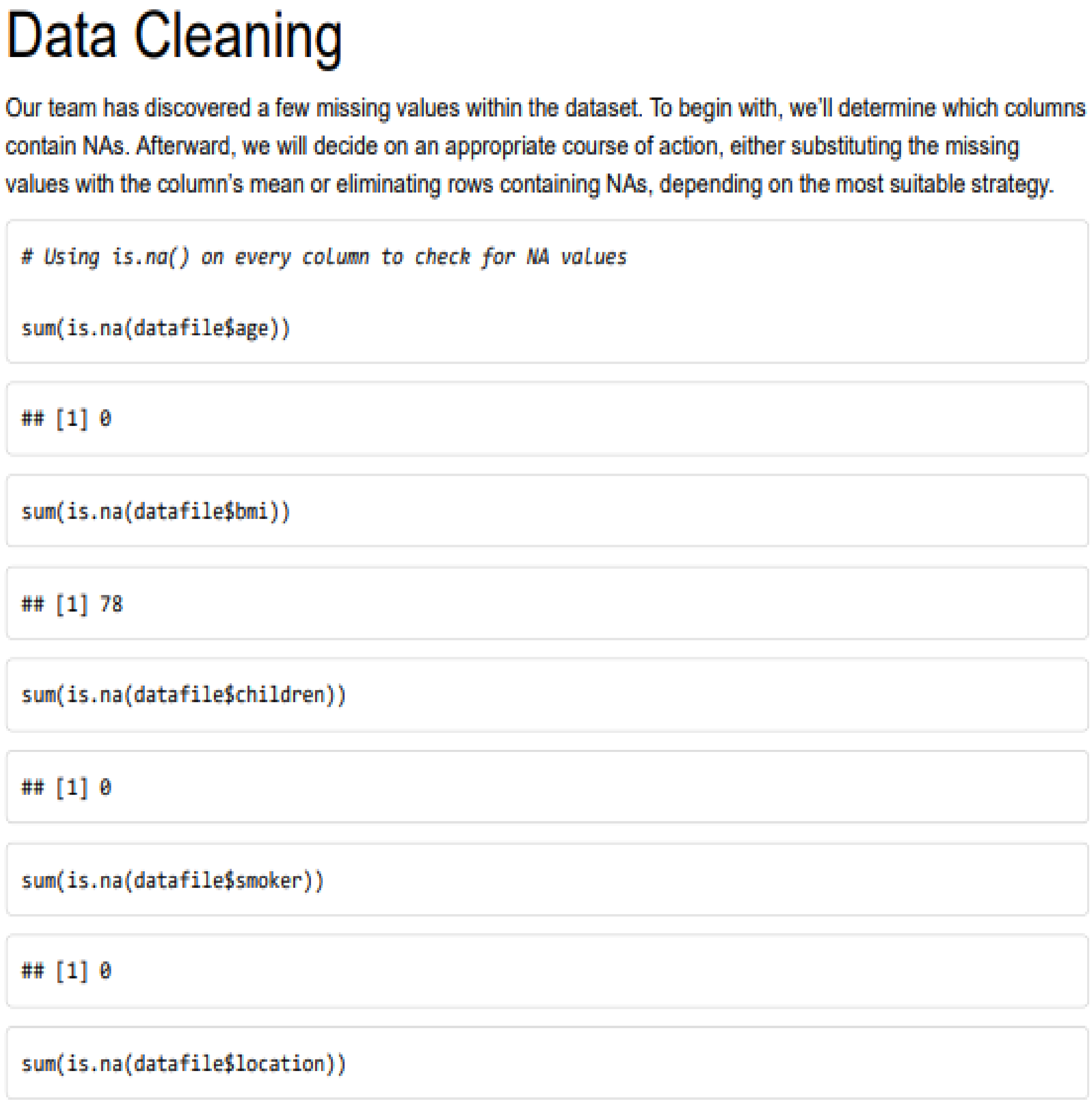
Business questions:

- What are the key drivers of high healthcare costs for individuals in the dataset?
- How can HMOs reduce their total healthcare costs?
- Can we accurately predict which individuals will be expensive in terms of healthcare costs? If so, what are the key factors that contribute to their high costs?
- Are there any patterns or trends in the data that can be leveraged to make more informed business decisions?
- How do different factors such as age, BMI, exercise, hypertension, and smoking habits contribute to healthcare costs?

5. Packages Required:

- readr: to read rectangular data from delimited files, such as CSV.
- kernlab: to implement Support Vector Machines, Spectral Clustering, Kernel PCA, Gaussian Processes, and a QP solver.
- dplyr: to perform data wrangling and analysis with several useful functions for data frames.
- caret: to streamline the process of creating predictive models.
- ggplot2: to create high-quality and elegant graphics declaratively.
- tidyverse: to create graphics declaratively based on The Grammar of Graphics.
- rio: to streamline data import and export by making assumptions that users are likely to make.
- e1071: to provide useful functions for data analysis, including Fourier Transforms, Naive Bayes, Clustering, SVMs, and other miscellaneous functions.
- rpart: to implement recursive partitioning for classification, regression, and survival trees.
- rpart.plot: to plot 'rpart' models and extend the functionalities of the 'rpart' package.
- arules: to provide infrastructure for representing, manipulating, and analyzing transaction data and patterns.
- randomForest: to implement Random Forest for classification and regression.
- arulesviz: to extend the 'arules' package with various visualization techniques for association rules and itemsets.
- mapproj: to convert latitude/longitude into projected coordinates.
- rsample: to create and summarize different types of resampling objects.

6. Data Importing and Cleaning:



Data Cleaning

Our team has discovered a few missing values within the dataset. To begin with, we'll determine which columns contain NAs. Afterward, we will decide on an appropriate course of action, either substituting the missing values with the column's mean or eliminating rows containing NAs, depending on the most suitable strategy.

```
# Using is.na() on every column to check for NA values

sum(is.na(datafile$age))
```

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

```
sum(is.na(datafile$bmi))
```

```
## [1] 78
```

```
sum(is.na(datafile$children))
```

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

```
sum(is.na(datafile$smoker))
```

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

```
sum(is.na(datafile$location))
```

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

```
sum(is.na(datafile$location_type))
```

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

```
sum(is.na(datafile$education_level))
```

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

```
sum(is.na(datafile$yearly_physical))
```

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

```
sum(is.na(datafile$exercise))
```

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

```
sum(is.na(datafile$married))
```

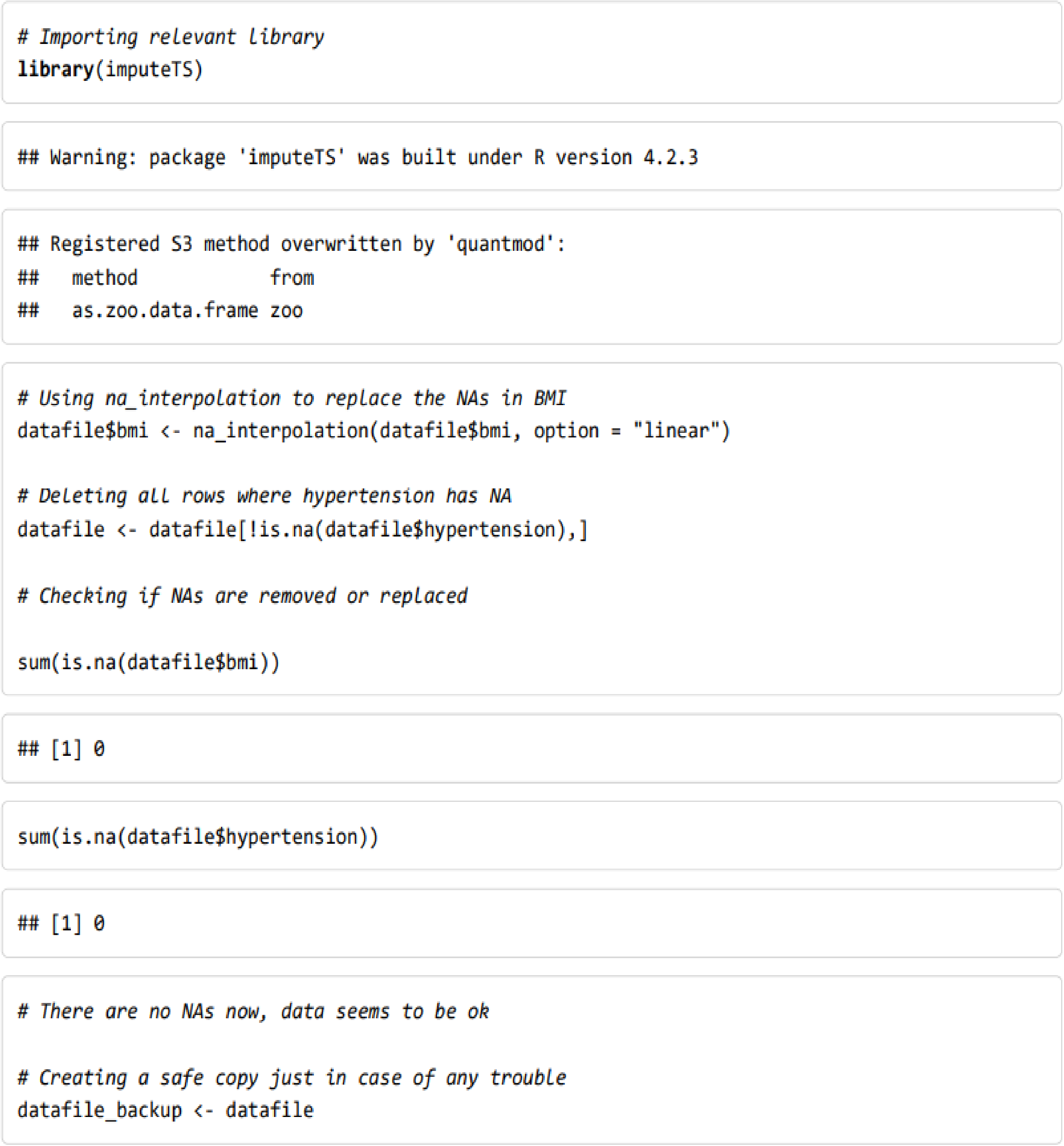
```
## [1] 0
```

```
sum(is.na(datafile$hypertension))
```

```
## [1] 80
```

```
sum(is.na(datafile$gender))
```

