|  |  |
| --- | --- |
|  | Healthcare cost analysis |
|  |  |
|  | Margil Shah  DATA SCIENCE WITH R  8/2/20 |

Source Code – Full Project

Contents

[DESCRIPTION 2](#_Toc47302789)

[Background and Objective: 2](#_Toc47302790)

[Domain 2](#_Toc47302791)

[Dataset Description 2](#_Toc47302792)

[Analysis to be done 2](#_Toc47302793)

[Screen Shots 3](#_Toc47302794)

[Load Dataset 3](#_Toc47302795)

[Data Analysis 3](#_Toc47302796)

[Analysis 1 4](#_Toc47302797)

[Analysis 2 4](#_Toc47302798)

[Analysis 3 5](#_Toc47302799)

[Analysis 4 5](#_Toc47302800)

[Analysis 5 6](#_Toc47302801)

[Analysis 6 6](#_Toc47302802)

[Source Code 7](#_Toc47302803)

# DESCRIPTION

## Background and Objective:

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.

Domain Healthcare

## Dataset Description

Here is a detailed description of the given dataset:

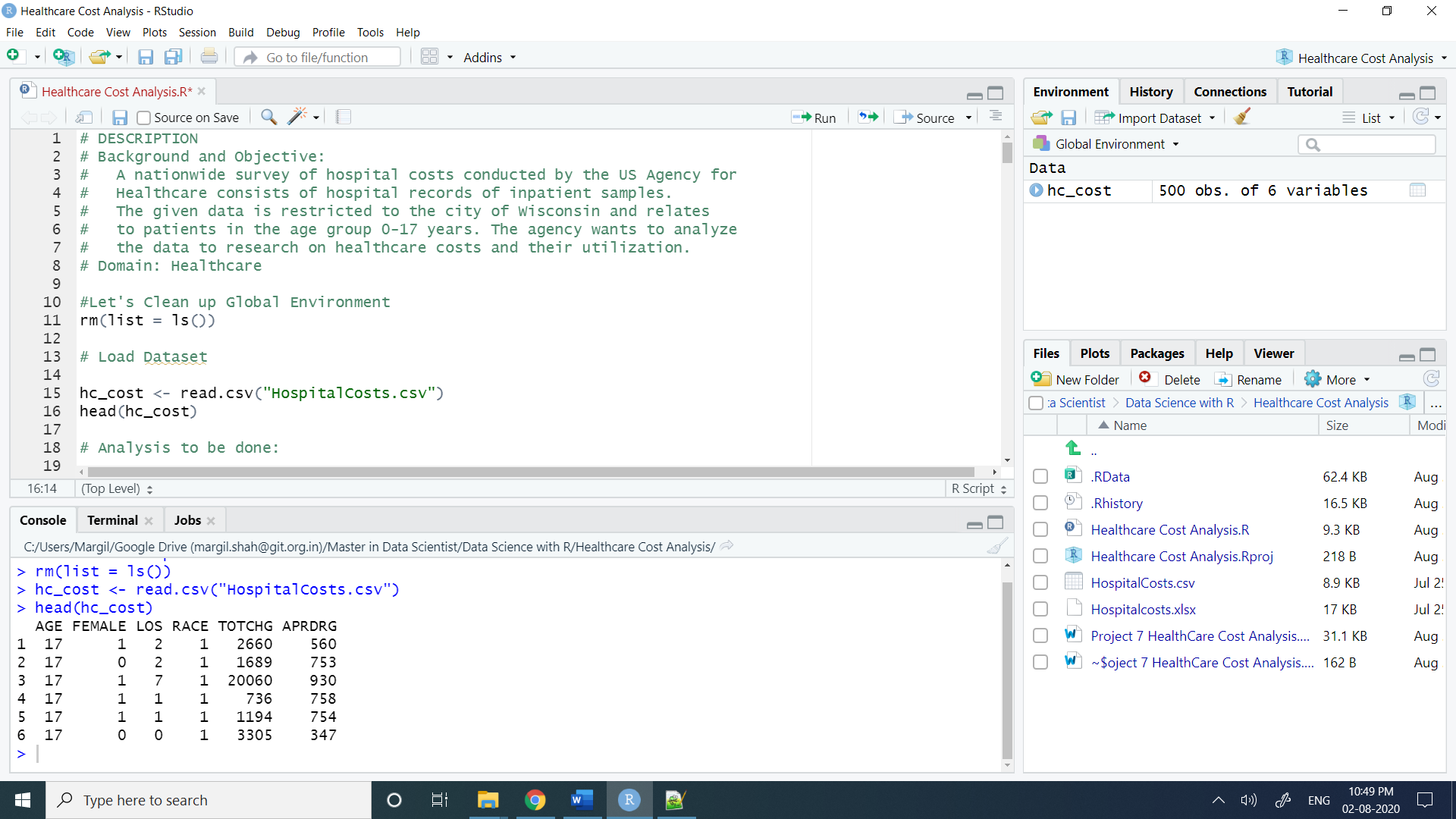
|  |  |
| --- | --- |
| **ATTRIBUTE** | **Description** |
| AGE | Age of the patient discharged |
| FEMALE | A binary variable that indicates if the patient is female |
| LOS | Length of stay in days |
| RACE | Race of the patient (specified numerically) |
| TOTCHG | Hospital discharge costs |
| APRDRG | All Patient Refined Diagnosis Related Groups |

