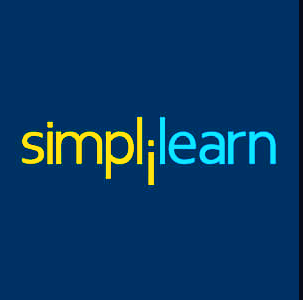
|  |
| --- |
| Project Report  Data Science with R |
|  |
| June 15  Project 7  Submitted by: ARKA PRAVA PANDA |



CONTENTS

1. Cover page ………………………………………………………………………. 1
2. Contents ……………………………………………………………………… 2
3. Obj. and description ………………………………………………………………………… 3
4. Required business queries …………………………………………………………………. 4
5. Project code ………………………………………………………………………………………. 5-8
6. Analysis ………………………………………………………………………………………………. 9-14

# Objective and Background

A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyze the data to research on healthcare costs and their utilization.

# Description and Dataset details

Based on the given dataset, we will have to perform statistical analysis and infer conclusions based on the business queries. The dataset details and descriptions are as follows:-

|  |  |
| --- | --- |
| Attribute | Description |
| Age | Age of the patient discharged |
| Female | A binary variable that indicates if the patient is female |
| Los | Length of stay in days |
| Race | Race of the patient (specified numerically) |
| Totchg | Hospital discharge costs |
| Aprdrg | All Patient Refined Diagnosis Related Groups |

Required Business queries

The following describes the analysis that has to be performed

1. To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

3. To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.

4. To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.

5. Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

6. To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.

# Project code

getwd()

setwd(choose.dir())

install.packages('tidyverse')

library(tidyverse)

#Loading the healthcare xls file

library(readxl)

Healthcare = read\_xlsx('E:/SIMPLILEARn/Project/Healthcare cost analysis p7/1555054100\_hospitalcosts.xlsx')

summary(Healthcare)

str(Healthcare)

# checking for na or missing values

sum(is.na(Healthcare$AGE))

sum(is.na(Healthcare$FEMALE))

sum(is.na(Healthcare$LOS))

sum(is.na(Healthcare$RACE))

sum(is.na(Healthcare$TOTCHG))

sum(is.na(Healthcare$APRDRG))

#Since the female column is a categorical variable it will be convenient to convert it into

#a factor

Healthcare$AGE = as.factor(Healthcare$AGE)

Healthcare$FEMALE = as.factor(Healthcare$FEMALE)

#Since the Race column has an NA field we will have to check it!

Healthcare = na.omit(Healthcare)

view(Healthcare)

str(Healthcare)

length(unique(Healthcare$APRDRG))

# There are 63 unique diagnosis related groups

#TASK 1

str(Healthcare)

library(ggplot2)

#total charges accumulated per age

totchgagewise = aggregate(TOTCHG~AGE,Healthcare,sum)

str(totchgagewise)

view(totchgagewise)

summary(Healthcare$AGE)

a1 = (summary(Healthcare$AGE))

a1 = as.data.frame(a1)

str(a1)

view(a1)

#creating a new data frame consisting of occurences of the ages and cumulative expenditures

totchgagewise = cbind(totchgagewise,summary(Healthcare$AGE))

view(totchgagewise)

names(Healthcare)

ggplot(totchgagewise)+

geom\_point(aes(x = AGE, y = summary(Healthcare$AGE),size = TOTCHG),color ='blue',shape = 22 ,fill = 'black')+

theme(plot.background = element\_rect(fill = 'white'),panel.background = element\_rect(fill = 'orange'))+

xlab('Age Group')+

ylab('Counts')+

ggtitle('Plot of Age group and counts and their corresponding expenditures')

#Question 2

#Finding which diagnosis grp has the highest expenditure and admission

unique(Healthcare$APRDRG)

#For that lets factorise the APRDRG

Healthcare$APRDRG = as.factor(Healthcare$APRDRG)

str(Healthcare)

H\_table = as.data.frame(table(Healthcare$APRDRG))

view(H\_table)

colnames(H\_table) = c('APRDRG', 'Counts')