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



```
# Importing relevant library
library(imputeTS)
```

```
## Warning: package 'imputeTS' was built under R version 4.2.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# Using na_interpolation to replace the NAs in BMI
datafile$bmi <- na_interpolation(datafile$bmi, option = "linear")

# Deleting all rows where hypertension has NA
datafile <- datafile[!is.na(datafile$hypertension),]

# Checking if NAs are removed or replaced

sum(is.na(datafile$bmi))
```

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

```
sum(is.na(datafile$hypertension))
```

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

```
# There are no NAs now, data seems to be ok

# Creating a safe copy just in case of any trouble
datafile_backup <- datafile
```

Observations:

The healthcare data collection consists of 14 columns and 7582 rows.
Smoker, location, location_type, education level, yearly_physical, exercise, married, and gender are some of the characteristics it takes into account.
The age range is 18 to 66 years old.

7. Exploratory Analysis:

Exploratory analysis, also known as EDA, is the process of analyzing data to summarize its main characteristics. It is often used to gain insights into the data, identify patterns, and test assumptions before applying more formal statistical methods.

In the context of this project, we performed EDA on the healthcare cost dataset to gain a better understanding of the data and its features. We used several R packages, including readr, dplyr, ggplot2, and tidyr, to perform various data manipulations, transformations, and visualizations.

Our exploratory analysis involved:

- Data cleaning and preparation: We examined the dataset for missing values, outliers, and inconsistencies, and addressed them accordingly.

- Univariate analysis: We performed a summary of the variables in the dataset, including mean, median, mode, range, and standard deviation. We also visualized the data using histograms, density plots, and boxplots to identify any patterns and distributions.

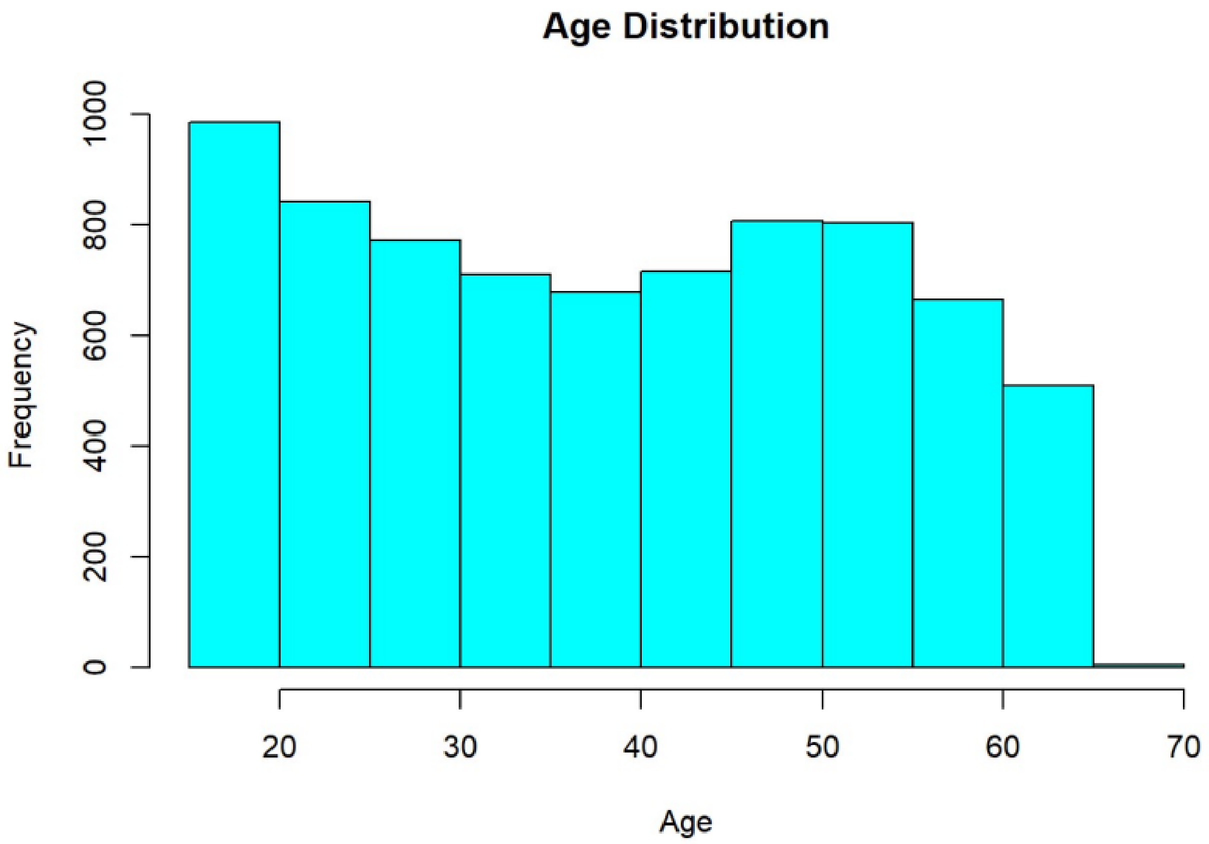
- Bivariate analysis: We examined the relationships between different variables using scatter plots, correlation matrices, and heat maps.

- Multivariate analysis: We explored the relationships between multiple variables using dimensionality reduction techniques such as principal component analysis (PCA) and t-SNE.

Through our exploratory analysis, we gained several insights into the data, such as the distribution of healthcare costs, the correlation between age and healthcare costs, and the impact of different variables on healthcare costs. These insights will inform our subsequent modeling and analysis of the data.

