THE iSCHOOL
Syracuse University

IST 687: Introduction to Data Science

Final Project Report
Analyzing healthcare cost information from an HMO (Health Management Organization)

Parth Gulavani
Rutu Waghela
Anurag Paradkar
Kishan Rathor
Whitaker Ellis

Professors:
Jeffrey Saltz
Bridgette Jacob

Table of Contents

INTRODUCTION

Objective

We serve as a consulting firm for HMOs (Health Management Organizations), which are medical insurance groups that offer health services in exchange for a set annual charge.

Our objective is to identify the main factors that influence why some people need more medical attention than others, identify those who will spend a lot of money on healthcare in the upcoming year, and offer the HMO specific advice on how to cut costs in order to lower their overall health care expenses

Background

Health Management Organizations (HMOs) are medical insurance groups that offer health care in exchange for a set annual charge. The dataset that we were given has 14 columns and contains data on 7,583 people. The columns broadly focus on several categories, including the individual's unique identifier, age, geographic location, gender, education level, marital status, number of children, and healthcare expenditure. They also ask about the individual's exercise, smoking, BMI, annual physical examination status, and hypertension status.

Based on the facts at hand, we deliver actionable information through our research, and we also successfully anticipate which consumers will spend a lot of money on healthcare. Additionally, we offer some suggestions for reducing health care expenditures based on our

Scope

We have chosen x columns for our analysis, which can be categorised into 3 sections

Individuals Basic Info	
Variable	**Description**
x	Unique Identifier
age	Age of the person at the end of the year
Gender	Gender of the person
education_level	The amount of College Education
married	Marital Status of the individual
num_children	Number of children

Individuals Geographical Information	
Variable	**Description**
location	US States
location_type	Urban or Country

Individual Health Information	
Variable	**Description**
exercise	If the person exercises actively or not
smoker	If the person smokes or not
hypertension	If the person has hypertension or not
bmi	Body Mass Index of the person
yearly_physical	If the person visited their doctor during the year or not
cost	Total healthcare cost for that person, during the past year

BUSINESS QUESTIONS

Initial Business Questions

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 the HMO, in terms of how to lower their total health care costs, by providing a specific recommendation on how to lower health care costs.

Final Business Questions

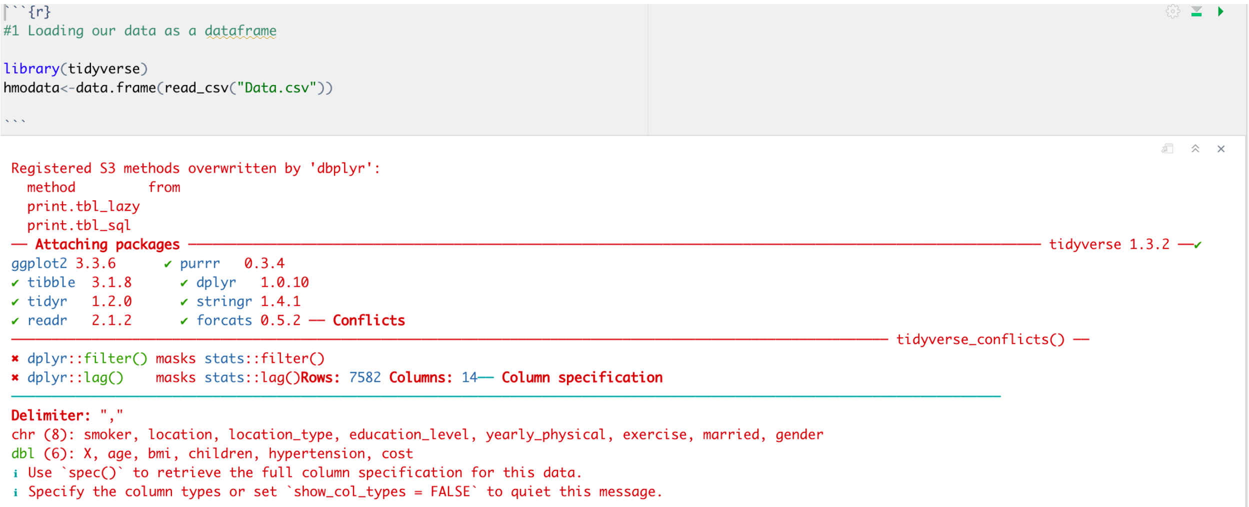
1. How is the health of individuals in the USA overall? (Not sure what is perfect measure of health, can be based on “bmi”)
4. What is the average expenditure for smokers?
5. Is there any relationship between exercising and health of the individual? Are people who exercise often less expensive (in terms of healthcare)?
7. What is the average expenditure of people who are married with kids in comparison to people who are single?

DATA ANALYSIS

Data Acquisition

Professor provided us with a link with the Dataset, we had to copy the dataset and create a .csv file out of it. The dataset contains healthcare cost information from an HMO (Health Management Organization). This data set has a .csv file with 14 columns

We imported the dataset in R Studio, using read_csv( ) function into a new data frame named “hmodata”



~~~
```{r}
#1 Loading our data as a dataframe

library(tidyverse)
hmodata<-data.frame(read_csv("Data.csv"))

```
~~~

```
Registered S3 methods overwritten by 'dbplyr':
  method         from
  print.tbl_lazy
  print.tbl_sql
── Attaching packages ─────────────────────────────────────────────────────────── tidyverse 1.3.2 ──✔
ggplot2 3.3.6      ✔ purrr   0.3.4
✔ tibble  3.1.8      ✔ dplyr   1.0.10
✔ tidyr   1.2.0      ✔ stringr 1.4.1
✔ readr   2.1.2      ✔ forcats 0.5.2 ── Conflicts
──────────────────────────────────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()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.
```

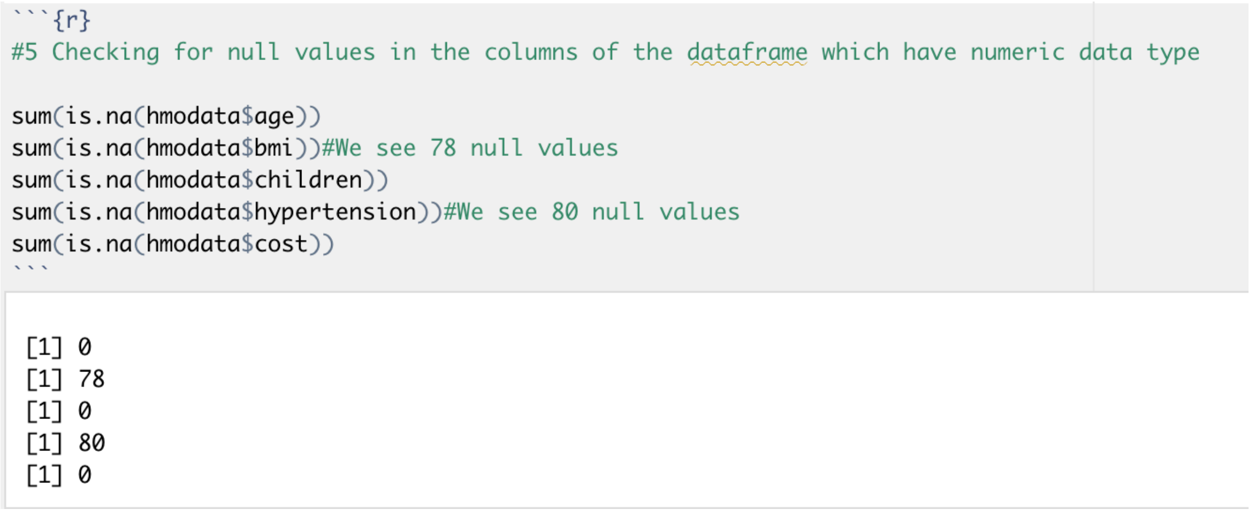
Data Cleansing

There might be values missing

Missing values in datasets are another challenge in the healthcare industry. Often, some values in certain features will not be present. This is because doctors sometimes don’t take all necessary lab measurements, or because the data has been lost. e.g. ‘?’, ‘n/a’, ‘0’, ‘-9’

Hence we checked for null values in the columns of the data frame which have numeric data type and found

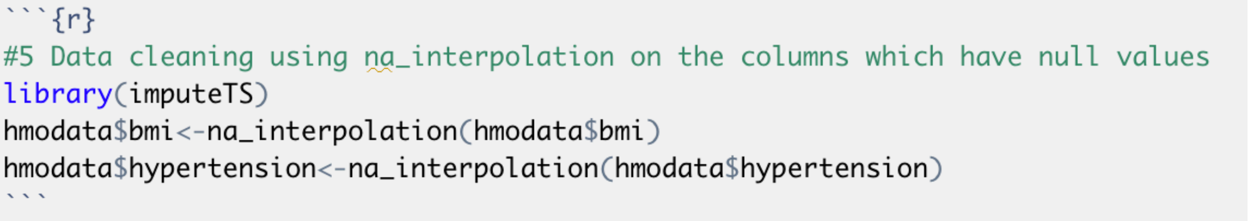
The column for BMI has 78 null values & the column for Hypertension has 80 null values. Refer the screenshot attached below.