## Analysis to be done

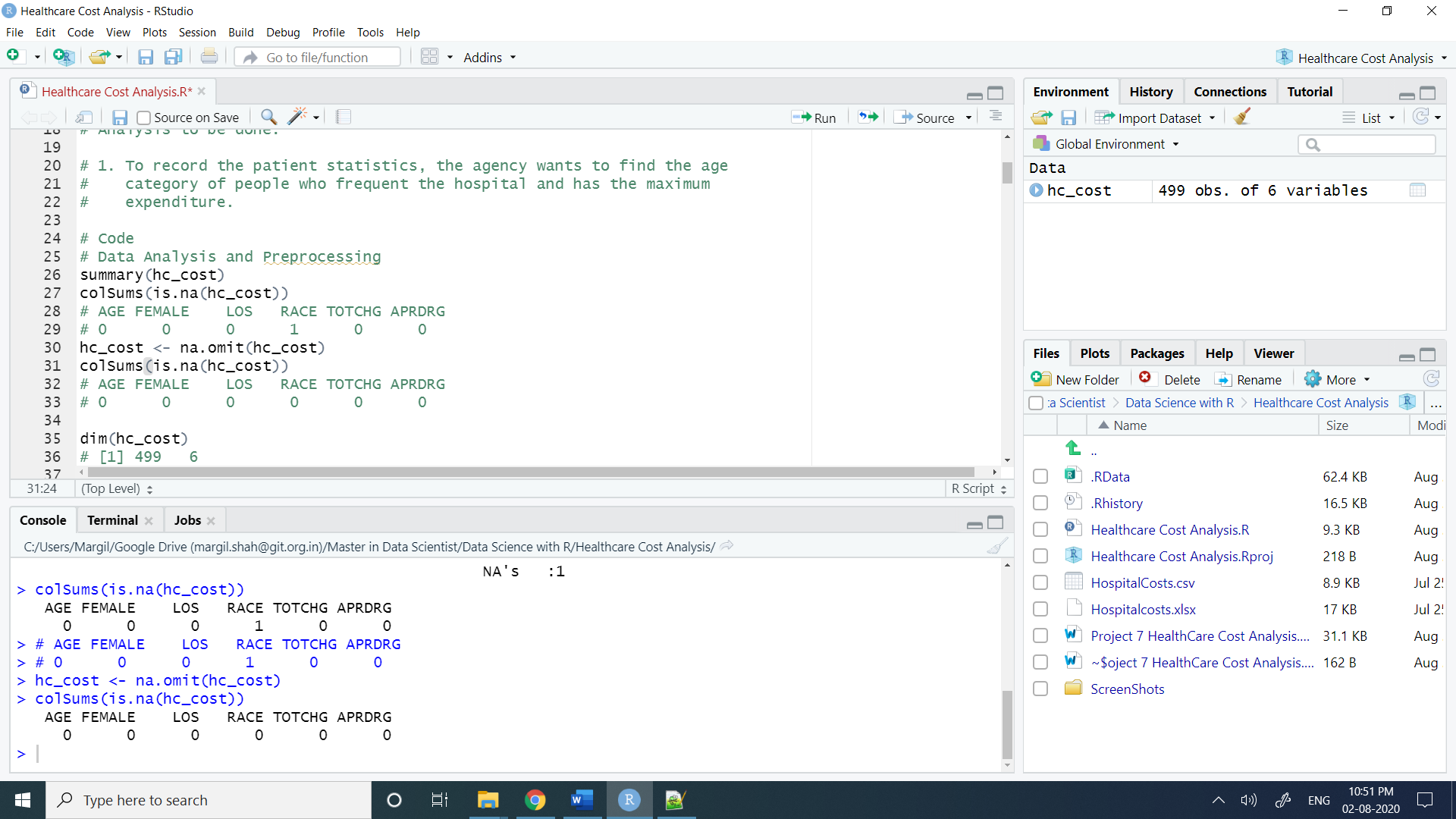
1. To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.
2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that 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.