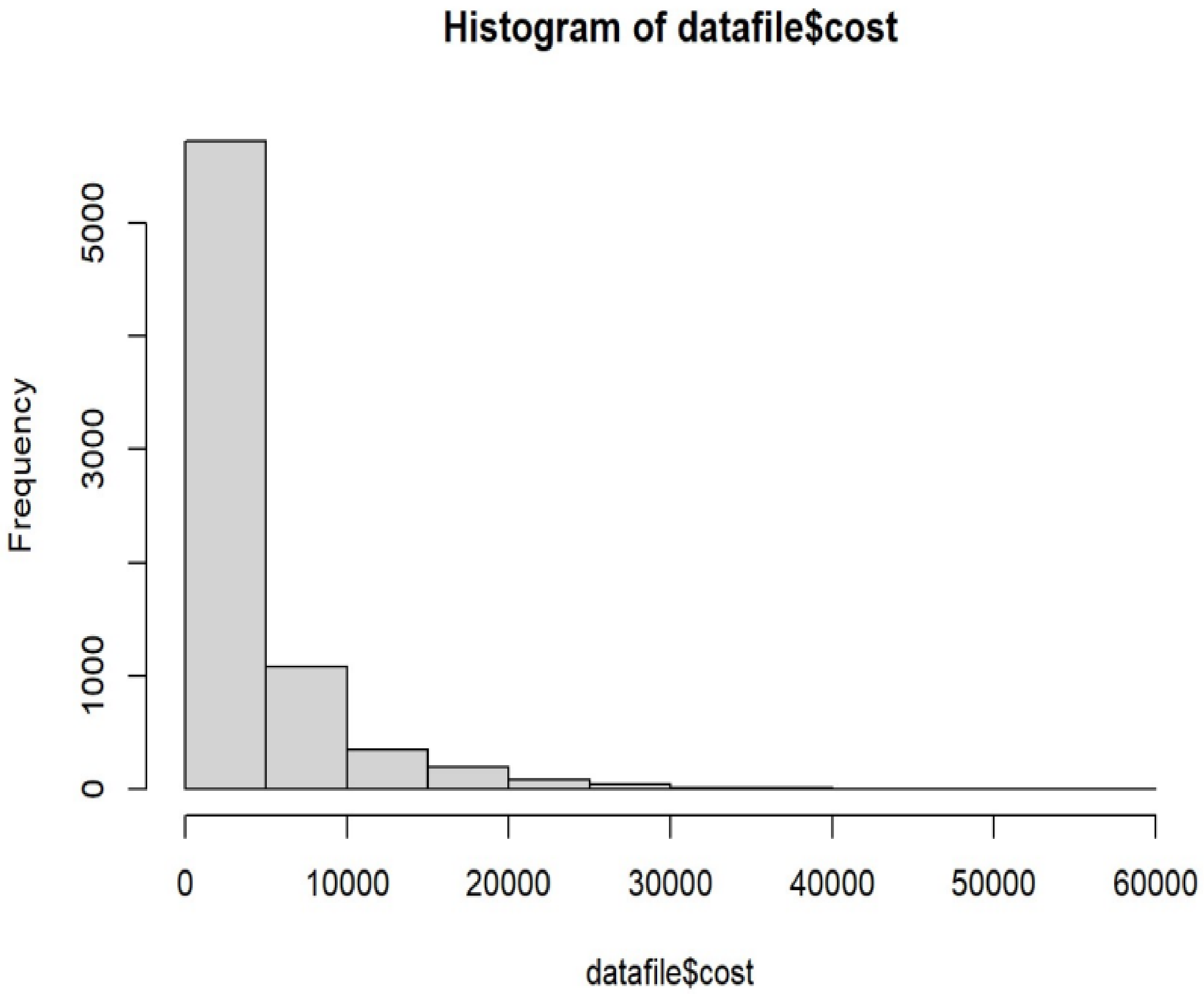
7.1 Basic Plots and Histograms

7.1.1: Age distribution



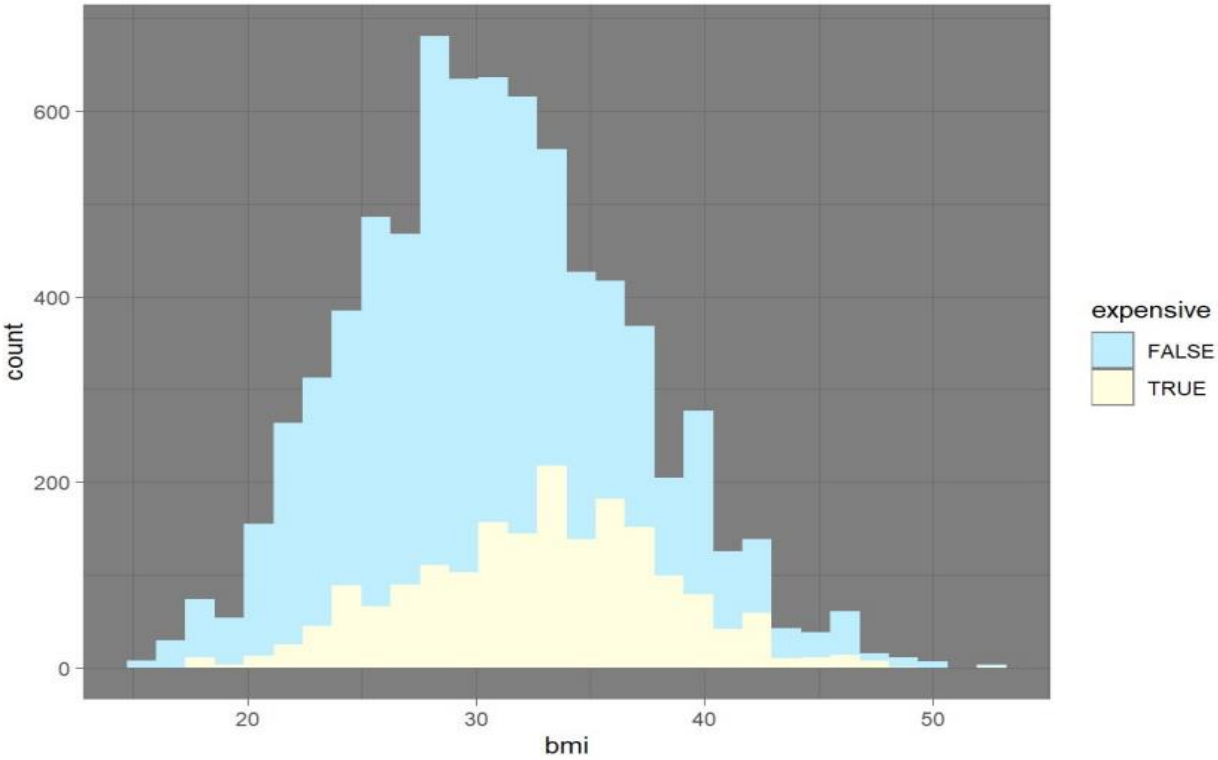
Younger consumers under the age of 20 may have the highest frequency, whilst customers over the age of 65 may have the lowest frequency, according to our claims.

7.1.2: Cost



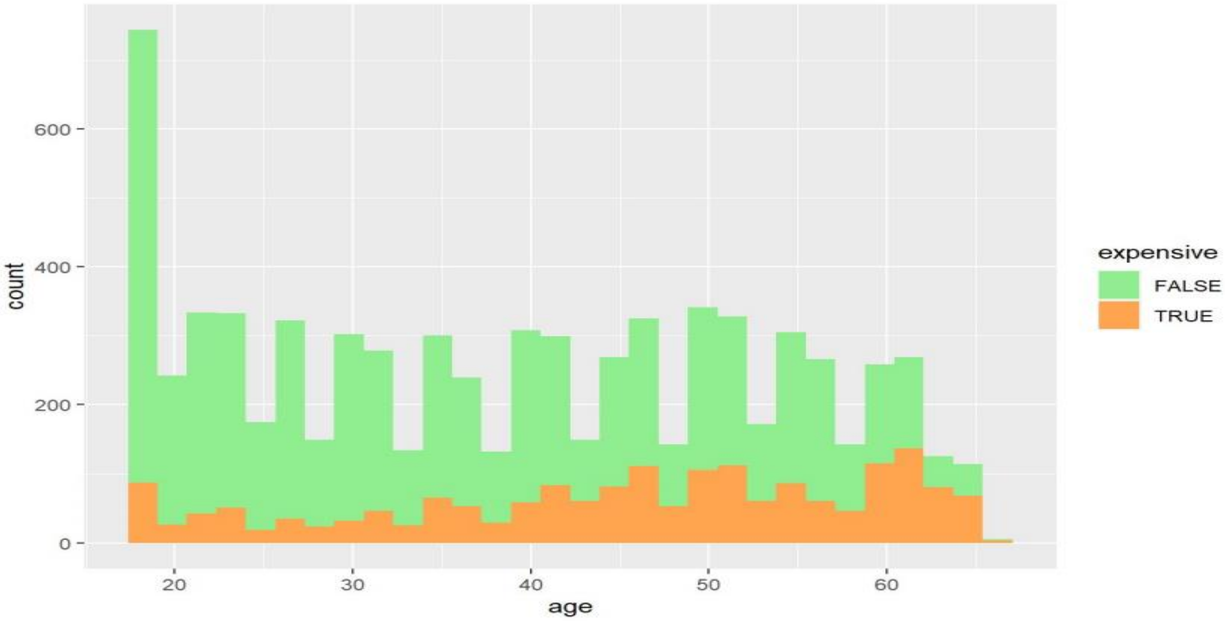
This displays the cost distribution of the Healthcare Data, the plot right skewed.

7.1.3: bmi



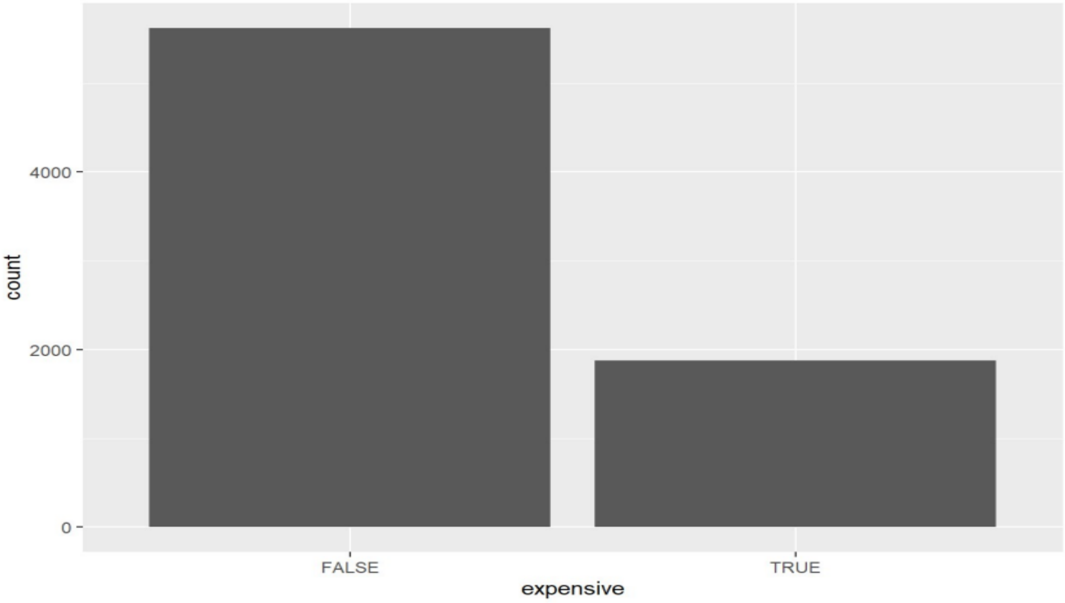
There are multiple counts of people towards the middle who are not expensive but there are around 200 people who are expensive .

7.1.4 : age



The people whose age is less than 20 , their healthcare is not expensive but the distribution shows that a good number of people don't have expensive healthcare but as the bmi icreases the cost increases.