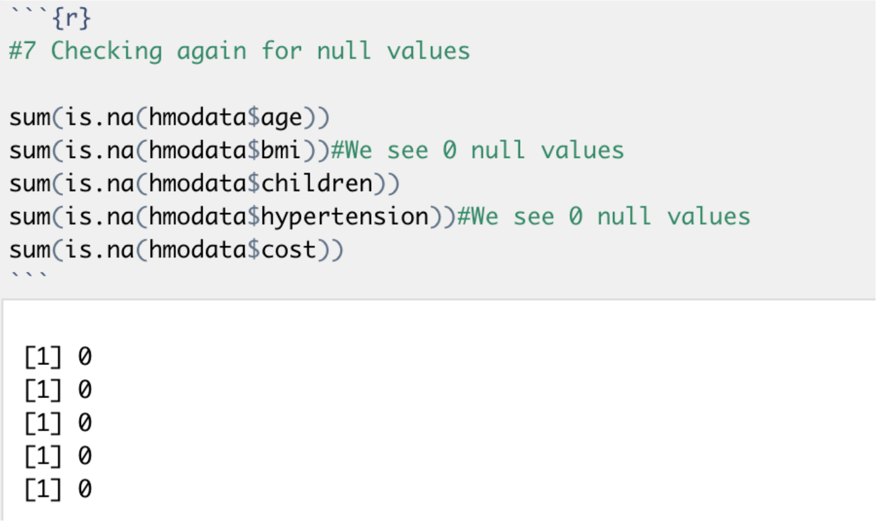
~~~
```{r}
#5 Checking for null values in the columns of the dataframe which have numeric data type

sum(is.na(hmodata$age))
sum(is.na(hmodata$bmi))#We see 78 null values
sum(is.na(hmodata$children))
sum(is.na(hmodata$hypertension))#We see 80 null values
sum(is.na(hmodata$cost))
```
~~~

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

Hence we used “na_interpolation” on the “bmi” & “hypertension” columns to get rid of the null values, which can be referred by the screenshot attached below.
After we cleaned the data using na_interpolation, we checked for the null values again, which were now 0



~~~
```{r}
#5 Data cleaning using na_interpolation on the columns which have null values
library(imputeTS)
hmodata$bmi<-na_interpolation(hmodata$bmi)
hmodata$hypertension<-na_interpolation(hmodata$hypertension)
```
~~~



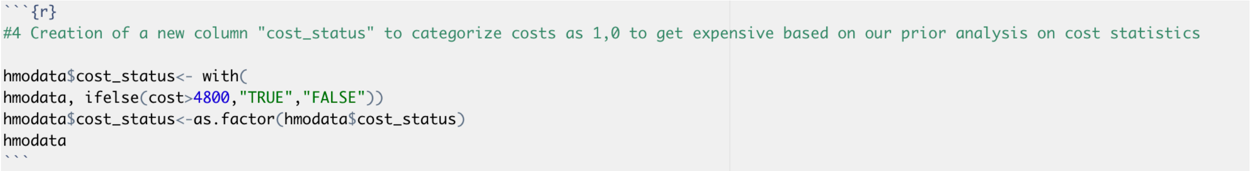
~~~
```{r}
#7 Checking again for null values

sum(is.na(hmodata$age))
sum(is.na(hmodata$bmi))#We see 0 null values
sum(is.na(hmodata$children))
sum(is.na(hmodata$hypertension))#We see 0 null values
sum(is.na(hmodata$cost))
```
~~~

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

Values in Data Can Be Incorrect
To quickly determine whether there are any incorrect values in the dataset, one of the methods is to use “Pandas function df.describe()” to see statistical properties of a certain feature. It works well for numerical features. Data Transformation

Data Transformation

We created a new column, named cost_status, where cost > 4800 is assigned value 1, else 0.

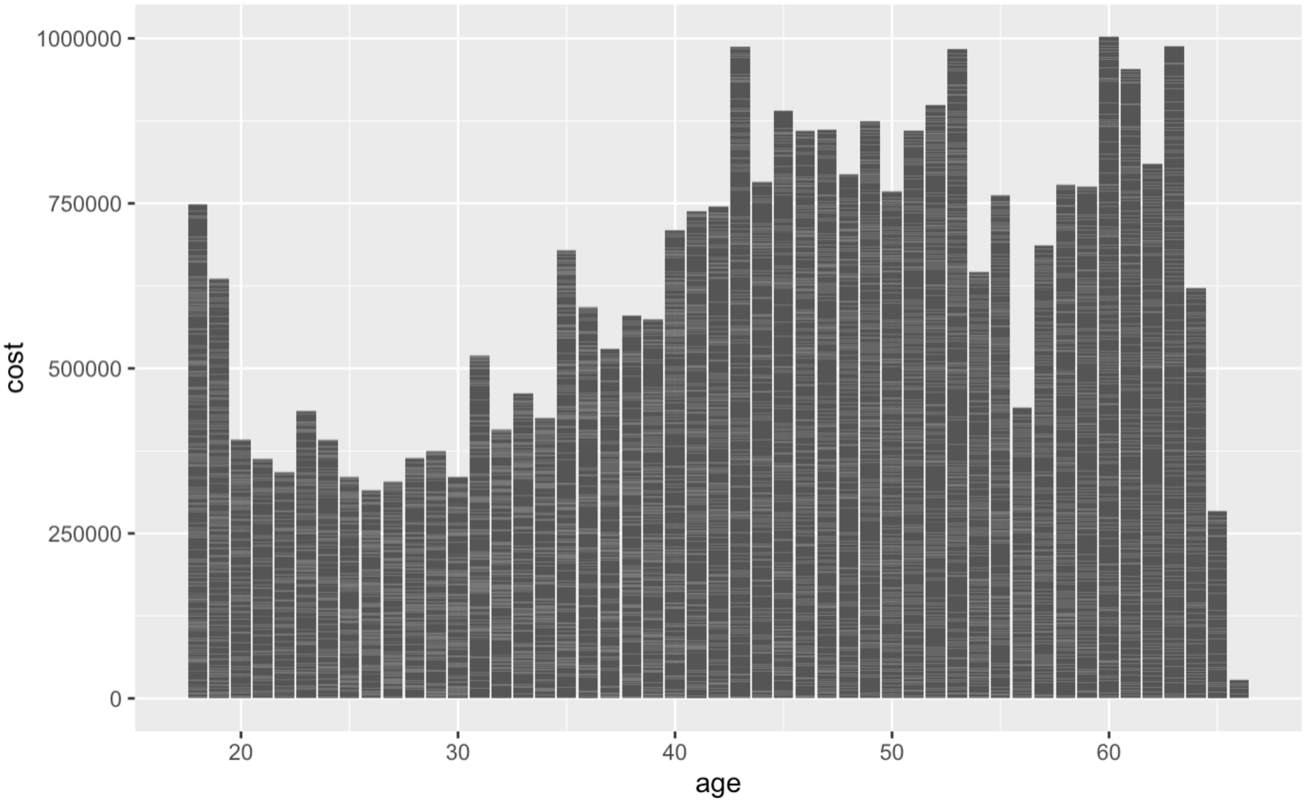


~~~
```{r}
#4 Creation of a new column "cost_status" to categorize costs as 1,0 to get expensive based on our prior analysis on cost statistics

hmodata$cost_status<- with(
hmodata, ifelse(cost>4800,"TRUE","FALSE"))
hmodata$cost_status<-as.factor(hmodata$cost_status)
hmodata
```
~~~

DESCRIPTIVE STATISTICS & VISUALISATIONS

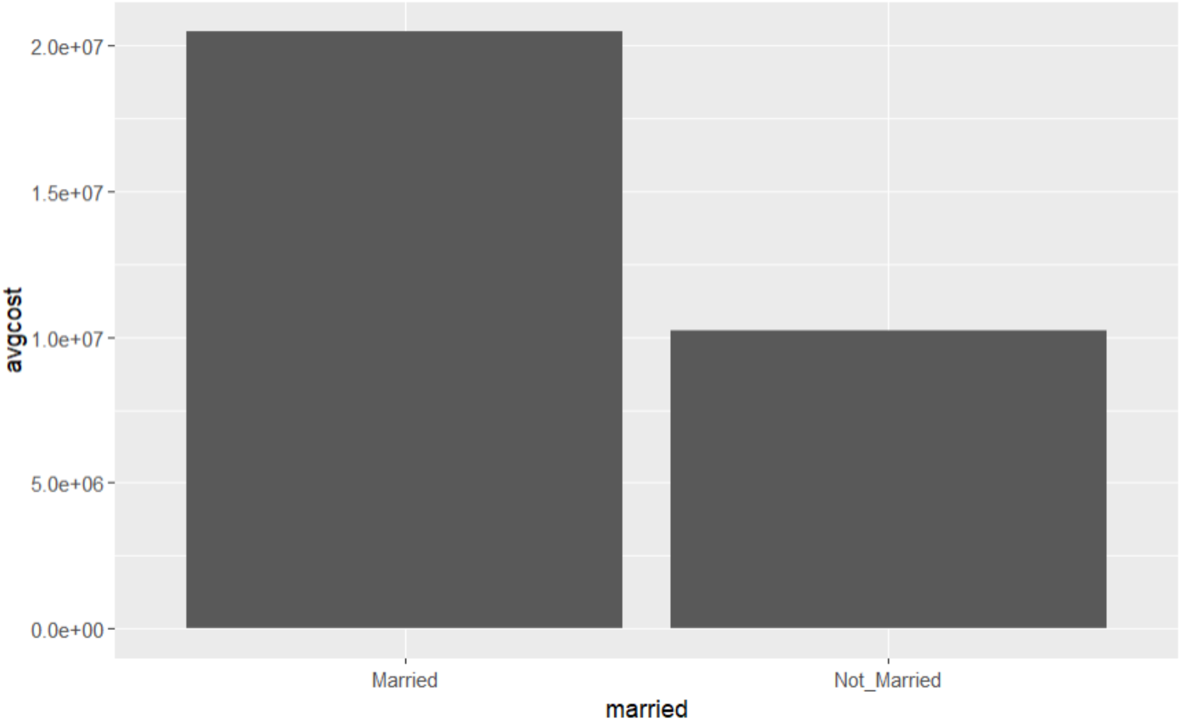
We first analyzed the dataset and visualized it for understanding

Age v/s Cost Bar Plot



And we noted down that, costs are high in teen years, dips down significantly for young adults, and then gradually increase with the age.