#From this table one can clearly observe that

#APRDRG group number 640 has the highest number of patient admission

H\_table\_1 = aggregate(TOTCHG~APRDRG,Healthcare,sum)

view(H\_table\_1)

# deleting the third column as it is only a repetition of the first column

H\_table = merge(H\_table,H\_table\_1)

view(H\_table)

#This H\_table represents the diagnosis groups and their case counts and their total expenditure

ggplot(H\_table)+

geom\_point(aes(x = APRDRG, y = TOTCHG,size = Counts),col = 'red')+

theme(axis.ticks.x = element\_line(color = 'purple'),axis.ticks.y = element\_line(color = 'purple'))+

theme(axis.title.x = element\_text(vjust = 0.6))+

theme(axis.text.x = element\_text(vjust = 0.4, angle = 45))

#Q3 whether the race of the patient is related to tot charges

str(Healthcare)

#Converting race to a factor

Healthcare$RACE = as.factor(Healthcare$RACE)

unique(Healthcare$RACE)

summary(Healthcare$RACE)

Race\_health = aggregate(TOTCHG~RACE,Healthcare,sum)

view(Race\_health)

Race\_health1 = as.data.frame(table(Healthcare$RACE))

colnames(Race\_health1) = c('RACE','COUNTS')

view(Race\_health1)

Race\_health = merge(Race\_health,Race\_health1)

view(Race\_health)

Race\_health$PPCOST = Race\_health$TOTCHG/Race\_health$COUNTS

View(Race\_health)

#lets make and lmtest to confirm the influence of diffrent variables influencing totchg

??lmtest

install.packages('lmtest')

library(lmtest)

?lm

fit1 = lm(TOTCHG~.,Healthcare)

summary(fit1)

#from this summary one can observe that Race has no significantr role to play in terms of charge

# We will also use an alternative mothod to verify it

#making fit yusing a decision tree

library(rpart)

fit2 = rpart(TOTCHG~.,Healthcare)

?rpart

library(rpart.plot)

rpart.plot(fit2, extra = 1)

?plot

summary(fit2)

#from tthe lmtest we can clearly infer that the race has no significant influence

#in prices

#also from the decision tree we can clearly see RACE as almost no influence in the

#price hence its safe to conclude that the healthcare is devoid of malpractice

#for the 4th question lets make a decision tree featuring the variables discussed

str(Healthcare)

Healthcare$FEMALE = as.factor(Healthcare$FEMALE)

fit3 = rpart(TOTCHG~AGE+FEMALE,Healthcare)

fit3

summary(fit3)

rpart.plot(fit3,cex= 0.6)

#using lmtest

fit4 = lm(TOTCHG~AGE+FEMALE,Healthcare)

fit4

summary(fit4)

table(Healthcare$FEMALE)

#The decision treee concludes that the age has the highest influence

#in total charges.

#The lmtest also specifies that Age group numbers 1,3,4,5,6,9,10,15,17 makes the most influence

#Q5 whether LOS is predictable from age ,gender and race

#lmtest

Healthcare$LOS = as.numeric(Healthcare$LOS)

str(Healthcare)

fit\_\_1 = lm(LOS~AGE+FEMALE+RACE,Healthcare)