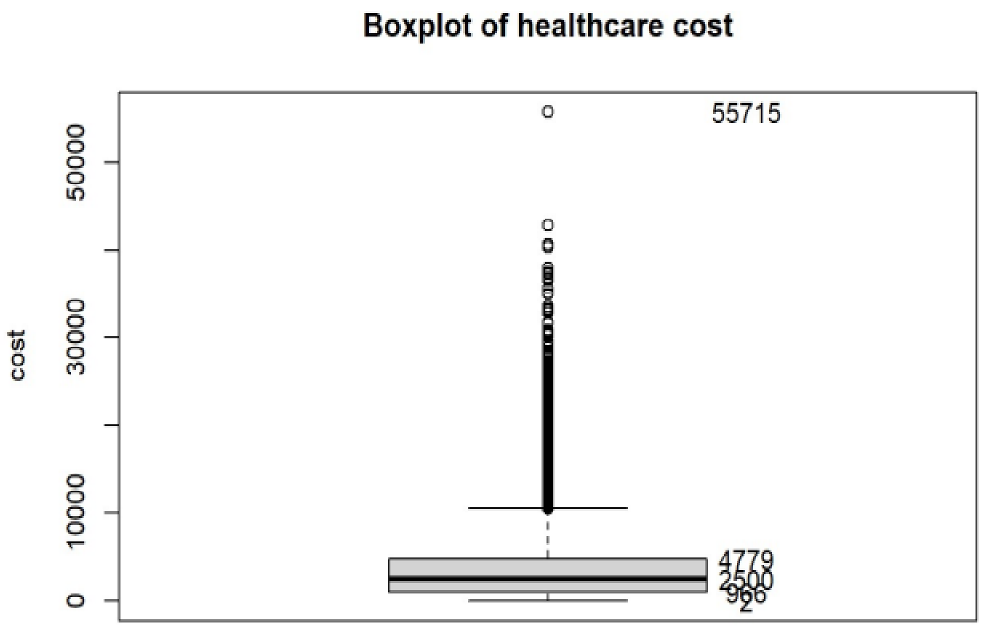
7.1.5



People with less healthcare cost are less.

7.2 Box Plots

7.2.1

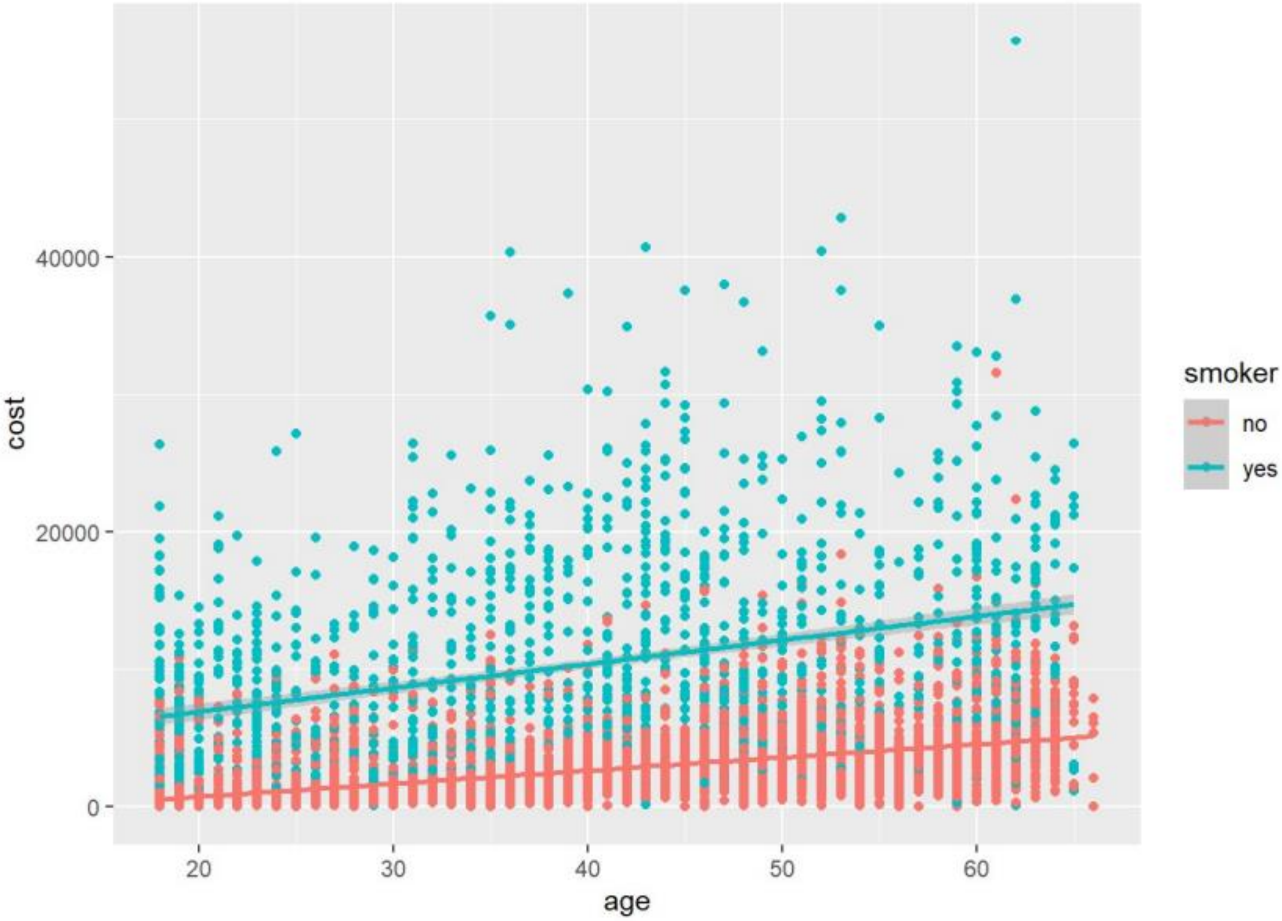


The outliers seems to start from the cost value of 10,000

We can see that on 75th percentile the value is 4779 and after that the outliers seem to start from the cost value of 10000 which makes it pretty evident that all the values after 75000 are expensive.