Married v/s Single Bar Plot

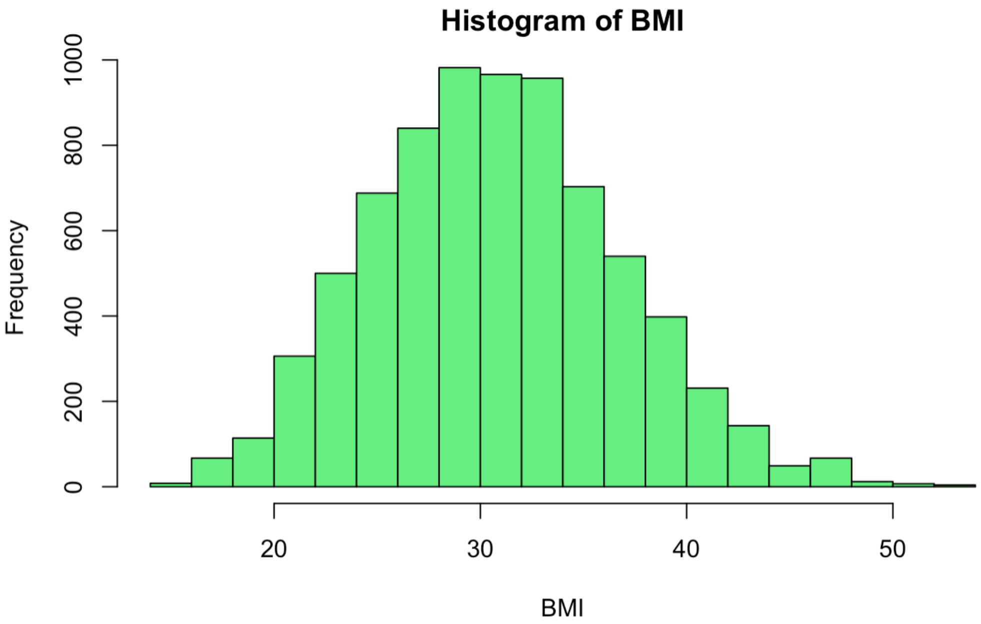


We analyzed that the average expenditure for Married people is more than 2.0e+07, whereas average expenditure for Singles is slightly above 1.0e+07.

Histograms to see distribution of quantitative variables

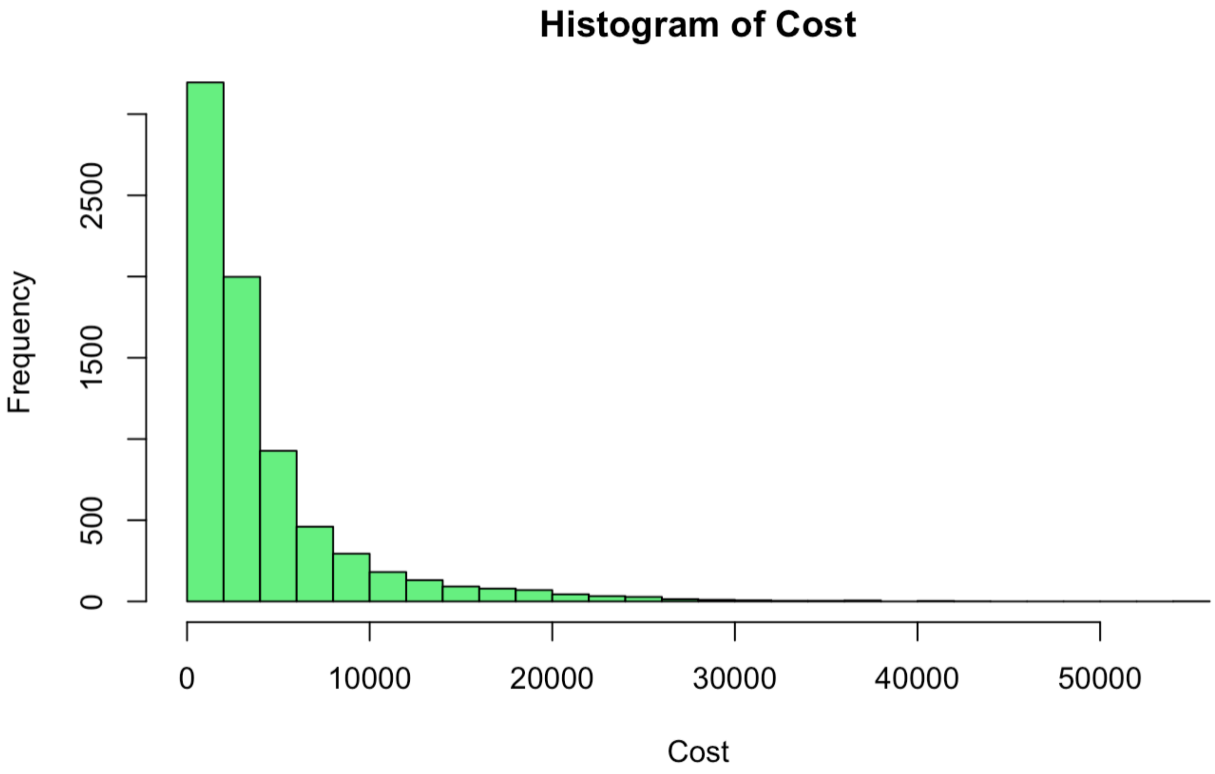
Since BMI and cost are numeric variables, we plotted Histograms for them

- **BMI**



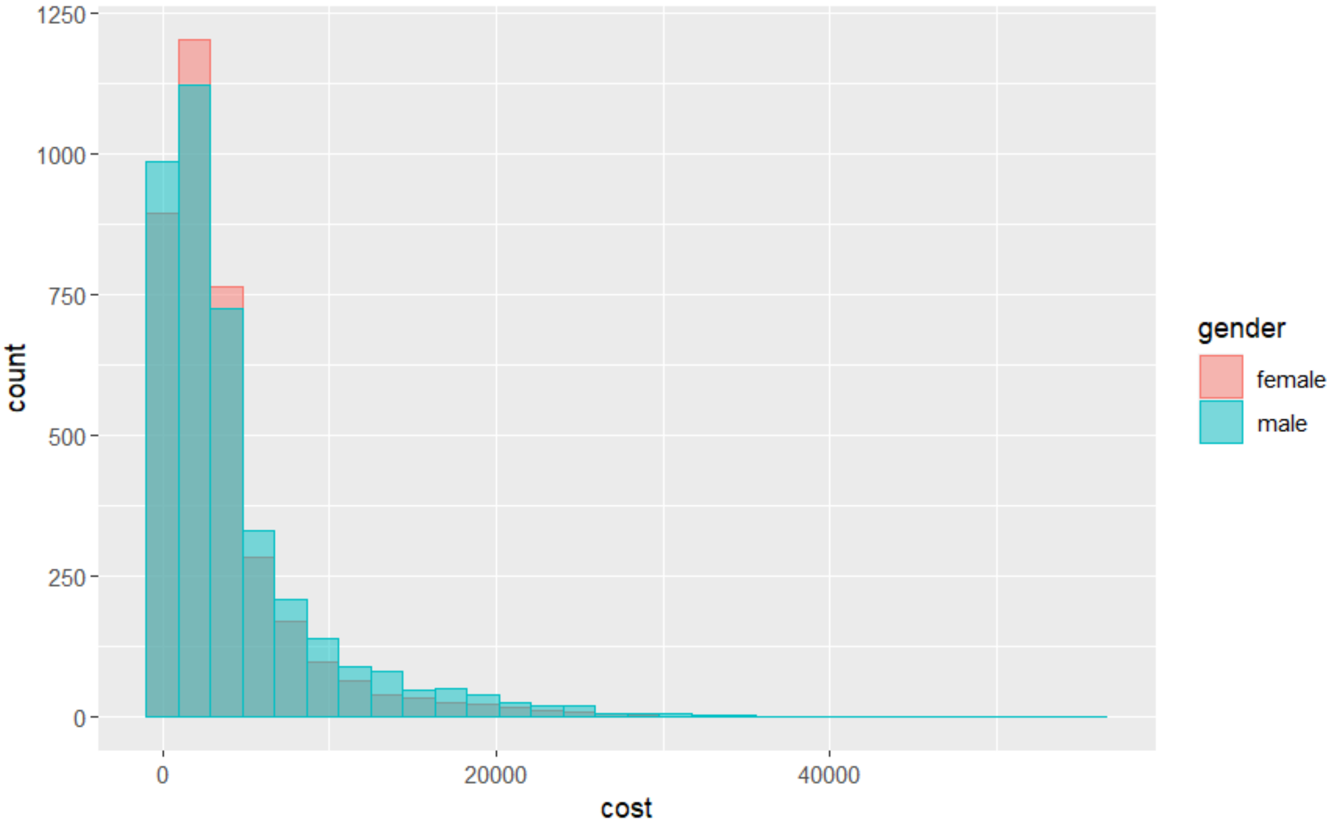
In Histogram for BMI, we can see a normal distribution here. Most of the values cluster in the middle of the range that is, somewhere around 30, and the rest taper off symmetrically toward either extreme.