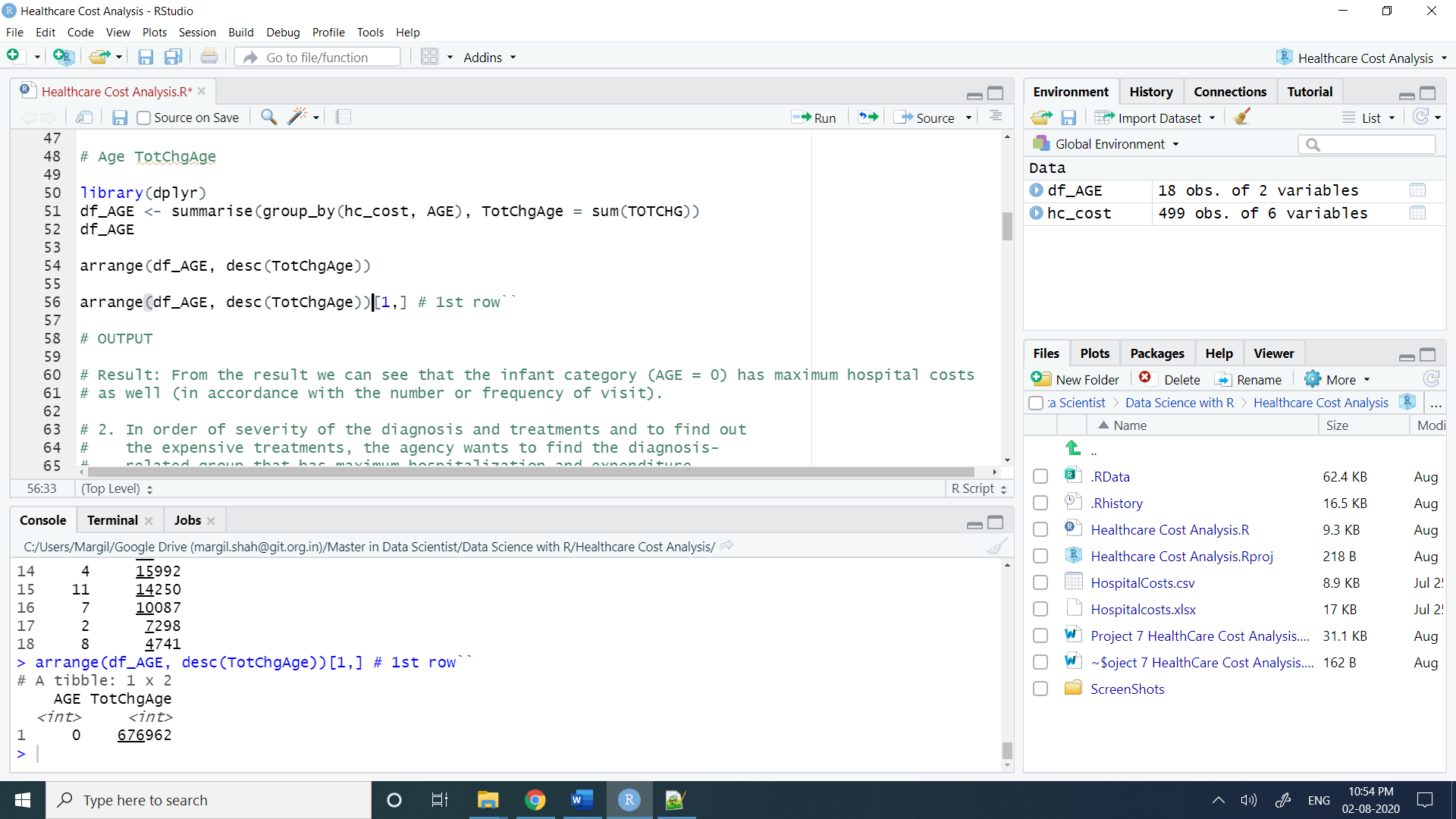
# Screen Shots

## Load Dataset

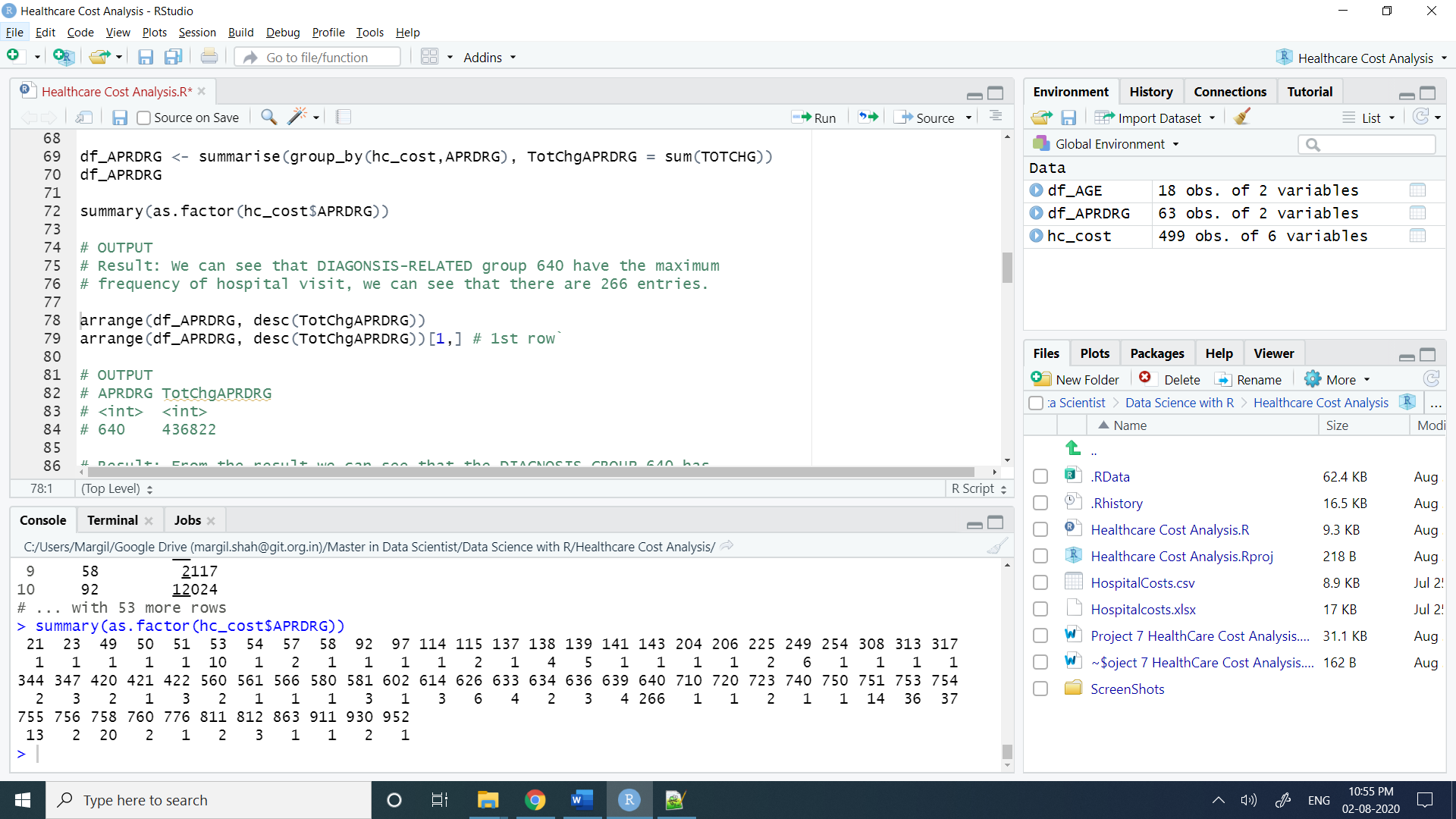


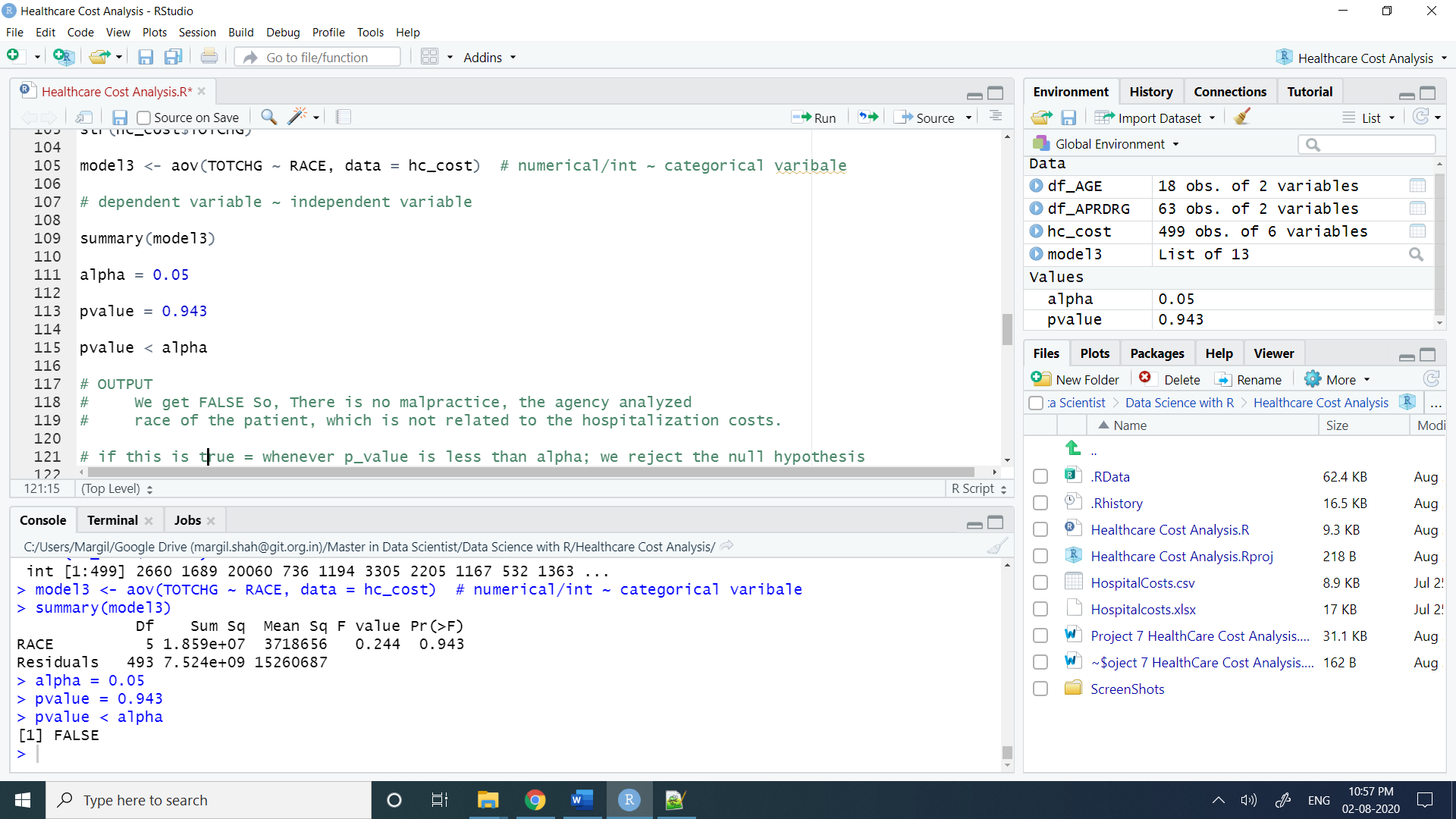
## Data Analysis



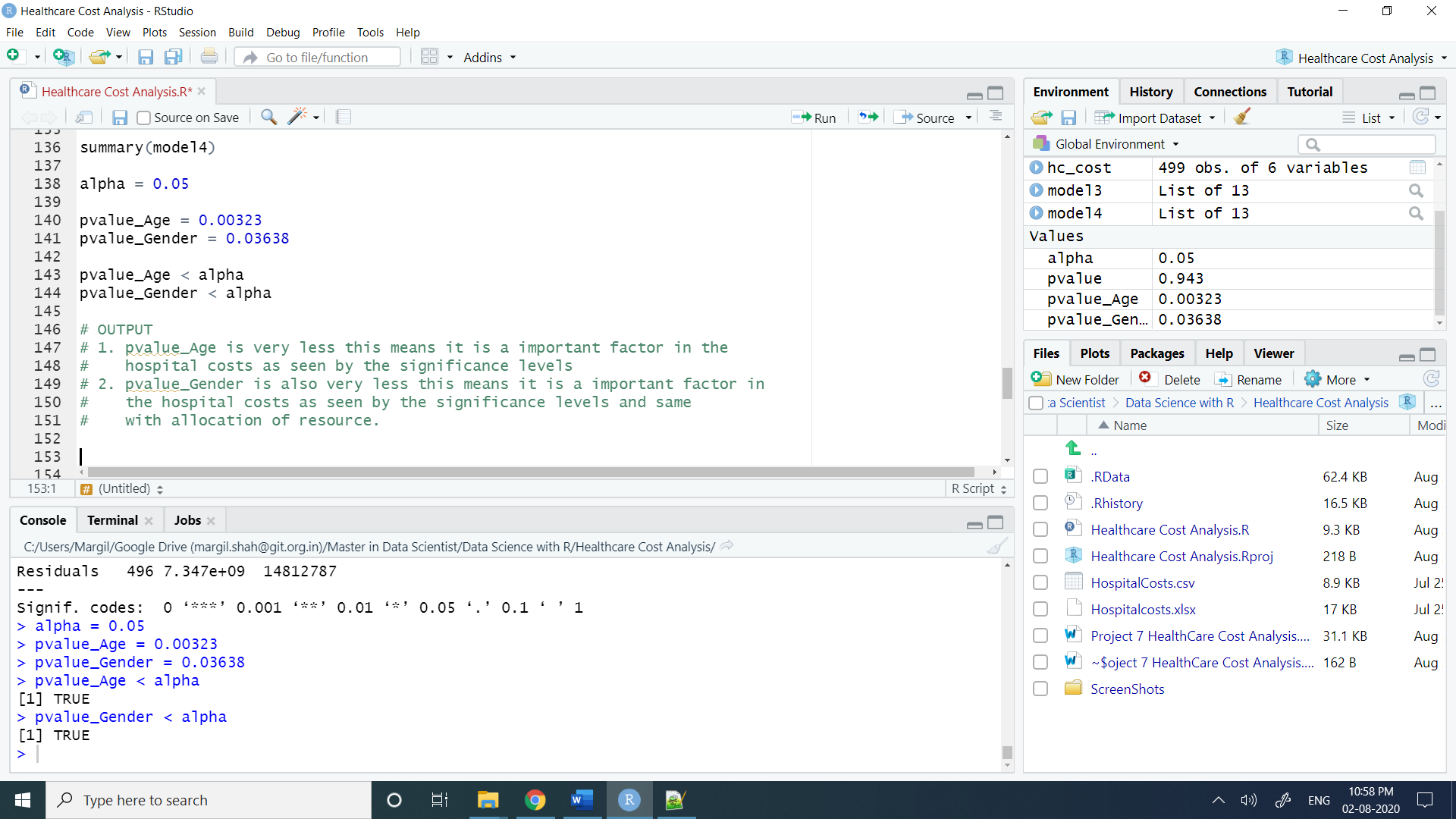
Analysis 1

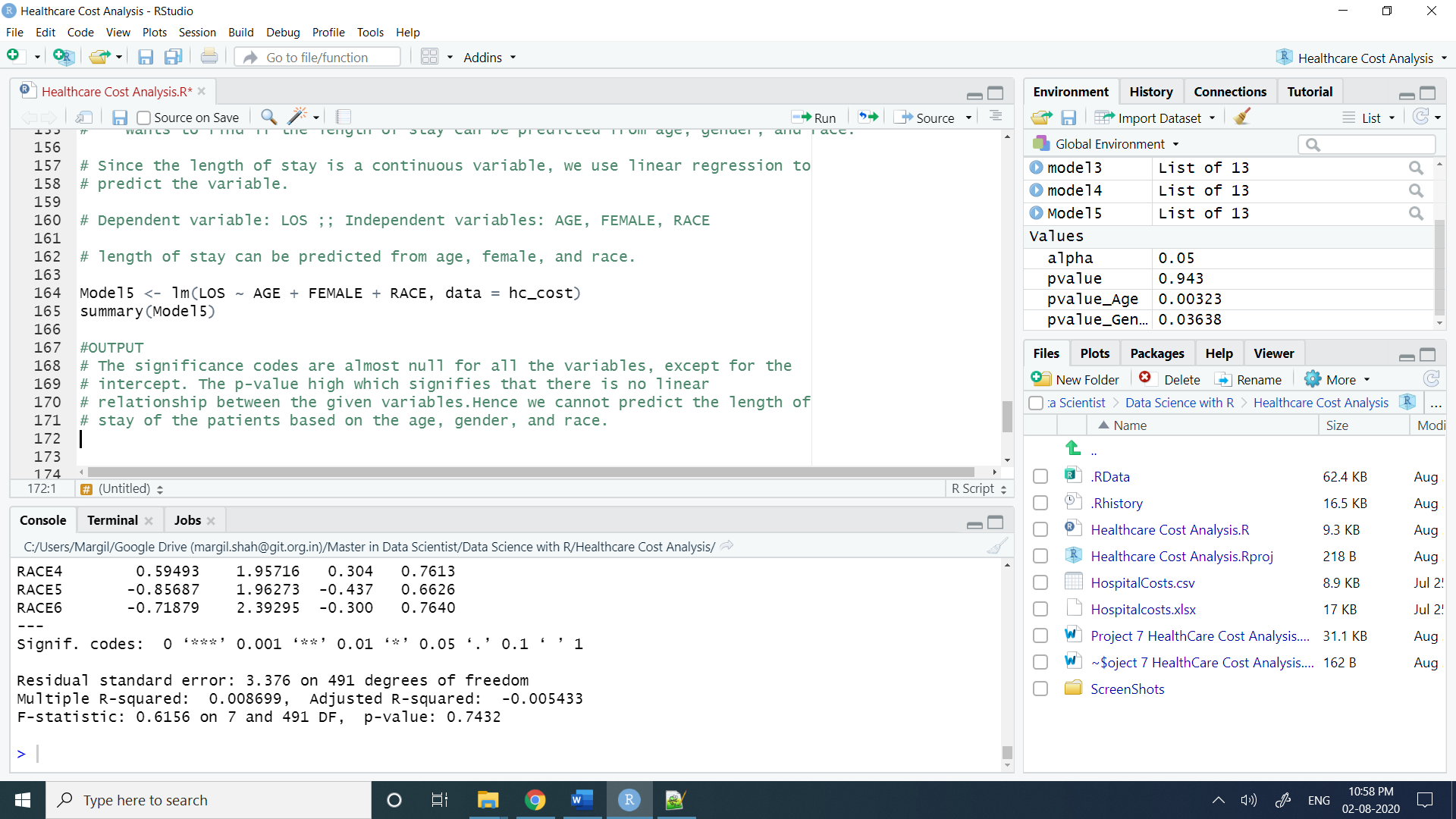
## Analysis 2



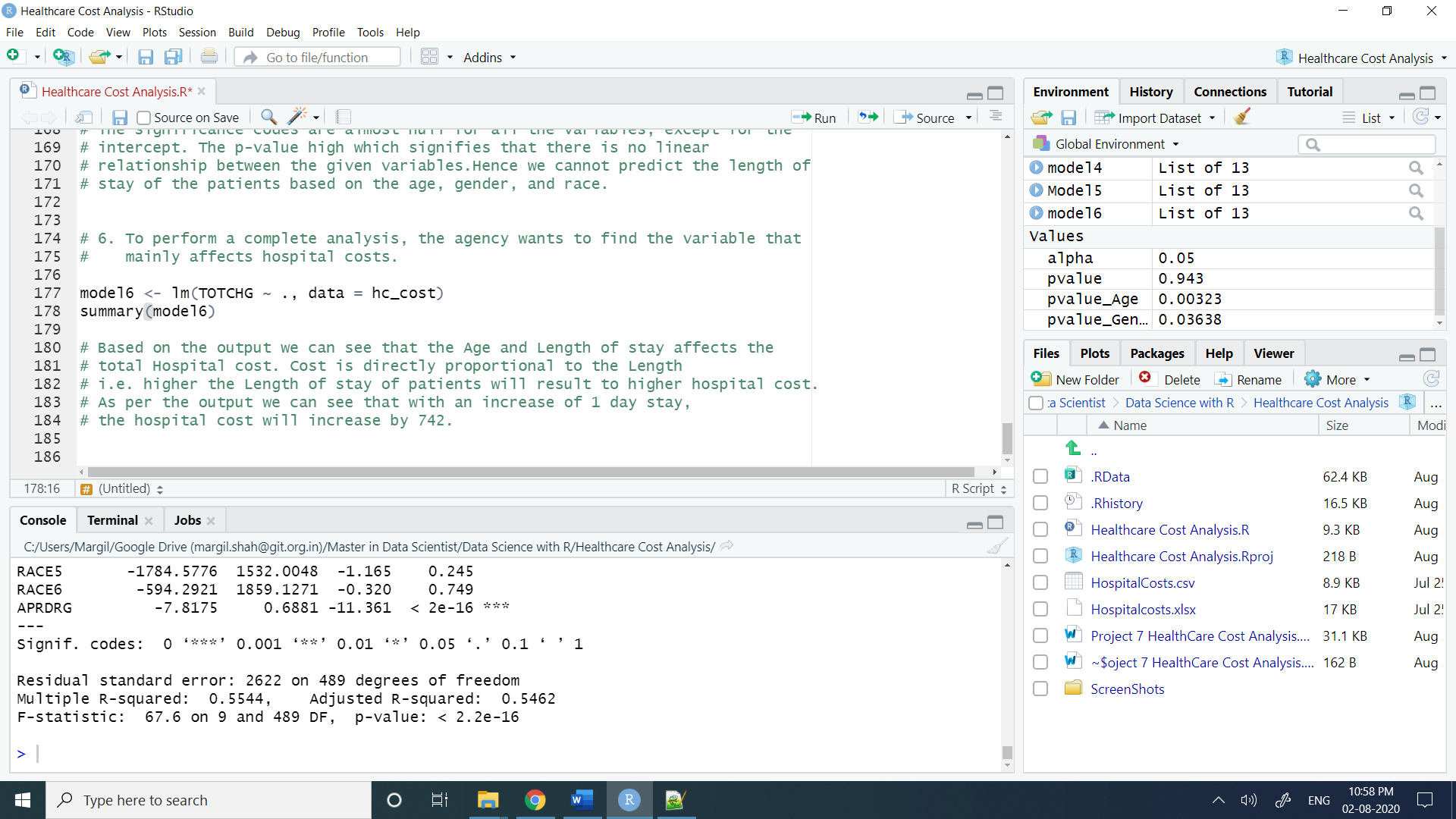
Analysis 3