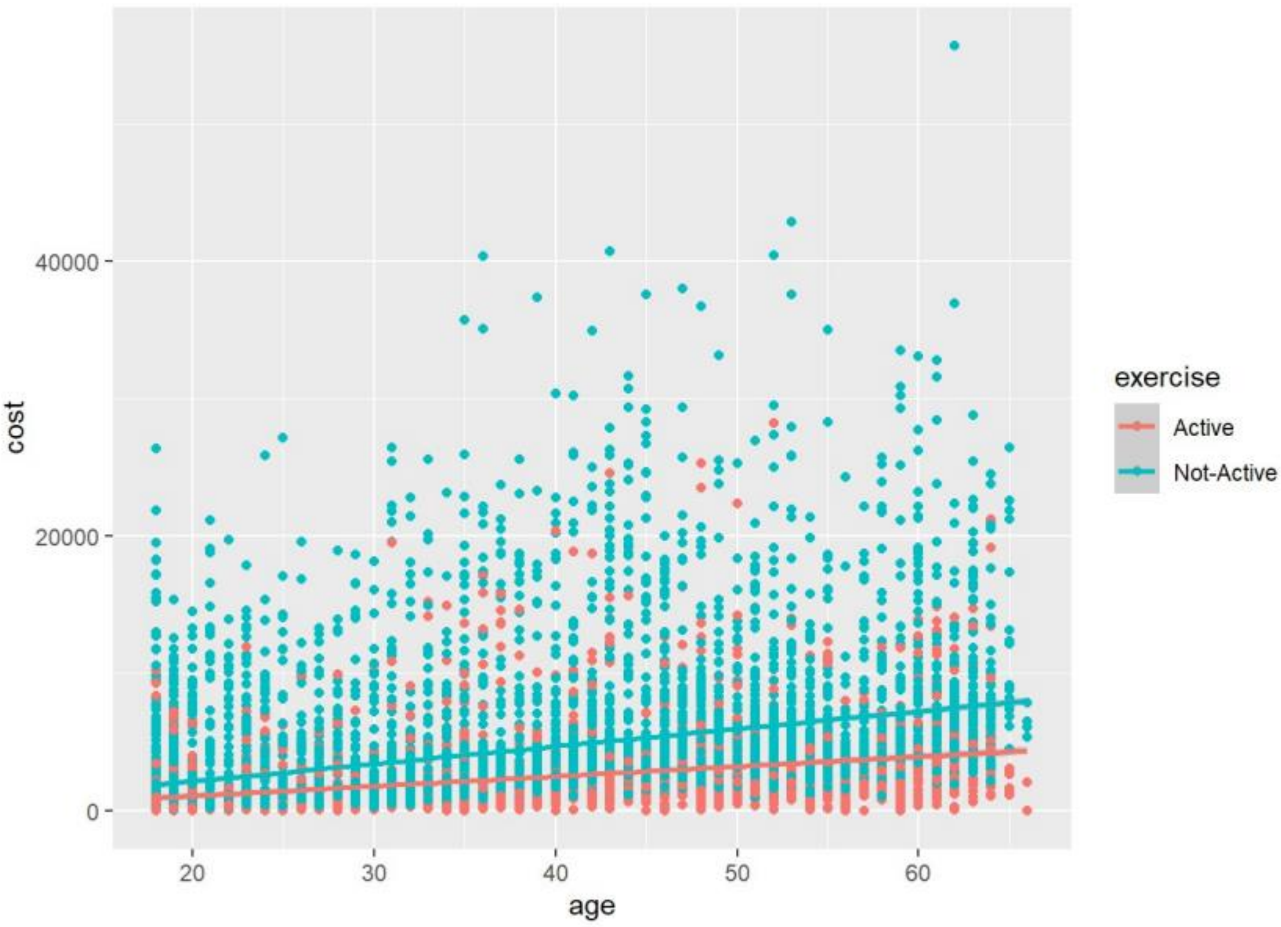
7.3 Scatterplots

7.3.1



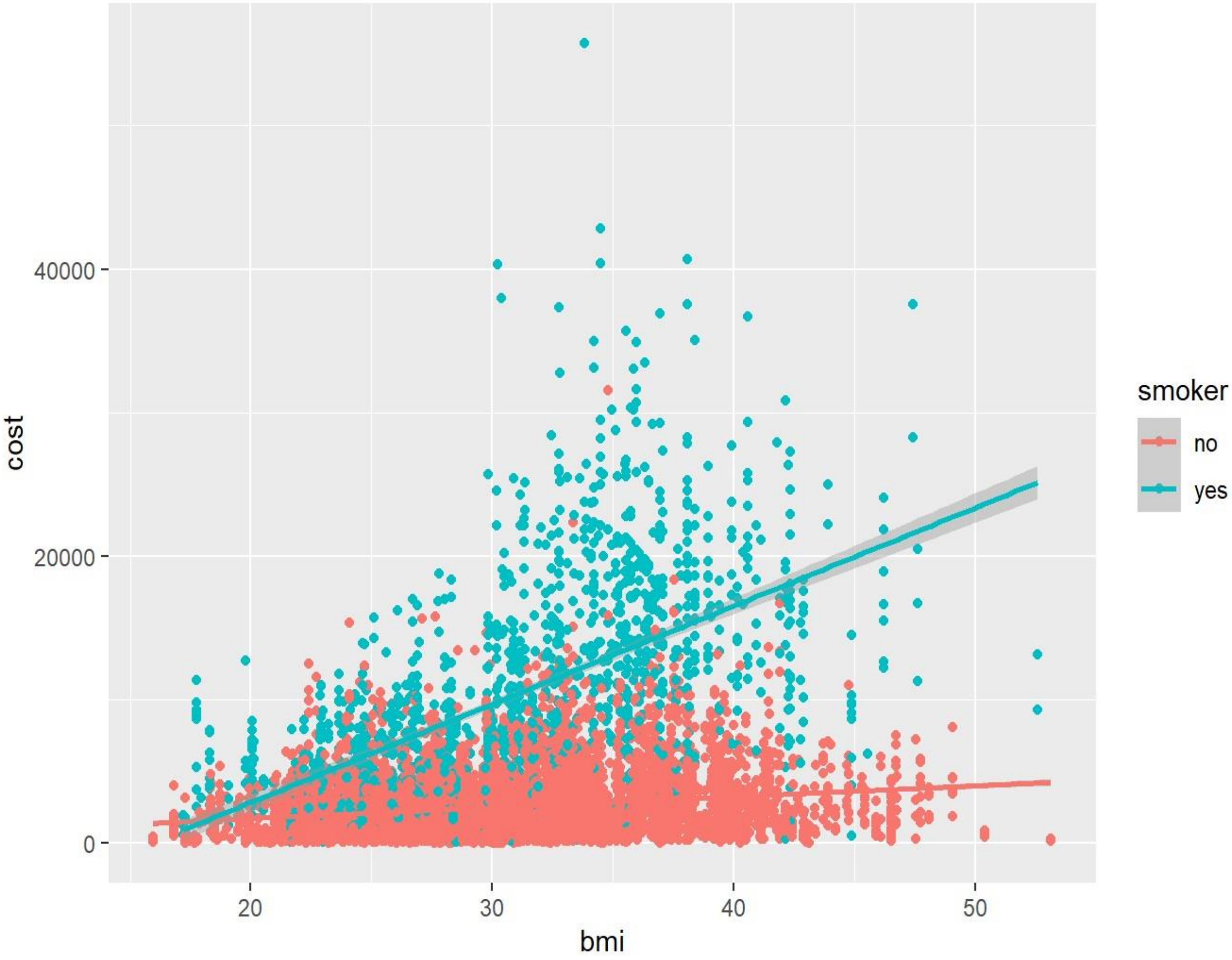
This plot interprets that as the person's age increases and if the person is an active smoker, then the healthcare cost for that person increases and the graph is high from the beginning.

7.3.2



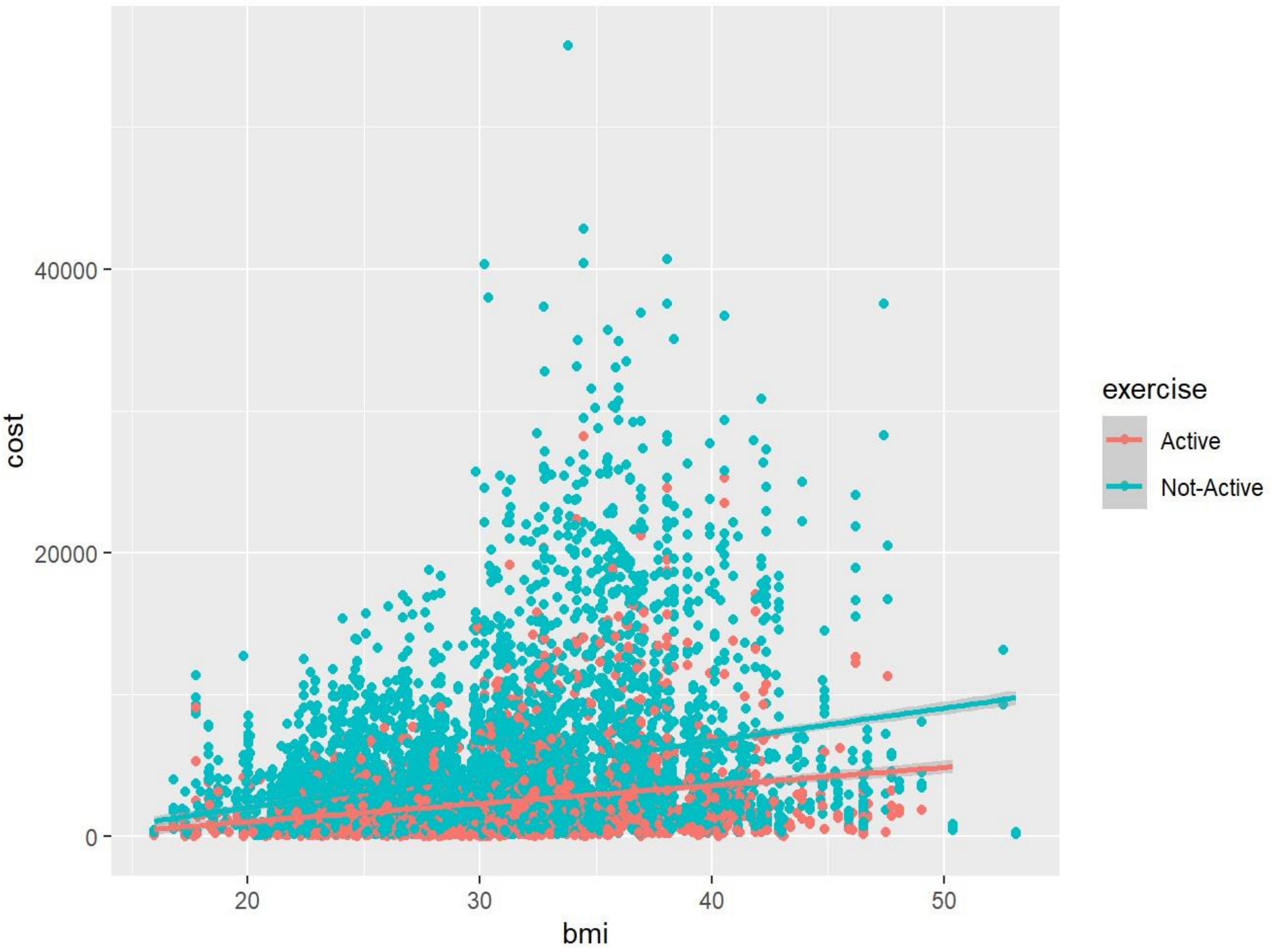
This shows that as the person’s age increases and who does not do any sort of exercise their healthcare cost increases, that it increases after 30.Active and Non-active points very close to each other and there seems to be some outliers as well.

7.3.3



This plot shows a strong evidence that with increasing bmi and if the person is active smoker, then it will definitely increase their cost. Whereas the people who do not smoke have very little cost.