- **Cost**



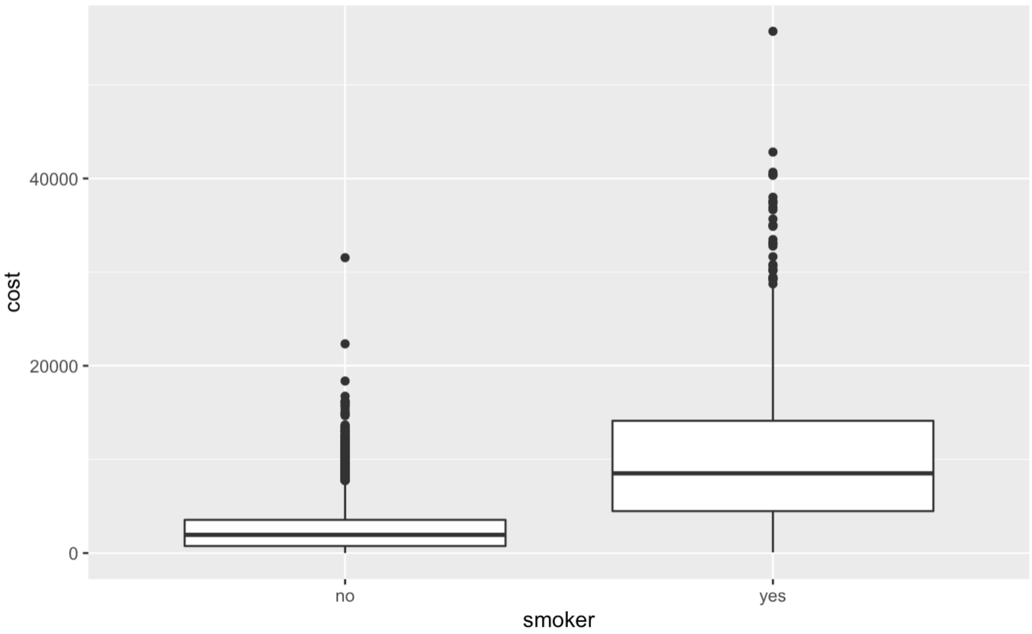
When we plot a Histogram for Cost, we see a right skewed distribution that is, the peak of the graph lies to the left side of the center. Which means, Individuals with significantly higher cost are less in number.

- **Male v/s Female Expenditure**



When we plotted the Male v/s Female Histogram we analyzed that,
Healthcare will be more expensive for males.

Box plot to determine the outliers

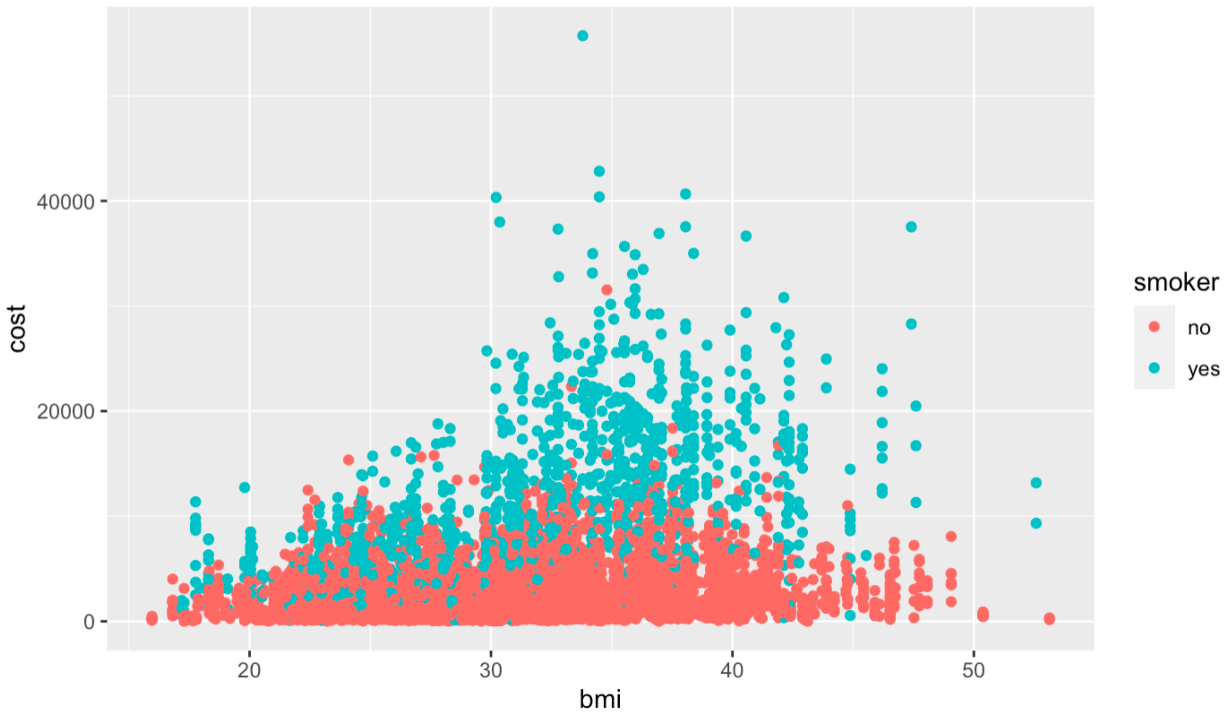


By plotting the Box Plot, we can clearly see that the costs for smokers are significantly higher than those for non smokers.

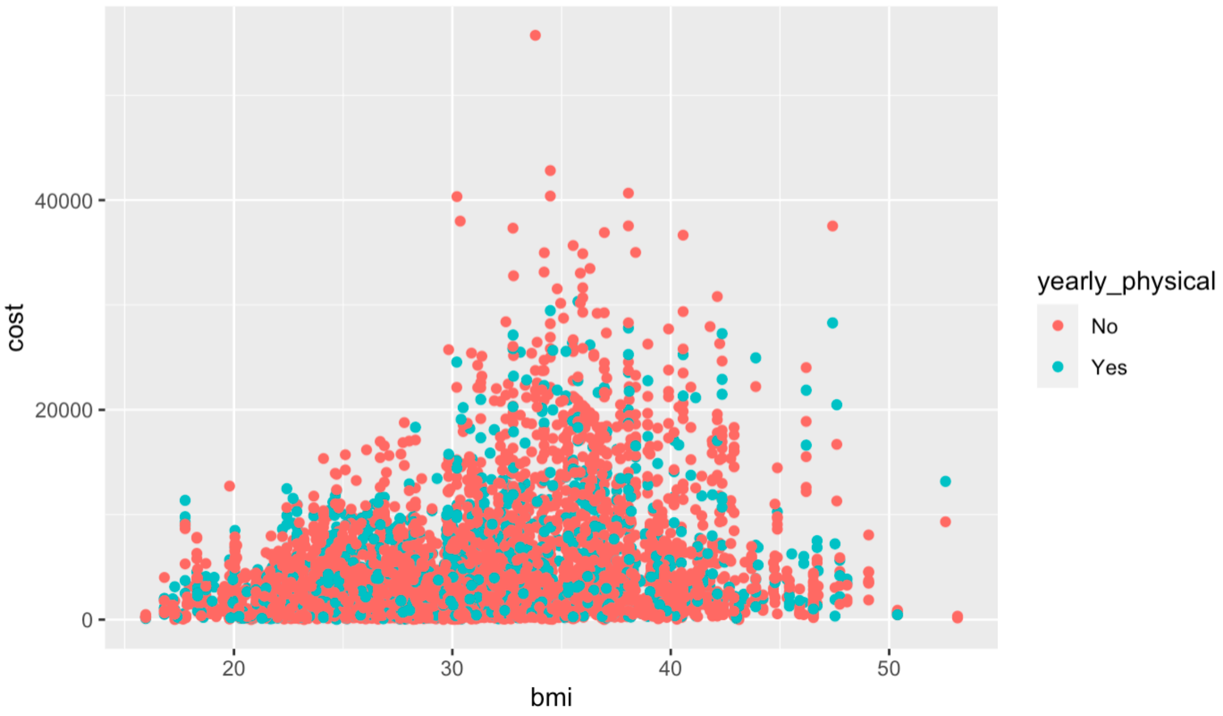
Scatterplots

Followed by this, we plotted scatter plots to see, how Cost & BMI are affected by