fit\_\_1

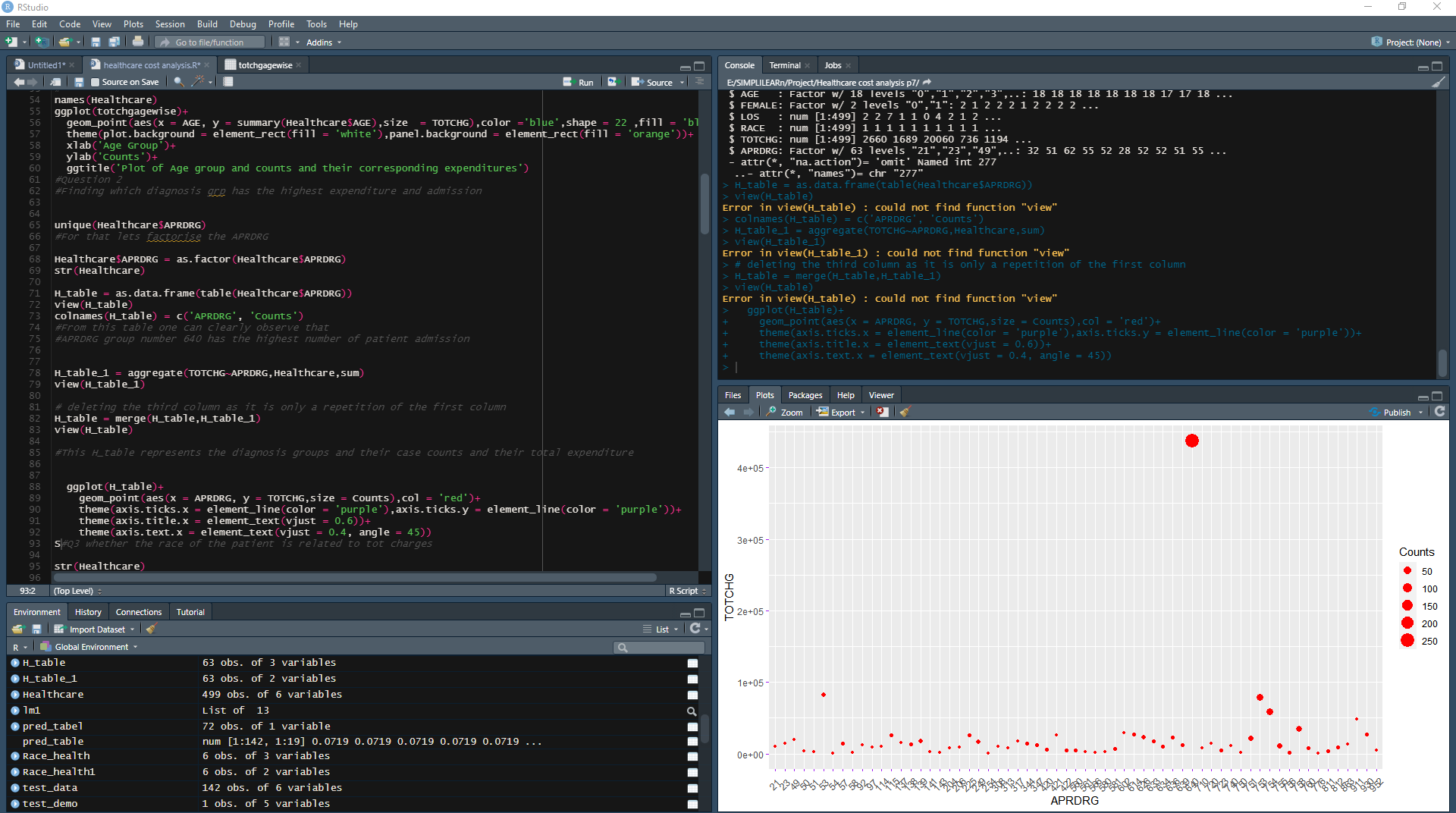
summary(fit\_\_1)