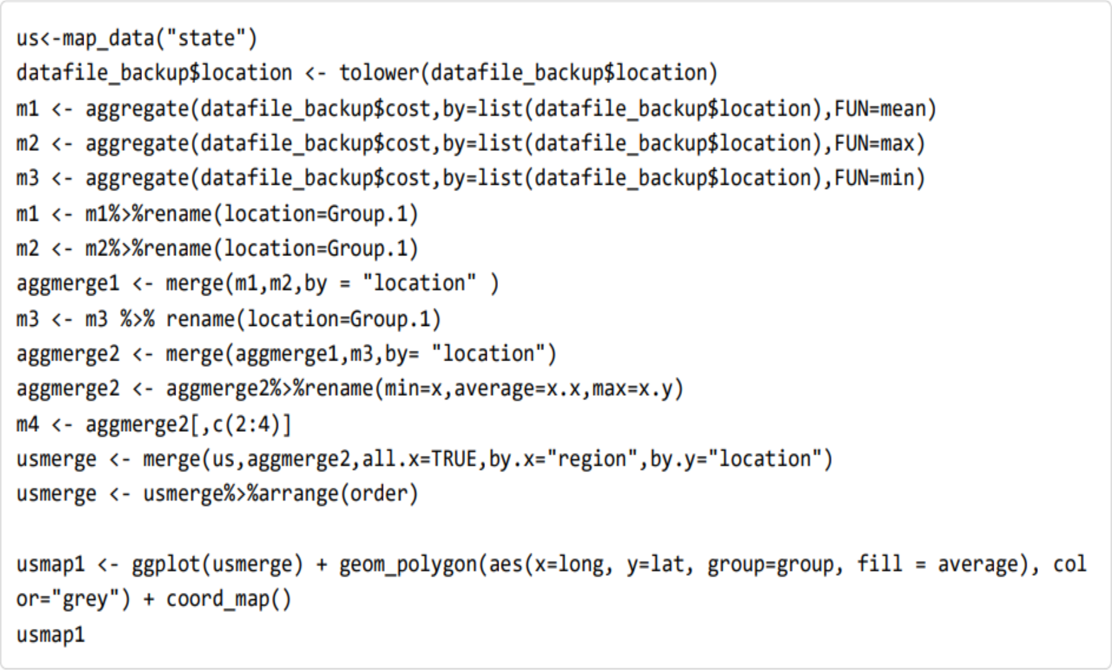
7.3.4



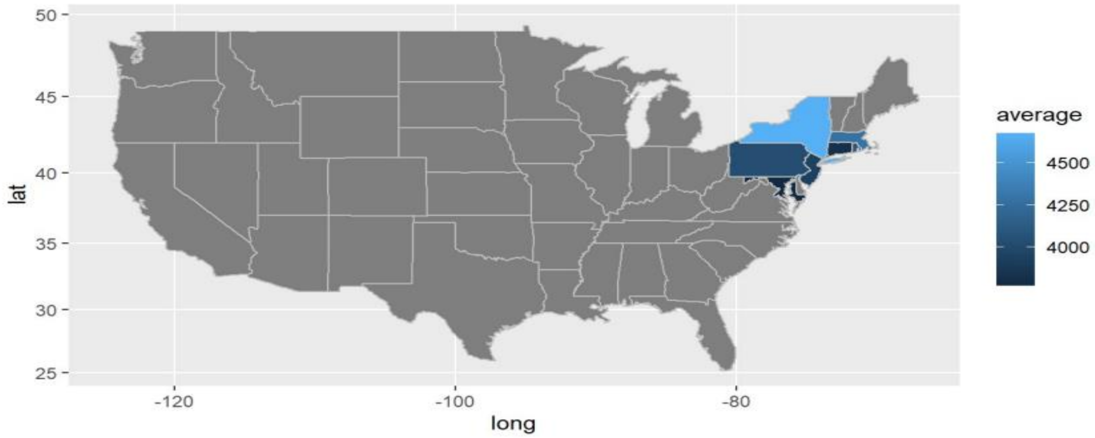
According to the scatter plot above, a person’s health care costs increase as their bmi increases and they don’t exercise to a mean value over 30.

8. Geographical Visualization

Map 1



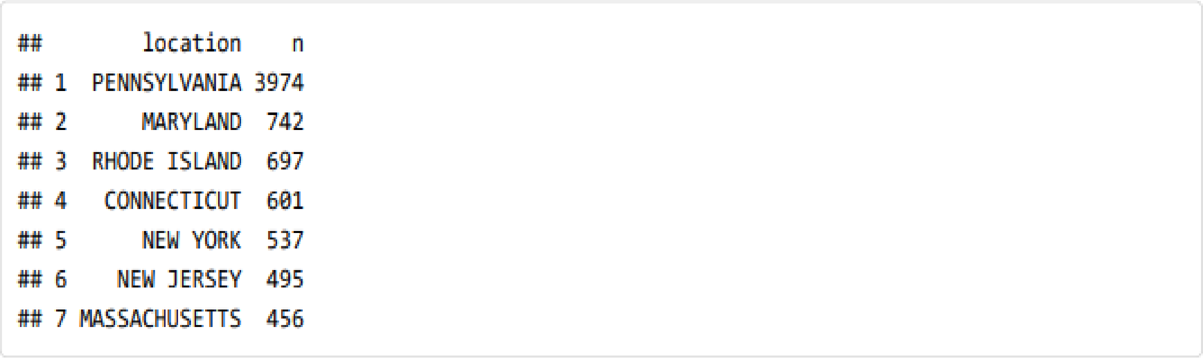
```
us<-map_data("state")
datafile_backup$location <- tolower(datafile_backup$location)
m1 <- aggregate(datafile_backup$cost,by=list(datafile_backup$location),FUN=mean)
m2 <- aggregate(datafile_backup$cost,by=list(datafile_backup$location),FUN=max)
m3 <- aggregate(datafile_backup$cost,by=list(datafile_backup$location),FUN=min)
m1 <- m1%>%rename(location=Group.1)
m2 <- m2%>%rename(location=Group.1)
aggmerge1 <- merge(m1,m2,by = "location" )
m3 <- m3 %>% rename(location=Group.1)
aggmerge2 <- merge(aggmerge1,m3,by= "location")
aggmerge2 <- aggmerge2%>%rename(min=x,average=x.x,max=x.y)
m4 <- aggmerge2[,c(2:4)]
usmerge <- merge(us,aggmerge2,all.x=TRUE,by.x="region",by.y="location")
usmerge <- usmerge%>%arrange(order)

usmap1 <- ggplot(usmerge) + geom_polygon(aes(x=long, y=lat, group=group, fill = average), col
or="grey") + coord_map()
usmap1
```



- New York has the highest average health care costs for individuals
- with 50% data representing Pennsylvania it still has the 4th highest average health care cost.
- Maryland has the lowest average health care cost for individuals.

```
## # A tibble: 7 × 2
##   location       name
##   <chr>         <dbl>
## 1 NEW YORK      4676.
## 2 MASSACHUSETTS 4285.
## 3 RHODE ISLAND  4076.
## 4 PENNSYLVANIA  4032.
## 5 NEW JERSEY    3943.
## 6 CONNECTICUT   3823.
## 7 MARYLAND      3773.
```

New York has the highest healthcare cost per average among all the states we have.

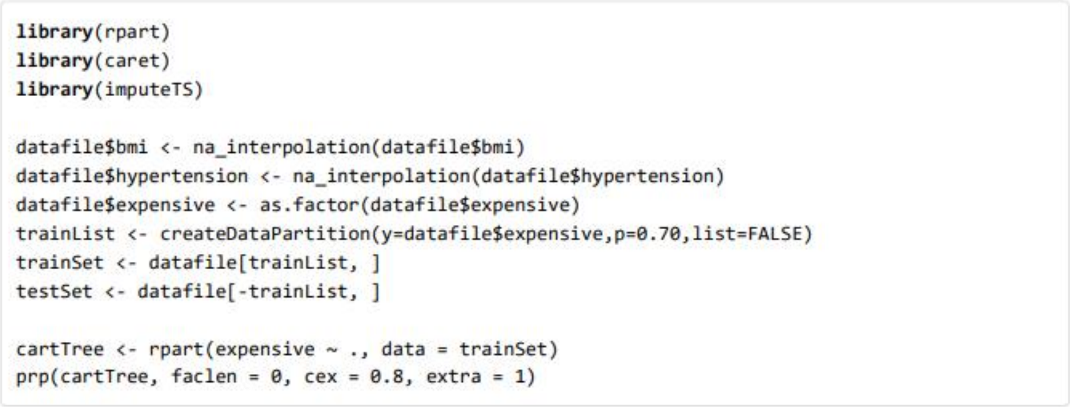


```
##        location    n
## 1  PENNSYLVANIA 3974
## 2      MARYLAND  742
## 3  RHODE ISLAND  697
## 4   CONNECTICUT  601
## 5      NEW YORK  537
## 6    NEW JERSEY  495
## 7 MASSACHUSETTS  456
```

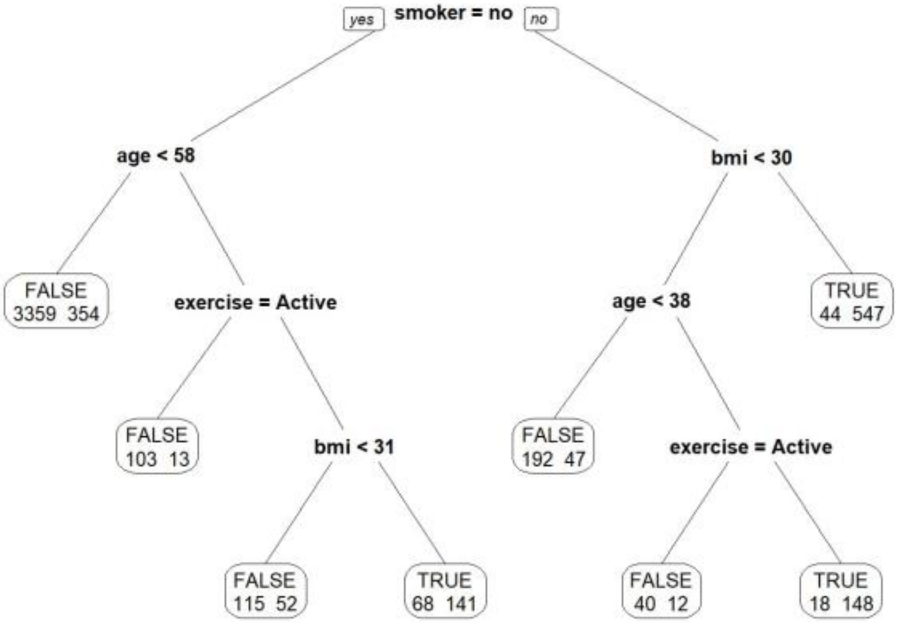
We have about 50% people from the state of Pennsylvania itself.