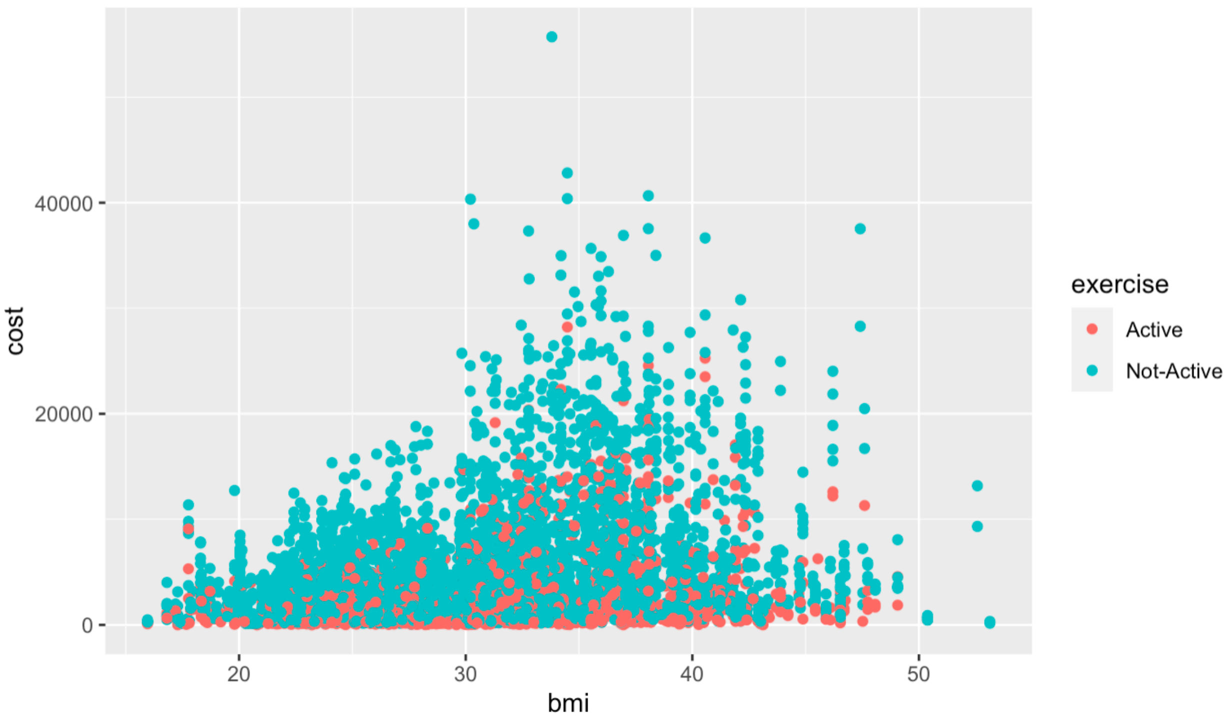
- **Smoking**



- **Getting Physically Tested regularly**



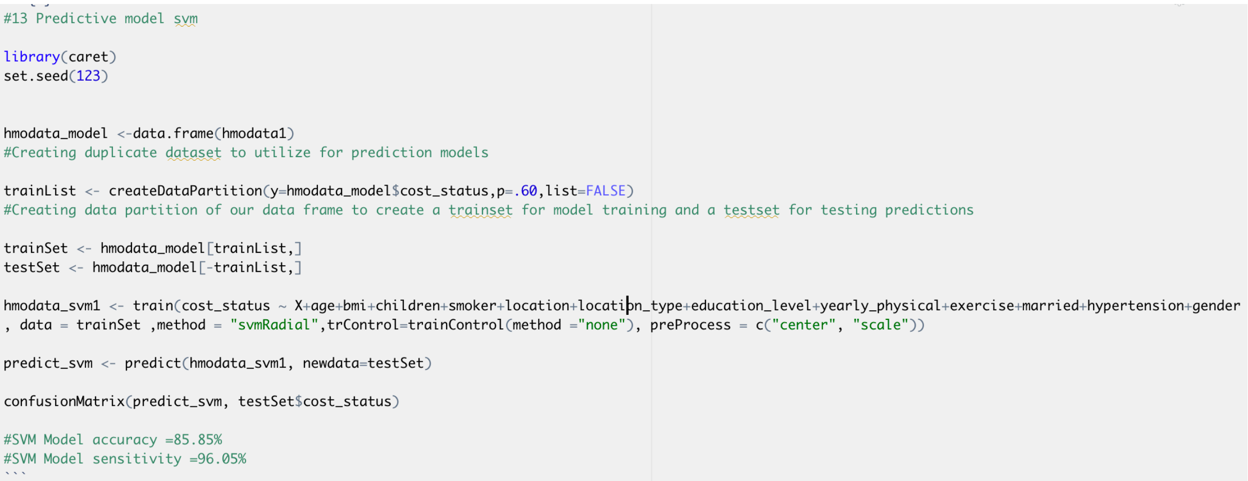
- **Exercising**



USE OF MODELING TECHNIQUES

First, we created a duplicate dataset from the original dataset to use for model training and then we applied Predictive models viz., svm, kscm & rpart models. To make our predictions.

SVM Model

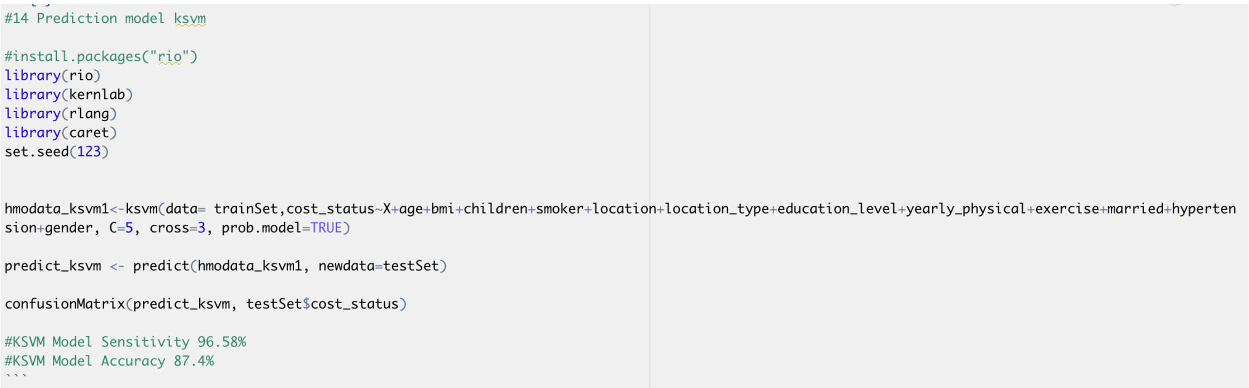


~~~
#13 Predictive model svm

library(caret)
set.seed(123)


hmodata_model <-data.frame(hmodata1)
#Creating duplicate dataset to utilize for prediction models

trainList <- createDataPartition(y=hmodata_model$cost_status,p=.60,list=FALSE)
#Creating data partition of our data frame to create a trainset for model training and a testset for testing predictions

trainSet <- hmodata_model[trainList,]
testSet <- hmodata_model[-trainList,]

hmodata_svm1 <- train(cost_status ~ X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical+exercise+married+hypertension+gender
, data = trainSet ,method = "svmRadial",trControl=trainControl(method ="none"), preProcess = c("center", "scale"))

predict_svm <- predict(hmodata_svm1, newdata=testSet)

confusionMatrix(predict_svm, testSet$cost_status)

#SVM Model accuracy =85.85%
#SVM Model sensitivity =96.05%
```
~~~

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  2205  308
     TRUE     74  445

               Accuracy : 0.874
                 95% CI : (0.8617, 0.8856)
    No Information Rate : 0.7516
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6234

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

            Sensitivity : 0.9675
            Specificity : 0.5910
         Pos Pred Value : 0.8774
         Neg Pred Value : 0.8574
             Prevalence : 0.7516
         Detection Rate : 0.7272
   Detection Prevalence : 0.8288
      Balanced Accuracy : 0.7792

       'Positive' Class : FALSE
```

- From this we can say that for our SVM model we consider these attributes from our dataset to predict cost status.
- We implemented a SVM radial method using the general SVM function.
- With this model we get:
- Accuracy of 85.88%
- Sensitivity of 96.84%