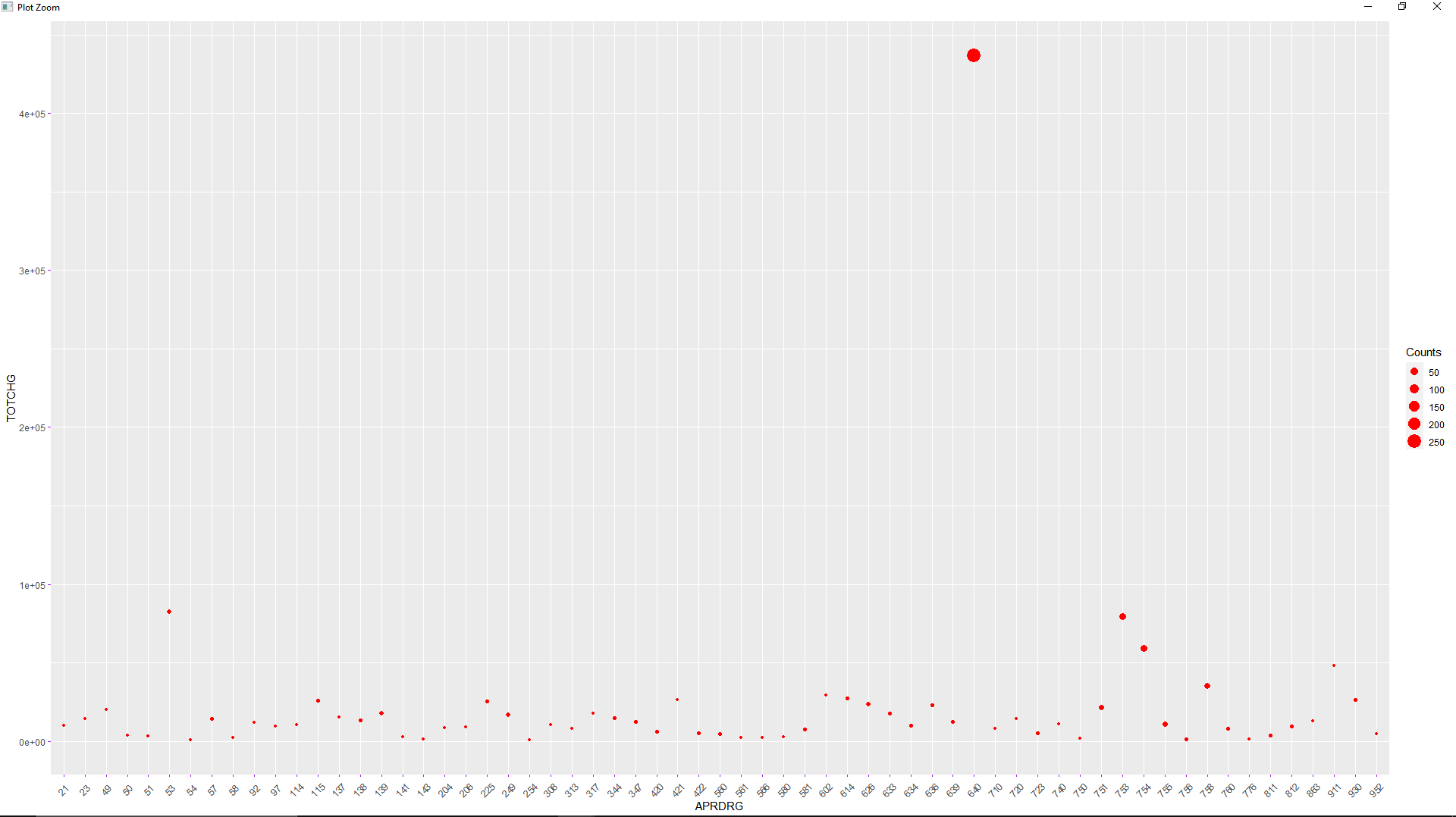
#|\_\_\_\_\_\_\_\_\_\_\_\_\_\_> Hence length of stay cannot be predicted from AGE,GENDER,RACE.