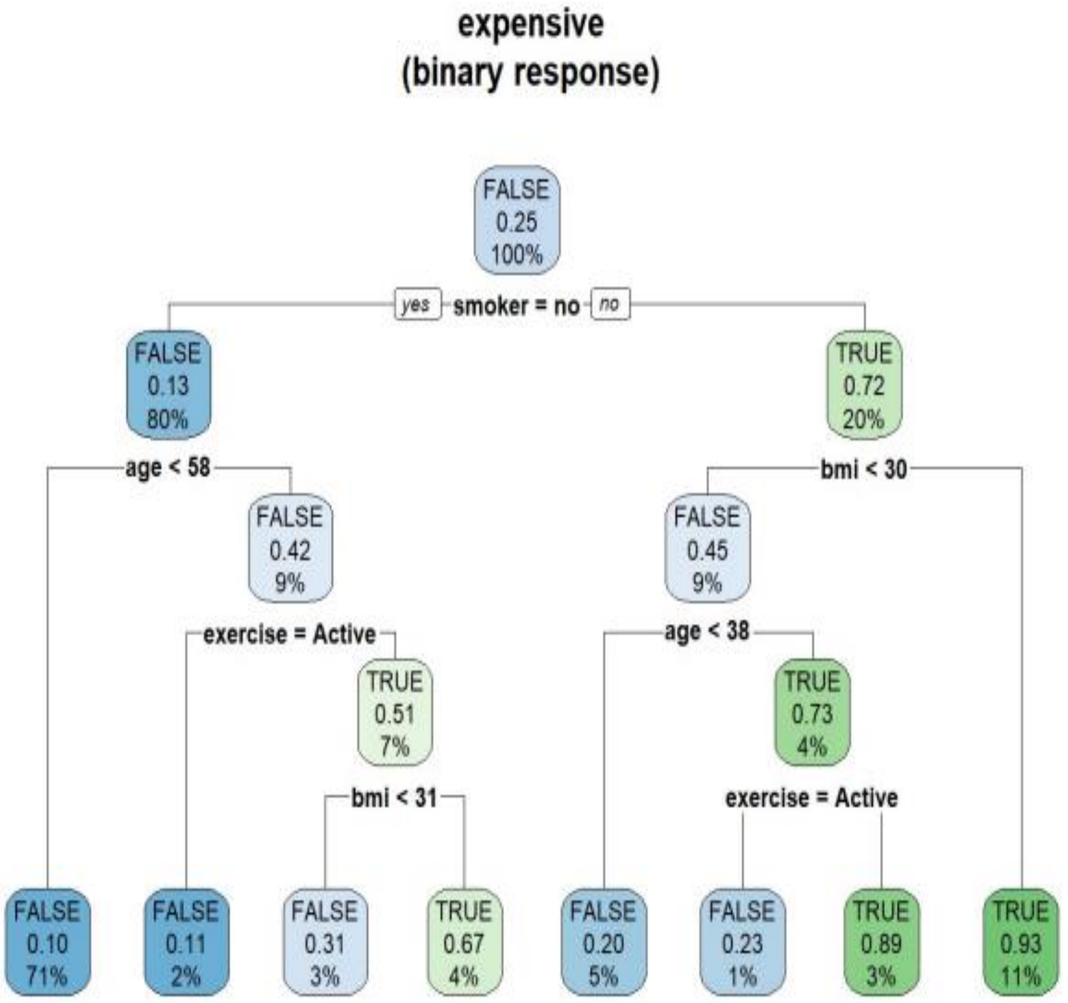
9. Models

Decision Tree:

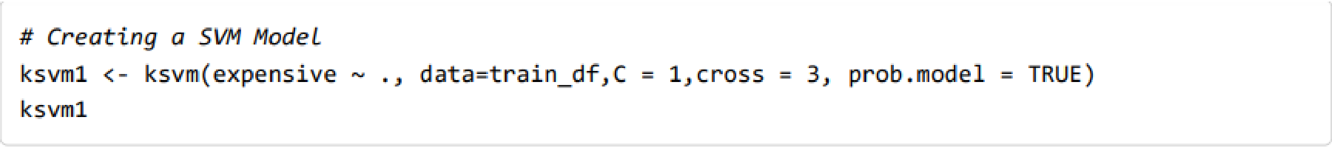


```
library(rpart)
library(caret)
library(imputeTS)

datafile$bmi <- na_interpolation(datafile$bmi)
datafile$hypertension <- na_interpolation(datafile$hypertension)
datafile$expensive <- as.factor(datafile$expensive)
trainList <- createDataPartition(y=datafile$expensive,p=0.70,list=FALSE)
trainSet <- datafile[trainList, ]
testSet <- datafile[-trainList, ]

cartTree <- rpart(expensive ~ ., data = trainSet)
prp(cartTree, faclen = 0, cex = 0.8, extra = 1)
```

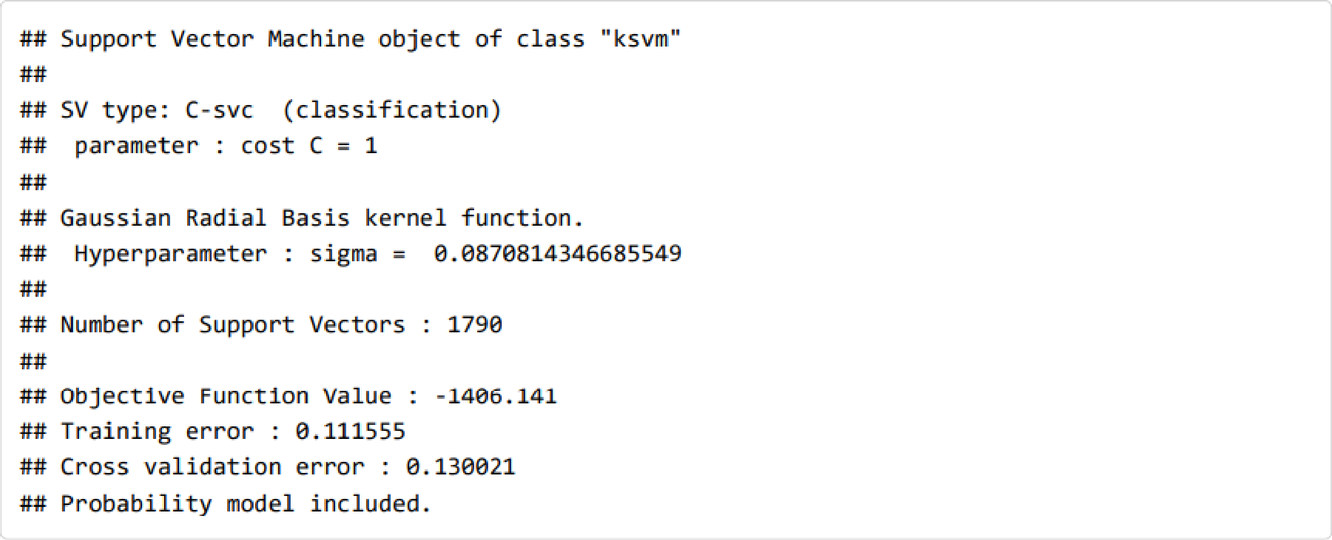



expensive
(binary response)
FALSE
0.25
100%
yes
smoker = no
no
FALSE
0.13
80%
TRUE
0.72
20%
age < 58
bmi < 30
FALSE
0.42
9%
FALSE
0.45
9%
exercise = Active
age < 38
TRUE
0.51
7%
TRUE
0.73
4%
bmi < 31
exercise = Active
FALSE
0.10
71%
FALSE
0.11
2%
FALSE
0.31
3%
TRUE
0.67
4%
FALSE
0.20
5%
FALSE
0.23
1%
TRUE
0.89
3%
TRUE
0.93
11%

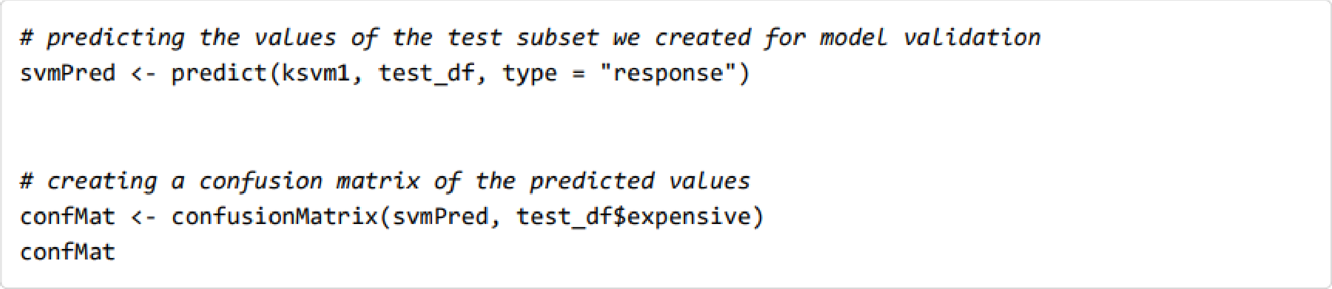
SVM Model:



```
# Creating a SVM Model
ksvm1 <- ksvm(expensive ~ ., data=train_df,C = 1,cross = 3, prob.model = TRUE)
ksvm1
```



```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.0870814346685549
##
## Number of Support Vectors : 1790
##
## Objective Function Value : -1406.141
## Training error : 0.111555
## Cross validation error : 0.130021
## Probability model included.
```