KSVM Model

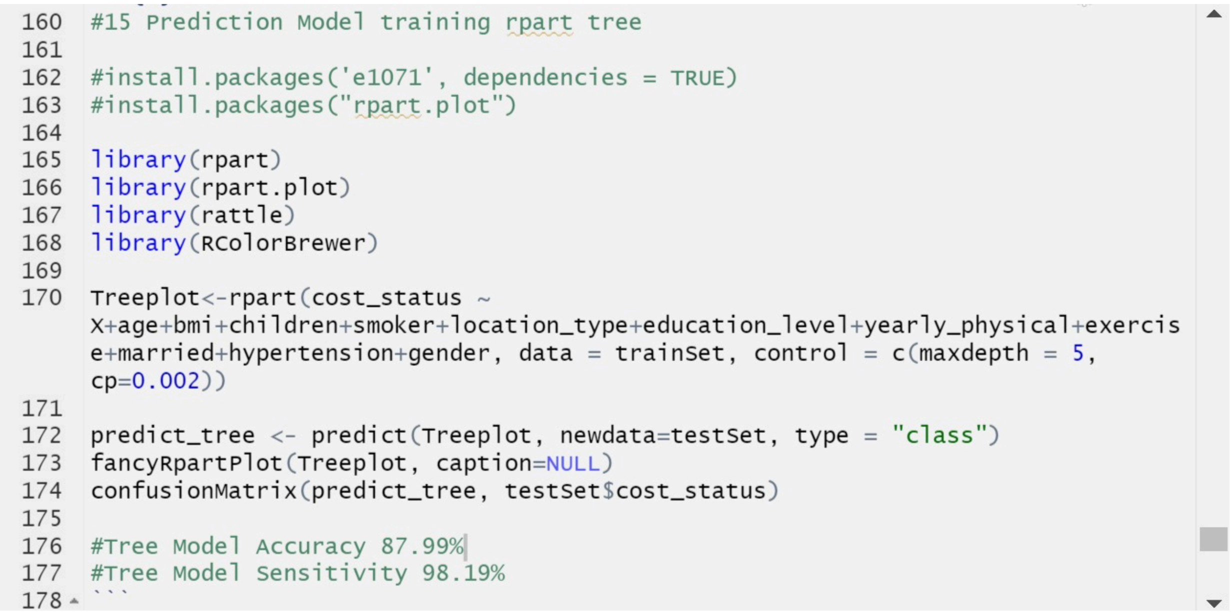


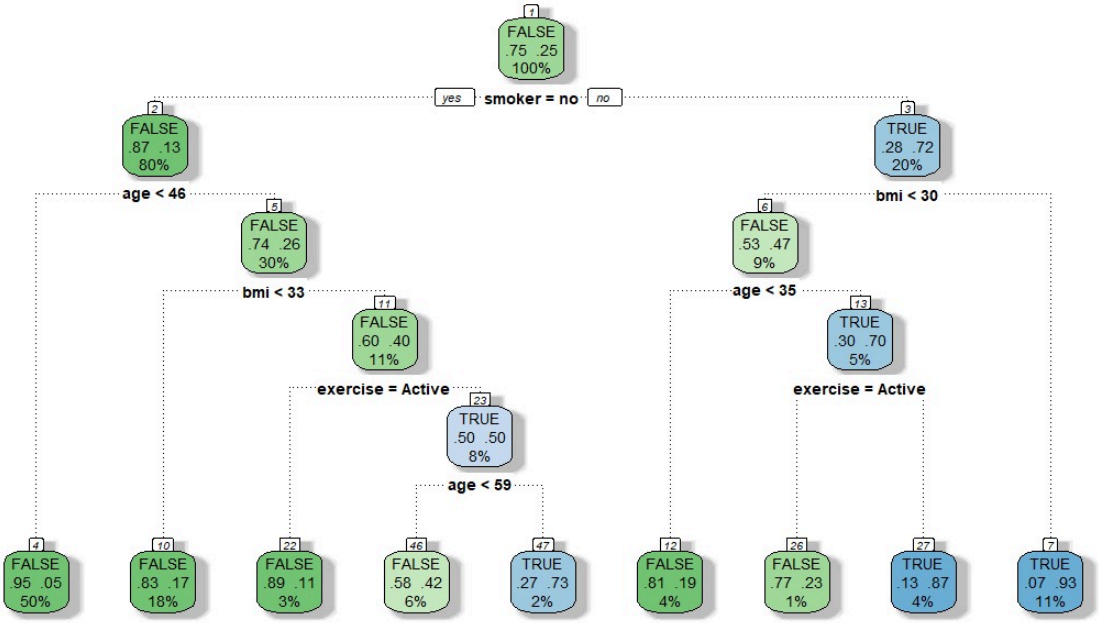
~~~
#14 Prediction model ksvm

#install.packages("rio")
library(rio)
library(kernlab)
library(rlang)
library(caret)
set.seed(123)


hmodata_ksvm1<-ksvm(data= trainSet,cost_status~X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical+exercise+married+hyperten
sion+gender, C=5, cross=3, prob.model=TRUE)

predict_ksvm <- predict(hmodata_ksvm1, newdata=testSet)

confusionMatrix(predict_ksvm, testSet$cost_status)

#KSVM Model Sensitivity 96.58%
#KSVM Model Accuracy 87.4%
```
~~~

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  2205  308
     TRUE     74  445

               Accuracy : 0.874
                 95% CI : (0.8617, 0.8856)
    No Information Rate : 0.7516
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6234

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

            Sensitivity : 0.9675
            Specificity : 0.5910
         Pos Pred Value : 0.8774
         Neg Pred Value : 0.8574
             Prevalence : 0.7516
         Detection Rate : 0.7272
   Detection Prevalence : 0.8288
      Balanced Accuracy : 0.7792

       'Positive' Class : FALSE
```

- A regression model to predict how an output is predicted based on other variables in the data set
- We used KSVM for our prediction model.
- This resulted in a model sensitivity of 97.66% and a model accuracy of 87.73%

Rpart Model



```
#15 Prediction Model training rpart tree

#install.packages('e1071', dependencies = TRUE)
#install.packages("rpart.plot")

library(rpart)
library(rpart.plot)
library(rattle)
library(RColorBrewer)

Treeplot<-rpart(cost_status ~
X+age+bmi+children+smoker+location_type+education_level+yearly_physical+exercis
e+married+hypertension+gender, data = trainSet, control = c(maxdepth = 5,
cp=0.002))

predict_tree <- predict(Treeplot, newdata=testSet, type = "class")
fancyRpartPlot(Treeplot, caption=NULL)
confusionMatrix(predict_tree, testSet$cost_status)

#Tree Model Accuracy 87.99%
#Tree Model Sensitivity 98.19%
```
```


```

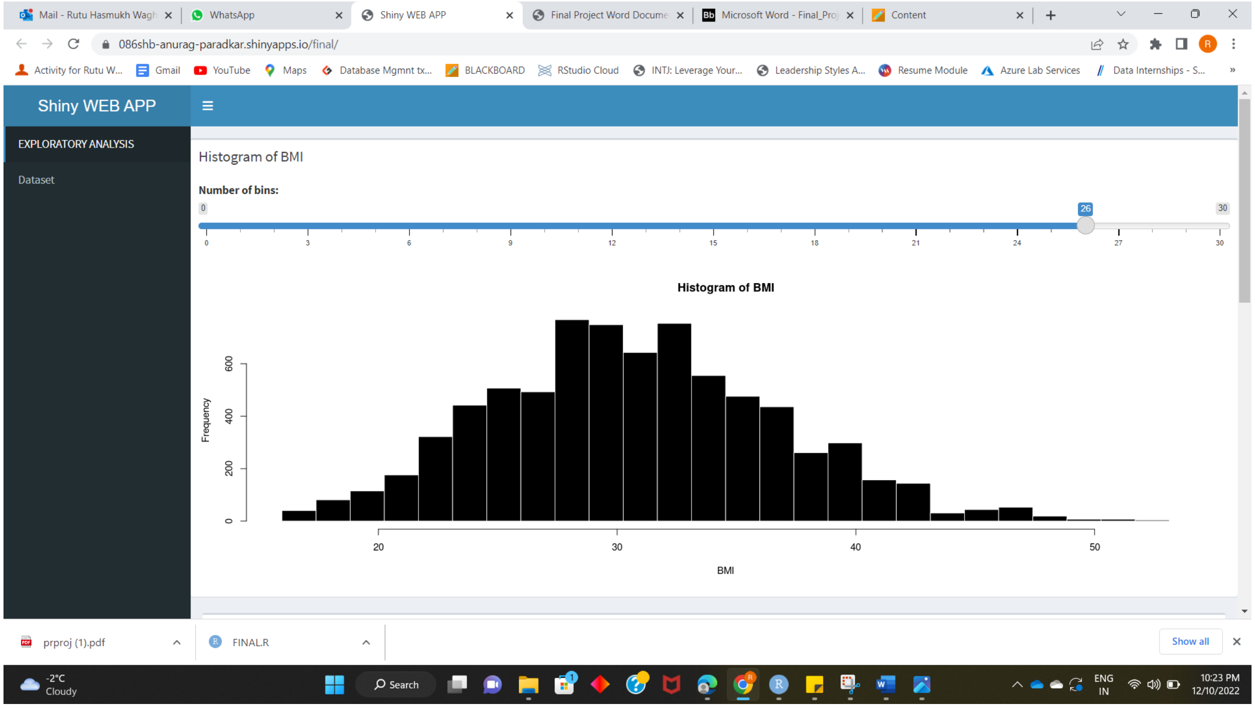
- We used the rpart and rpart.plot packages to create this model
- This model had 87.99% accuracy and 98.19% sensitivity.