# Analysis

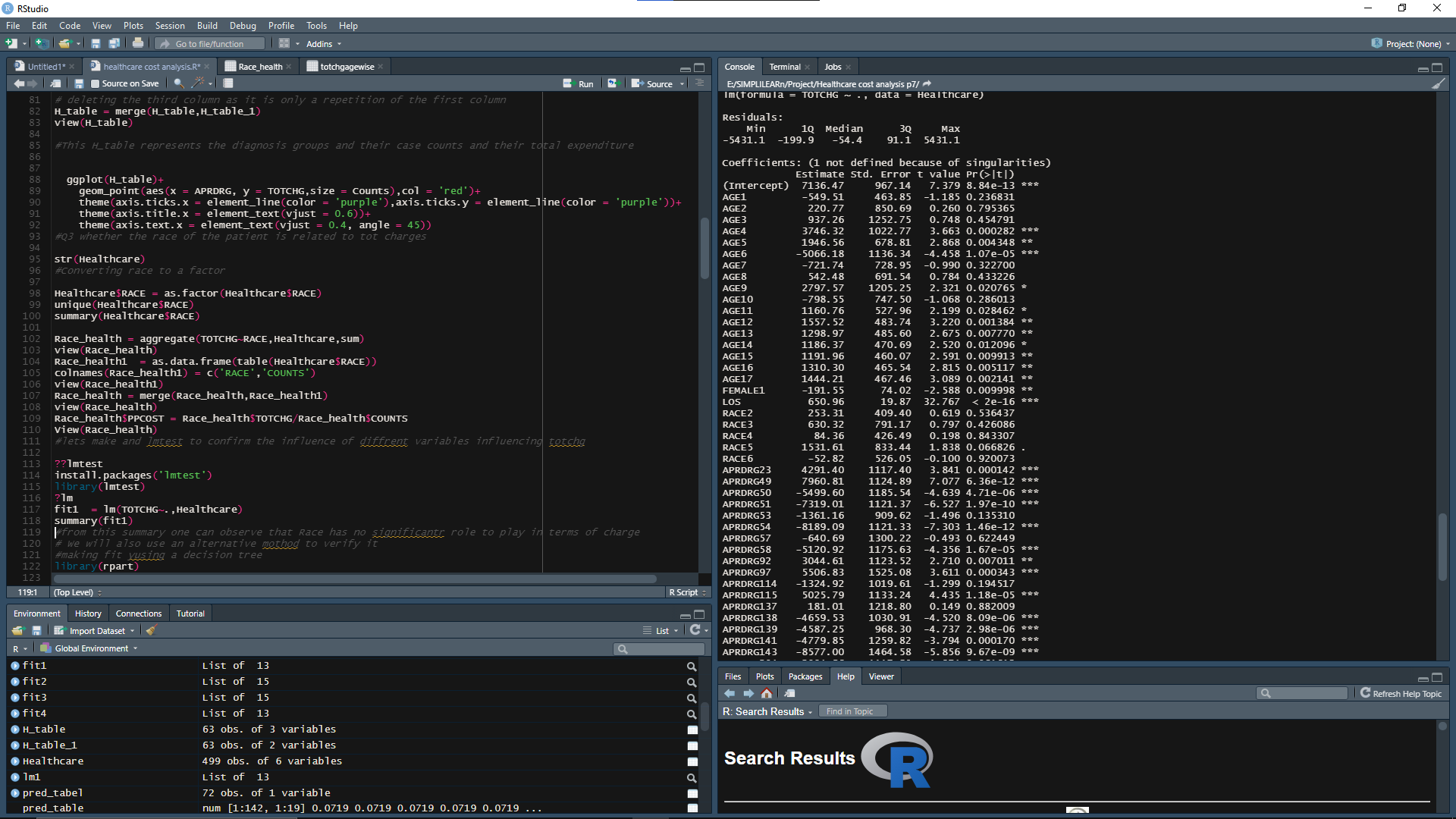
# 

For the first question, the plot clearly shows that the number of patients of the age group 0 is the highest, and their total expenditure is also highest among all the age groups as shown by the size of the respective boxes.