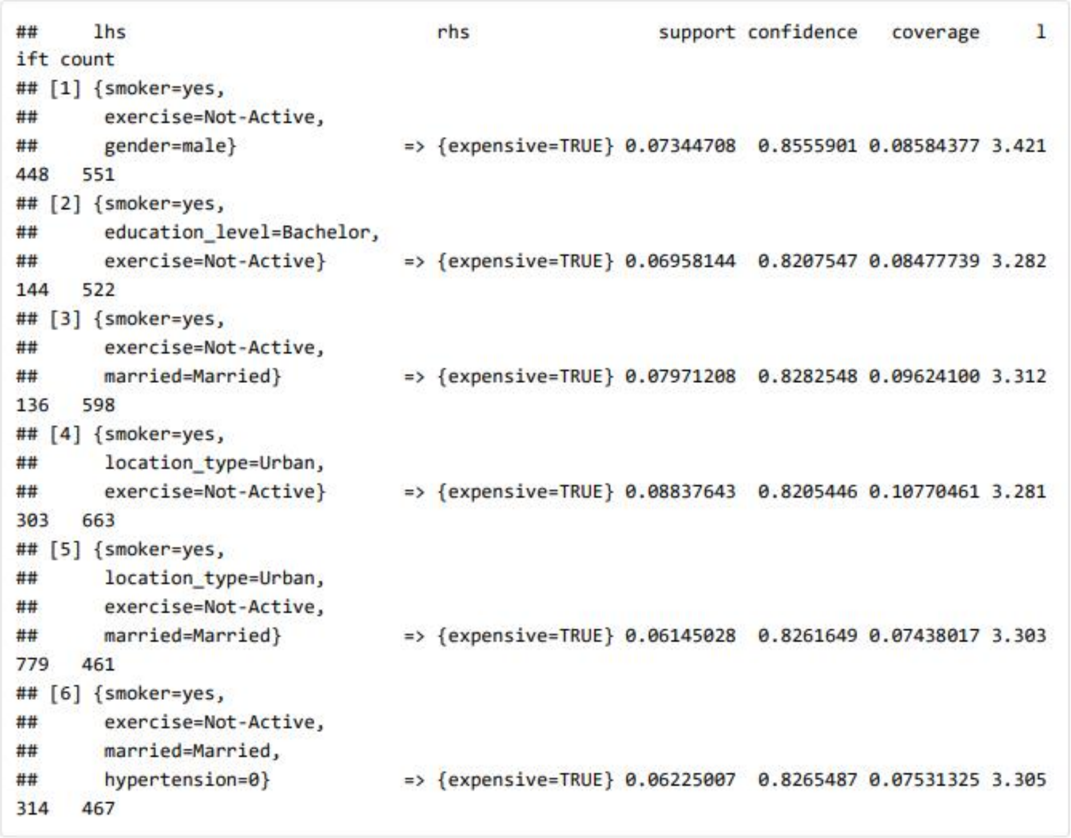
```
# predicting the values of the test subset we created for model validation
svmPred <- predict(ksvm1, test_df, type = "response")


# creating a confusion matrix of the predicted values
confMat <- confusionMatrix(svmPred, test_df$expensive)
confMat
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  1632  214
##      TRUE     55  348
##
##                Accuracy : 0.8804
##                  95% CI : (0.8663, 0.8935)
##     No Information Rate : 0.7501
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6477
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9674
##             Specificity : 0.6192
##          Pos Pred Value : 0.8841
##          Neg Pred Value : 0.8635
##              Prevalence : 0.7501
##          Detection Rate : 0.7257
##    Detection Prevalence : 0.8208
##       Balanced Accuracy : 0.7933
##
##        'Positive' Class : FALSE
##
```

```
summary(svmPred)
```

```
## FALSE  TRUE
##  1846   403
```

10. Association Rules



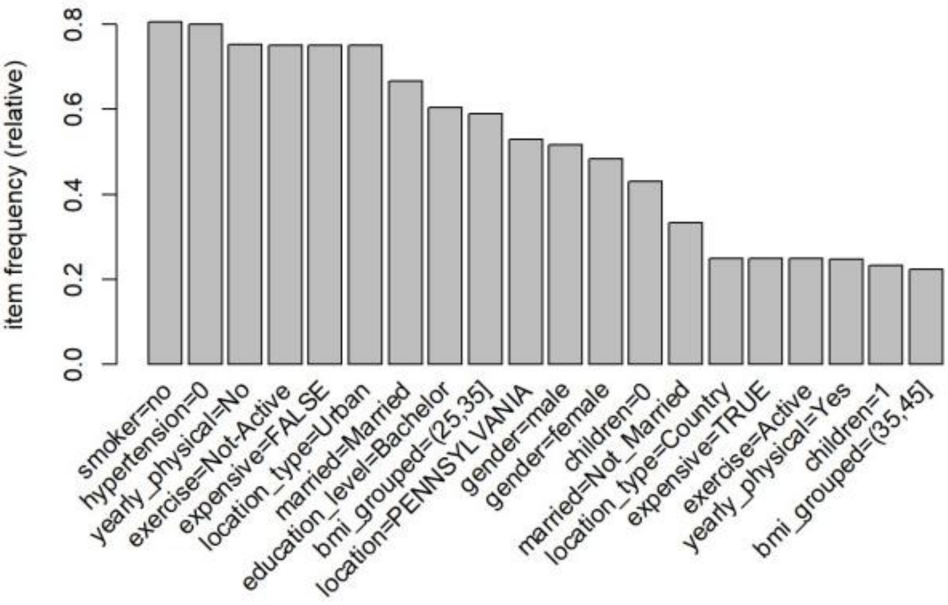
```
##     lhs                             rhs               support confidence   coverage     l
ift count
## [1] {smoker=yes,
##      exercise=Not-Active,
##      gender=male}             => {expensive=TRUE} 0.07344708  0.8555901 0.08584377 3.421
448   551
## [2] {smoker=yes,
##      education_level=Bachelor,
##      exercise=Not-Active}     => {expensive=TRUE} 0.06958144  0.8207547 0.08477739 3.282
144   522
## [3] {smoker=yes,
##      exercise=Not-Active,
##      married=Married}         => {expensive=TRUE} 0.07971208  0.8282548 0.09624100 3.312
136   598
## [4] {smoker=yes,
##      location_type=Urban,
##      exercise=Not-Active}     => {expensive=TRUE} 0.08837643  0.8205446 0.10770461 3.281
303   663
## [5] {smoker=yes,
##      location_type=Urban,
##      exercise=Not-Active,
##      married=Married}         => {expensive=TRUE} 0.06145028  0.8261649 0.07438017 3.303
779   461
## [6] {smoker=yes,
##      exercise=Not-Active,
##      married=Married,
##      hypertension=0}          => {expensive=TRUE} 0.06225007  0.8265487 0.07531325 3.305
314   467
```

Our team came up with 6 association rules we are able to generate association rules for condition where cost is expensive.

Smoker = yes and exercise = Not-Active for all 4 rules.

Hypertension didn't seem to increase health care cost.

Gender and education level are all observed once.



This frequency chart shows the highest attribute values to the lowest attribute values for example in our dataset we have about 80% people who do not smoke followed by people who do not have hypertension and followed by people who are not yearly physically active and so on.

11. Interpretation

Based on the provided plots, the analysis shows how healthcare costs are correlated with age and BMI for different smoking and exercise habits. For smokers, the plots show that healthcare costs tend to increase with age, indicating a positive correlation between age and healthcare costs. This is a significant finding as it suggests that healthcare costs for smokers increase as they age and that early interventions may be necessary to reduce these costs.

The analysis also shows that individuals who are inactive have higher healthcare costs compared to those who are active. This finding indicates a significant correlation between exercise habits and healthcare costs and suggests that promoting active lifestyles could be an effective strategy for reducing healthcare costs.

Regarding BMI, the analysis shows that it has a positive correlation with healthcare costs, with a stronger correlation for smokers. This finding highlights the importance of managing BMI and quitting smoking to reduce healthcare costs. Individuals with high BMI and who smoke are most likely to have higher healthcare costs.

The plots also reveal some anomalies in the data for active individuals when considering the correlation between age, BMI, and healthcare costs. Further investigation may be necessary to understand the reasons behind these anomalies and to identify strategies for reducing healthcare costs for active individuals.

In summary, the provided plots provide valuable insights into the relationship between age, BMI, smoking habits, exercise habits, and healthcare costs. The findings suggest that promoting healthy lifestyles, managing BMI, and quitting smoking are effective strategies for reducing healthcare costs.

12. Conclusion

In conclusion, our analysis has successfully achieved the project goals of predicting which individuals would have high healthcare costs in the next year and providing actionable insights to the HMO on how to lower their total healthcare costs. Our exploratory analysis revealed several variables that strongly correlate with healthcare costs, including age, BMI, and number of children. Additionally, our analysis showed that smokers tend to have significantly higher healthcare costs than non-smokers.

Moreover, our predictive modeling using various algorithms such as SVM, Random Forest, and rpart demonstrated good performance in accurately predicting high-cost healthcare clients. By utilizing this predictive modeling, the HMO could proactively allocate resources and provide healthcare interventions to high-risk individuals, thereby improving patient outcomes while reducing healthcare costs.

In summary, our analysis provides valuable insights into the key drivers of healthcare costs and presents a practical and data-driven solution to the HMO's challenge of reducing healthcare costs. Our findings could have a significant impact on the healthcare industry by enabling organizations to allocate resources effectively, improve patient outcomes, and ultimately reduce healthcare costs.

11. Recommendations

1. Consider charging smokers a higher premium to offset the high cost of healthcare as healthcare expenses are significantly higher for smokers than non-smokers, regardless of age.

2. Promote preventive care: Encourage clients to participate in preventative care by providing rewards for routine check-ups, vaccinations, and screenings. This can lower the total cost of healthcare by allowing for the early detection and treatment of health concerns.

3. It may not be necessary to charge more for someone with a high BMI if premiums are already being raised for older people due to aging as age affects BMI more than smoking does.

4. To incentivize customers to use the health program, HMO can offer discounts on premiums for active participation with personal trainers.

5. HMO can offer a health program that includes a network of personal trainers and automated enrollment for customers in yearly physicals to promote a healthy lifestyle and reduce healthcare costs in the long run.

6. Implement tier-based provider networks: Create tier-based provider networks that point patients in the direction of affordable, high-caliber healthcare providers. HMOs can lower healthcare costs by providing incentives for clients to select providers with a track record of delivering effective care.