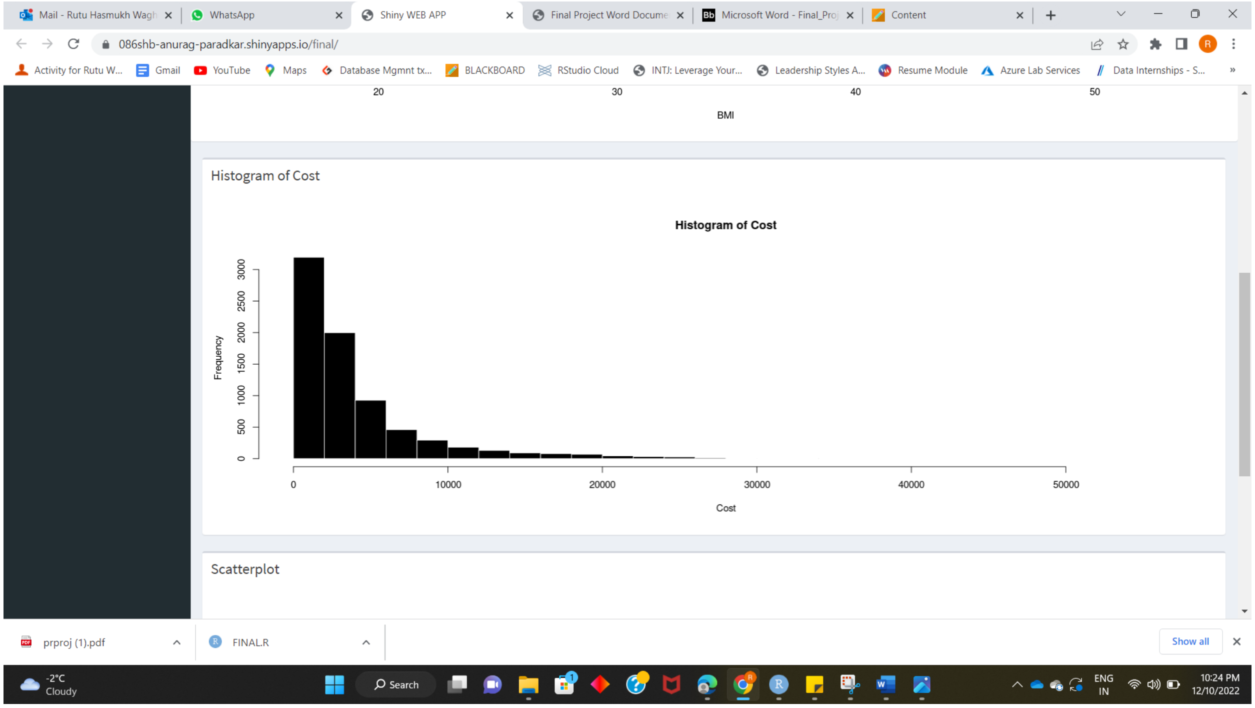
ACTIONABLE INSIGHTS

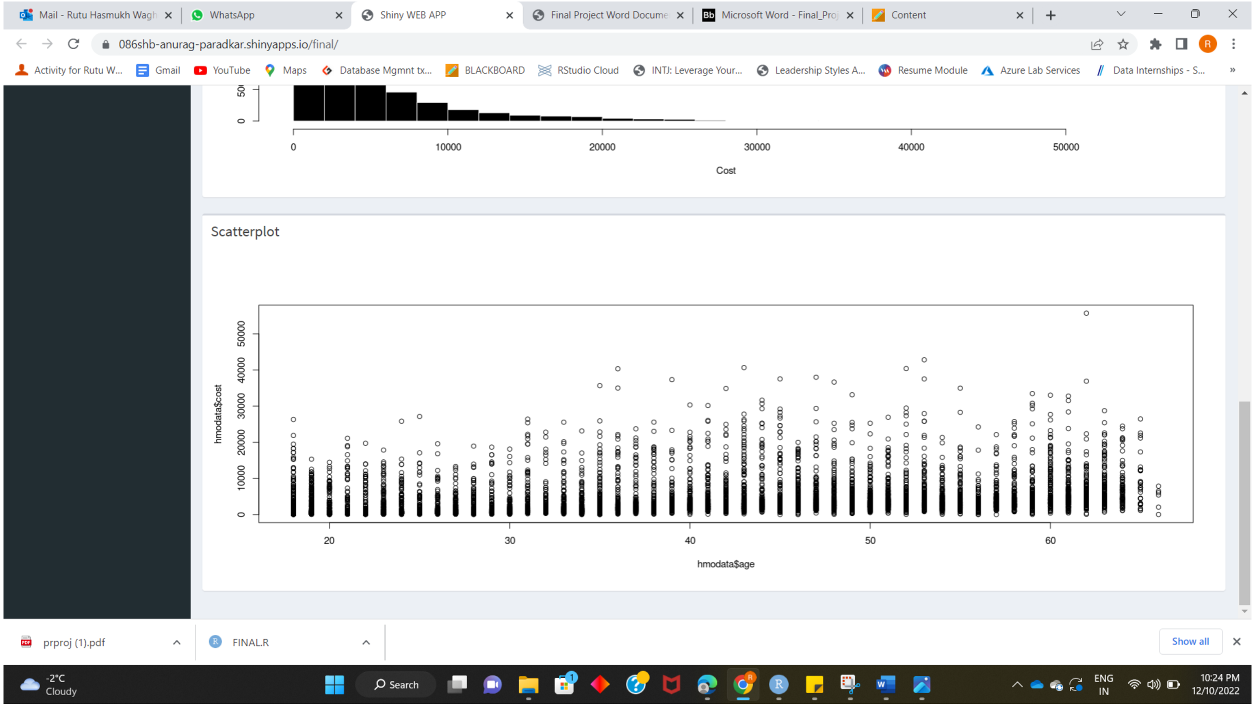
- We can see from the graph that Younger adults have higher costs, and hence we can start offering them free health checkups to promote the hmo and lower their cost by taking the required steps in their teens itself
- As we have seen from our scatterplots, individuals who are smokers and don't exercise have significantly higher costs on healthcare when compared with costs for non-smokers. Therefore, we can make a recommendation to suggest smokers join a partnered gym giving heavy discounted rates, motivating individuals to adopt a healthy lifestyle and minimize their cigarette consumption.
- From our analysis we can see that people who undergo yearly physical exams have comparatively lower costs hence we can suggest mandatory yearly physical health checkups to get an understanding of any factors that could be harmful to individuals and help prevent any such circumstances from occurring beforehand

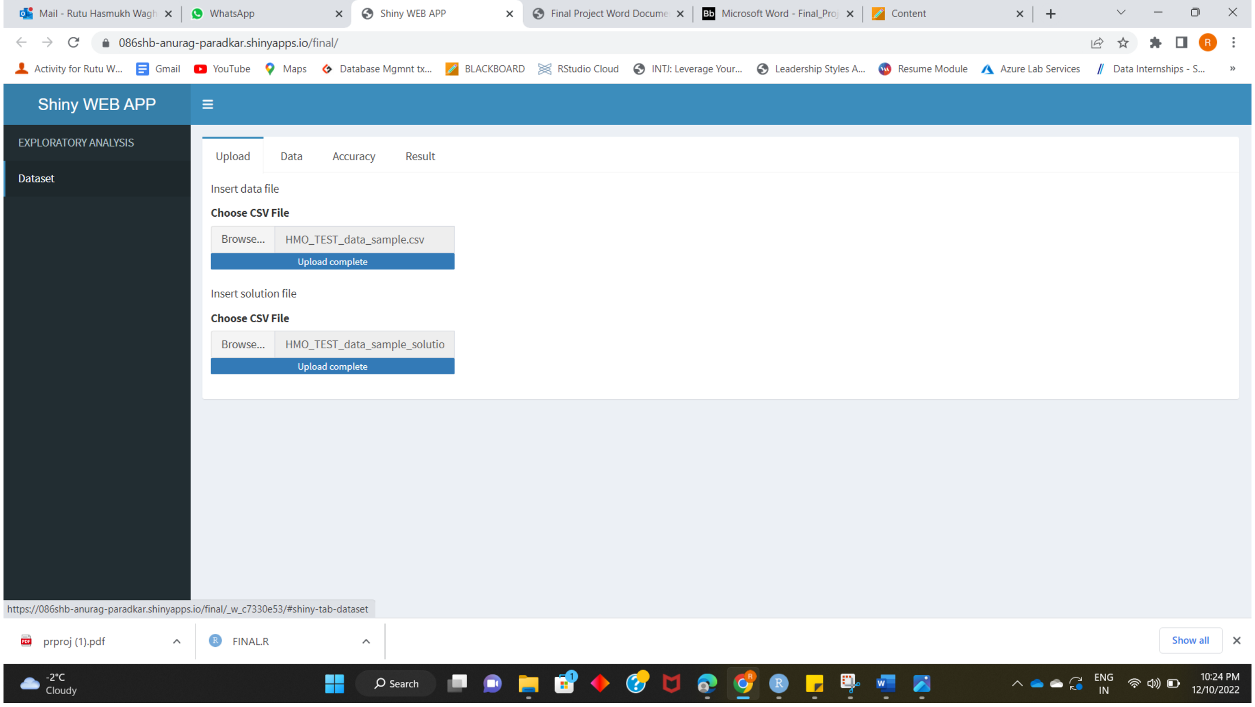
SHINY APP

Screenshots




Mail - Rutu Hasmukh Wagh
WhatsApp
Shiny WEB APP
Final Project Word Docume
Microsoft Word - Final_Proj
Content
086shb-anurag-paradkar.shinyapps.io/final/
Activity for Rutu W...
Gmail
YouTube
Maps
Database Mgmnt tx...
BLACKBOARD
RStudio Cloud
INTJ: Leverage Your...
Leadership Styles A...
Resume Module
Azure Lab Services
Data Internships - S...
20
30
40
50
BMI
Histogram of Cost
Histogram of Cost
Frequency
0
500
1000
1500
2000
2500
3000
0
10000
20000
30000
40000
50000
Cost
Scatterplot
prproj (1).pdf
FINAL.R
Show all
-2°C
Cloudy
Search
ENG
IN
10:24 PM
12/10/2022


Mail - Rutu Hasmukh Wagh
WhatsApp
Shiny WEB APP
Final Project Word Docume
Microsoft Word - Final_Proj
Content
086shb-anurag-paradkar.shinyapps.io/final/
Activity for Rutu W...
Gmail
YouTube
Maps
Database Mgmnt tx...
BLACKBOARD
RStudio Cloud
INTJ: Leverage Your...
Leadership Styles A...
Resume Module
Azure Lab Services
Data Internships - S...
500
0
0
10000
20000
30000
40000
50000
Cost
Scatterplot
hmodata$cost
0
10000
20000
30000
40000
50000
20
30
40
50
60
hmodata$age
prproj (1).pdf
FINAL.R
Show all
-2°C
Cloudy
Search
ENG
IN
10:24 PM
12/10/2022