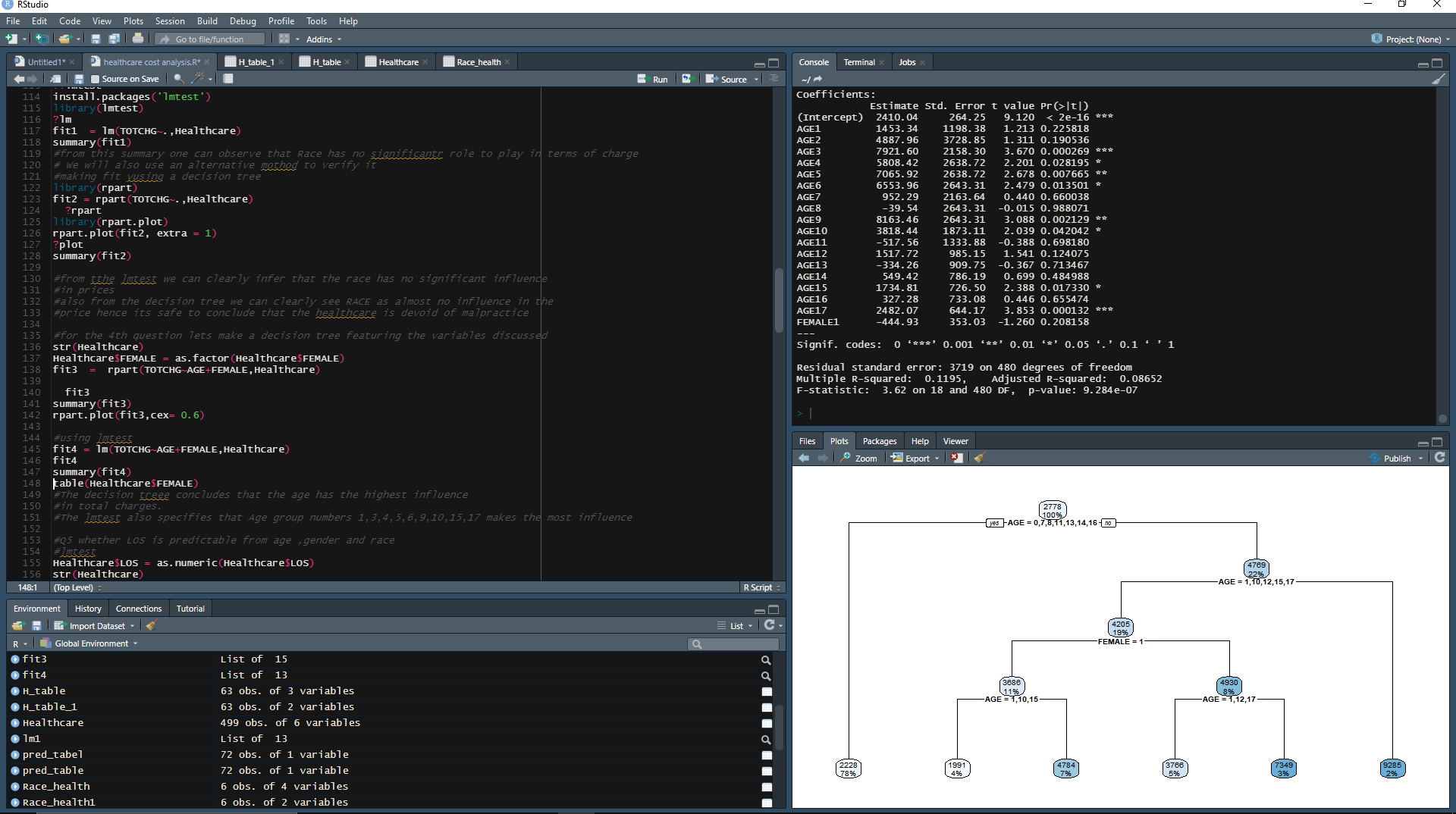




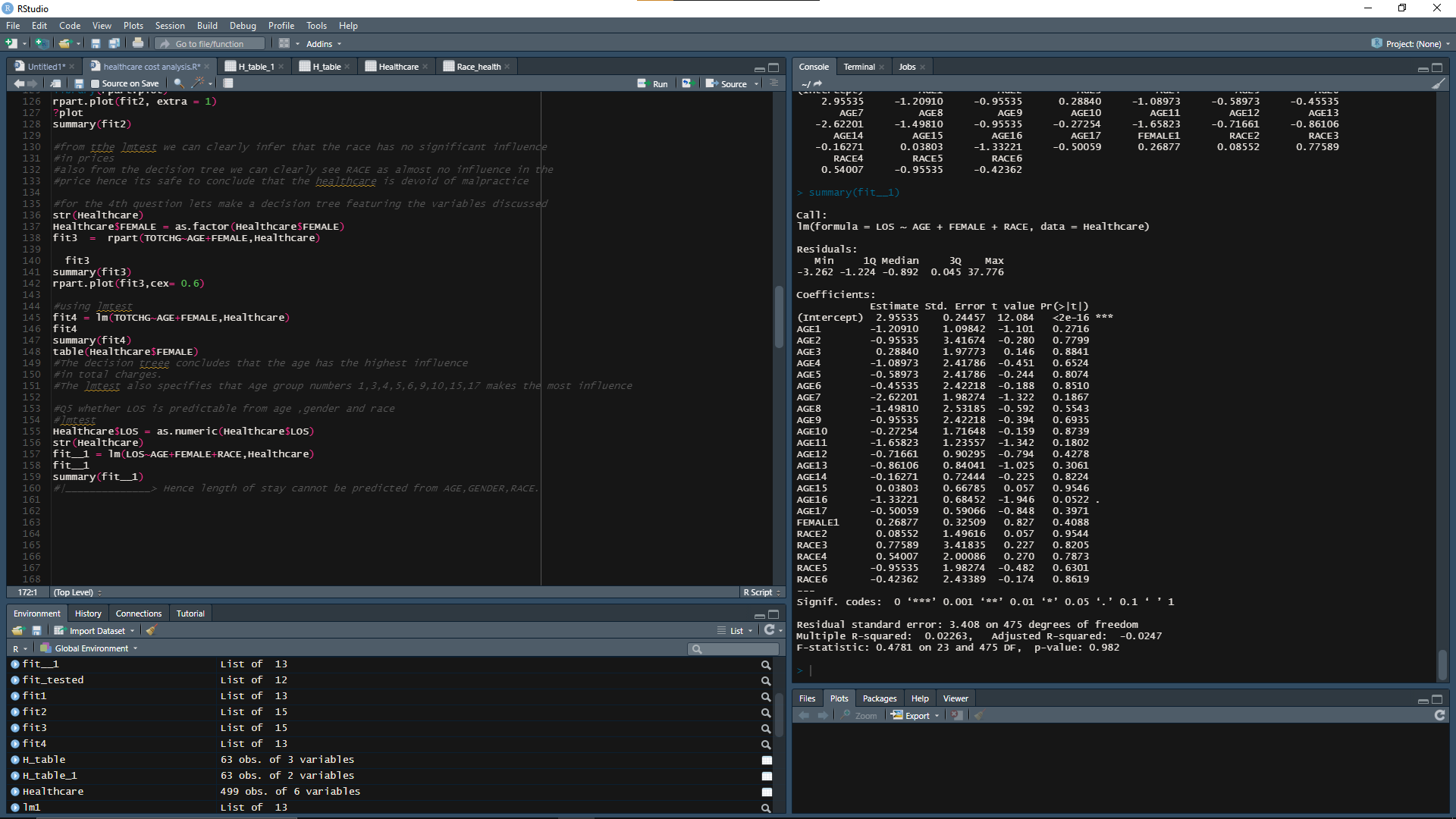
From the plot it can be easily identified that the diagnostic group number 640 has the highest number of hospitalizations (defined by the respective circle sizes) as well has highest total charges among all diagnostic related groups. Hence this answers the second question.