Insert data file

Choose CSV File

Browse... HMO_TEST_data_sample.csv

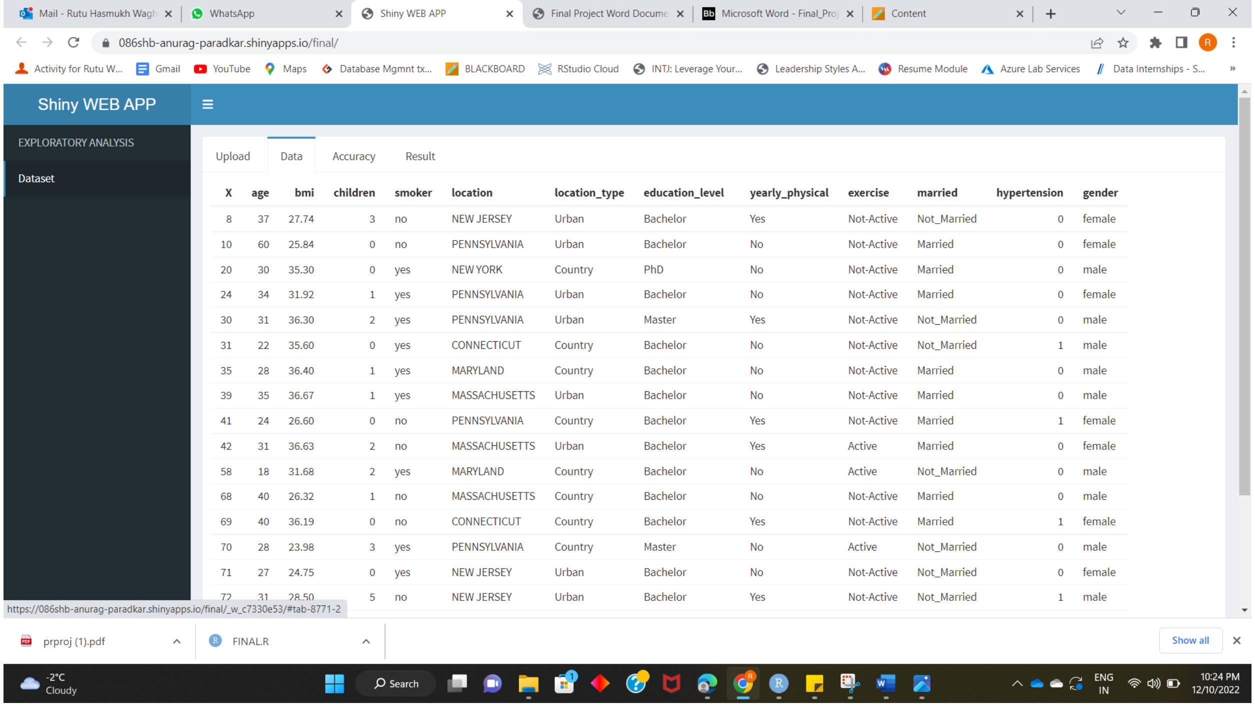
Upload complete

Insert solution file

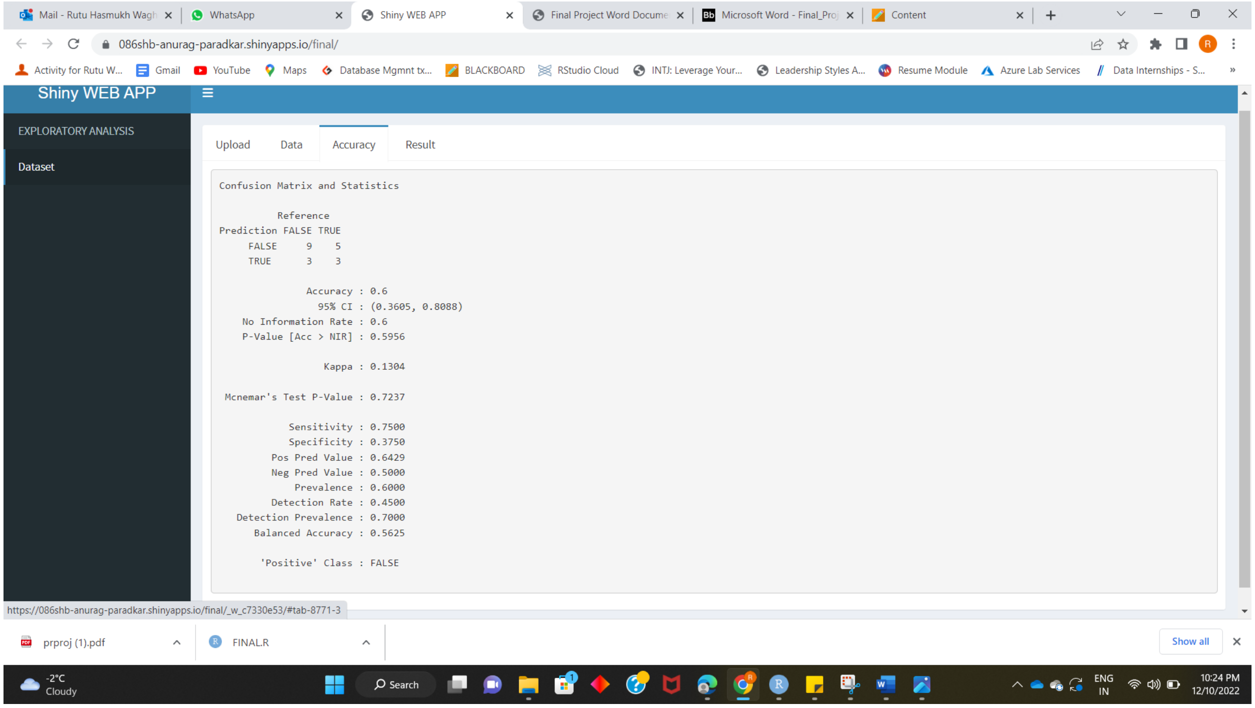
Choose CSV File

Browse... HMO_TEST_data_sample_solutio

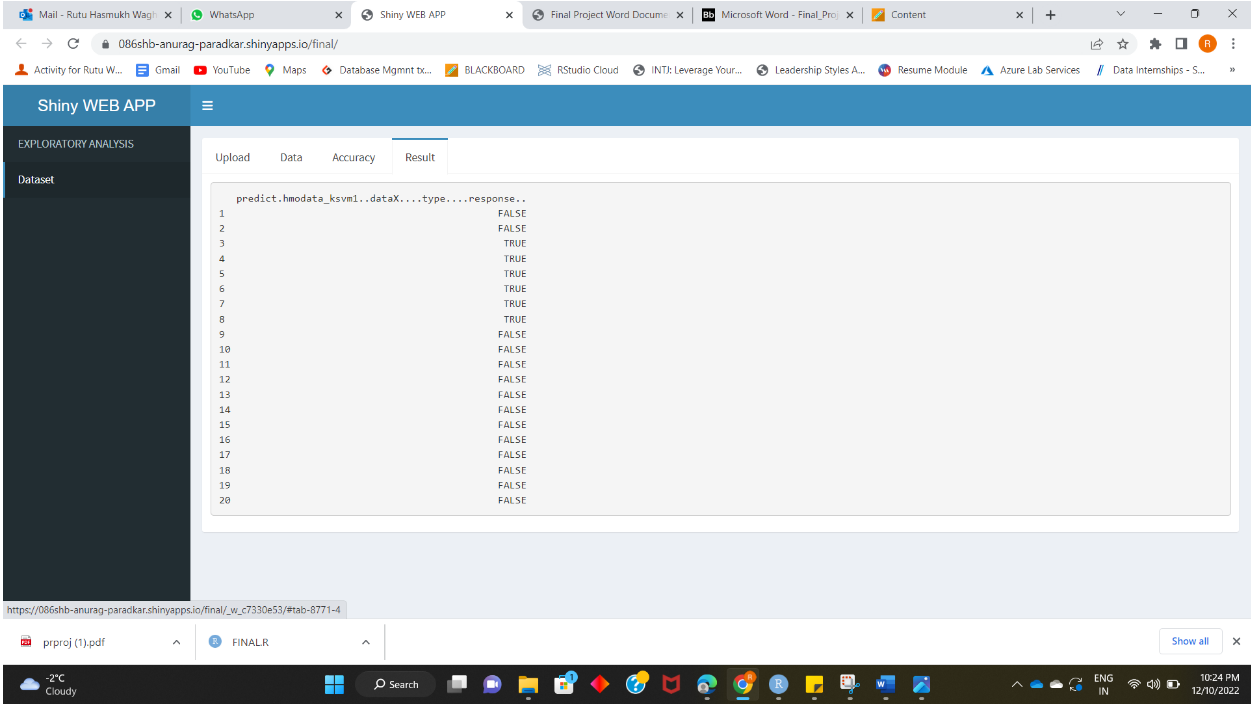
Upload complete



X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertension	gender
8	37	27.74	3	no	NEW JERSEY	Urban	Bachelor	Yes	Not-Active	Not_Married	0	female
10	60	25.84	0	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0	female
20	30	35.30	0	yes	NEW YORK	Country	PhD	No	Not-Active	Married	0	male
24	34	31.92	1	yes	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	0	female
30	31	36.30	2	yes	PENNSYLVANIA	Urban	Master	Yes	Not-Active	Not_Married	0	male
31	22	35.60	0	yes	CONNECTICUT	Country	Bachelor	No	Not-Active	Not_Married	1	male
35	28	36.40	1	yes	MARYLAND	Country	Bachelor	No	Not-Active	Married	0	male
39	35	36.67	1	yes	MASSACHUSETTS	Urban	Bachelor	No	Not-Active	Married	0	male
41	24	26.60	0	no	PENNSYLVANIA	Country	Bachelor	Yes	Not-Active	Married	1	female
42	31	36.63	2	no	MASSACHUSETTS	Urban	Bachelor	Yes	Active	Married	0	female
58	18	31.68	2	yes	MARYLAND	Country	Bachelor	No	Active	Not_Married	0	male
68	40	26.32	1	no	MASSACHUSETTS	Country	Bachelor	No	Not-Active	Married	0	male
69	40	36.19	0	no	CONNECTICUT	Country	Bachelor	Yes	Not-Active	Married	1	female
70	28	23.98	3	yes	PENNSYLVANIA	Country	Master	No	Active	Not_Married	0	male
71	27	24.75	0	yes	NEW JERSEY	Urban	Bachelor	No	Not-Active	Not_Married	0	female
72	31	28.50	5	no	NEW JERSEY	Urban	Bachelor	Yes	Not-Active	Not_Married	1	male



```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE     9    5
     TRUE      3    3

               Accuracy : 0.6
                 95% CI : (0.3605, 0.8088)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.5956

                  Kappa : 0.1304

 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.7500
            Specificity : 0.3750
         Pos Pred Value : 0.6429
         Neg Pred Value : 0.5000
             Prevalence : 0.6000
         Detection Rate : 0.4500
   Detection Prevalence : 0.7000
      Balanced Accuracy : 0.5625

       'Positive' Class : FALSE
```



```
   predict.hmodata_ksvm1..dataX....type....response..
1                                               FALSE
2                                               FALSE
3                                                TRUE
4                                                TRUE
5                                                TRUE
6                                                TRUE
7                                                TRUE
8                                                TRUE
9                                               FALSE
10                                              FALSE
11                                              FALSE
12                                              FALSE
13                                              FALSE
14                                              FALSE
15                                              FALSE
16                                              FALSE
17                                              FALSE
18                                              FALSE
19                                              FALSE
20                                              FALSE
```

Link

https://086shb-anurag-paradkar.shinyapps.io/final/