****

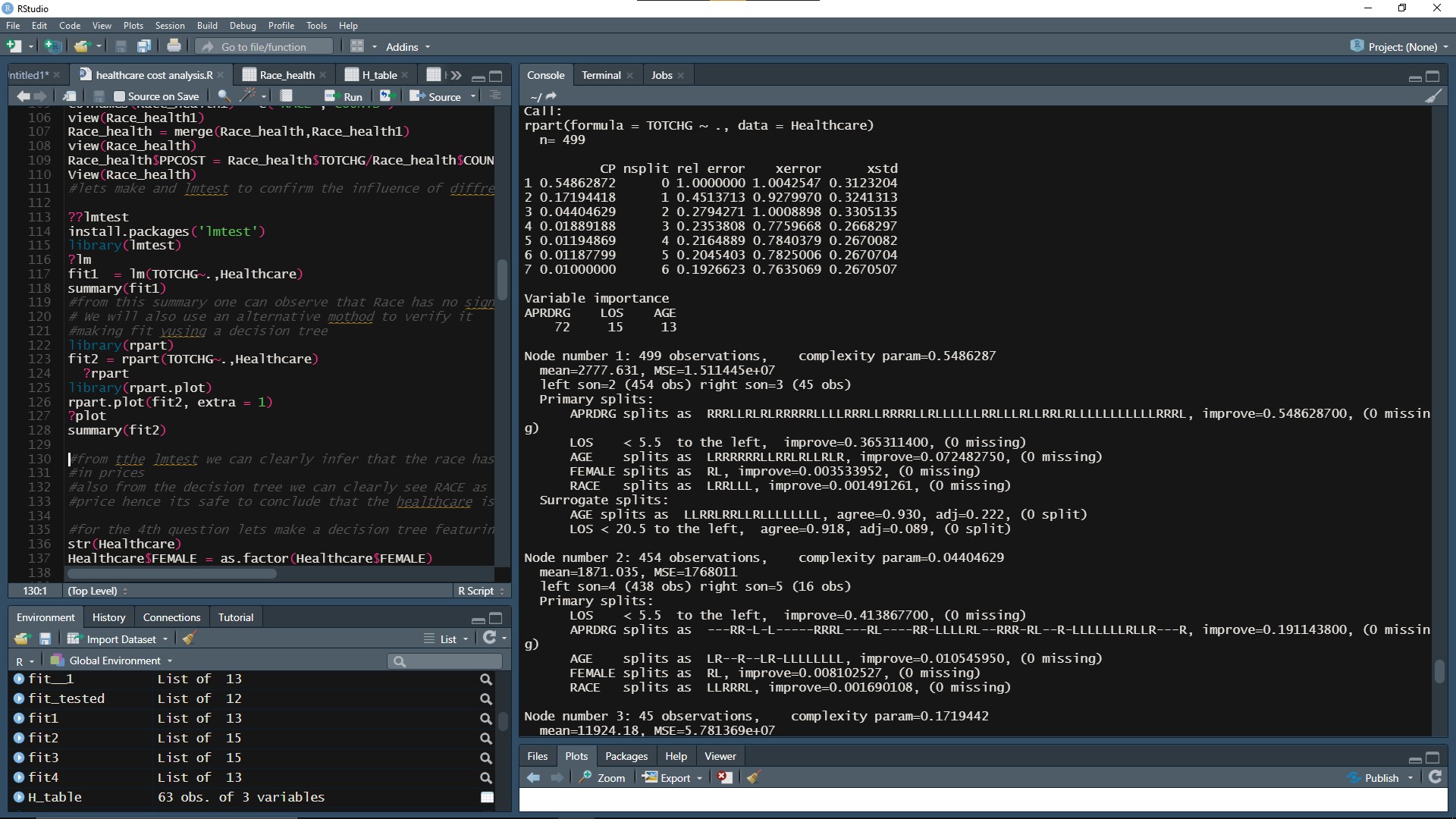
As the linear-regression model shows, race has no significant role to play (p>0.05) in determining patient charges, hence we can safely say there is no racial malpractice in the hospital in terms of patient charges, hence this analysis answers the 3rd question.



To answer the 4th question both decision tree modelling and linear-regression modelling has been performed on the data. The console displays the linear-regression modelling result, which clearly shows that the severity of hospital charges can be clearly defined by certain age groups i.e- 3,4,5,6,9,10,15 & 17,however gender has a minimal role to play in severity of hospital charges.



To answer the 5th question, a linear-regression modelling has been performed featuring LOS (Length of stay) as a dependent variable and other three predictors as independent variables. And the result on the right-hand side reflects the fact that LOS cannot be predicted using Age, Gender and Race.



To answer the Q6 we would go few steps back when we made a decision tree where the total charges were dependent variable and the rest was independent, we can clearly see that the variable of importance is APRDRG i.e. the diagnostic related group, followed by the LOS (Length of stay) and AGE. Hence we can successfully answer the Q6 by concluding that it is the APRDRG which mainly affects the hospital costs, followed by LOS and AGE.