## Analysis 4



Analysis 5

## Analysis 6



# Source Code

#Let's Clean up Global Environment

rm(list = ls())

# Load Dataset

hc\_cost <- read.csv("HospitalCosts.csv")

head(hc\_cost)

# Analysis to be done:

# 1. To record the patient statistics, the agency wants to find the age

# category of people who frequent the hospital and has the maximum

# expenditure.

# Code

# Data Analysis and Preprocessing

summary(hc\_cost)

colSums(is.na(hc\_cost))

hc\_cost <- na.omit(hc\_cost)

colSums(is.na(hc\_cost))

dim(hc\_cost)

# [1] 499 6

str(hc\_cost)

hc\_cost$RACE <- as.factor(hc\_cost$RACE)

hc\_cost$FEMALE <- as.factor(hc\_cost$FEMALE)

str(hc\_cost)

summary(as.factor(hc\_cost$AGE))

# Result: We can see that infants (AGE = 0) have the maximum

# frequency of hospital visit, going above 300. we can see that

# there are 306 entries for those in the range of 0-1 year.

# To find the age category with the maximum expenditure

# Age TotChgAge

library(dplyr)

df\_AGE <- summarise(group\_by(hc\_cost, AGE), TotChgAge = sum(TOTCHG))

df\_AGE

arrange(df\_AGE, desc(TotChgAge))

arrange(df\_AGE, desc(TotChgAge))[1,] # 1st row``

# OUTPUT

# Result: From the result we can see that the infant category (AGE = 0) has maximum hospital costs

# as well (in accordance with the number or frequency of visit).

# 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.

# Code

df\_APRDRG <- summarise(group\_by(hc\_cost,APRDRG), TotChgAPRDRG = sum(TOTCHG))

df\_APRDRG

summary(as.factor(hc\_cost$APRDRG))

# OUTPUT

# Result: We can see that DIAGONSIS-RELATED group 640 have the maximum

# frequency of hospital visit, we can see that there are 266 entries.

arrange(df\_APRDRG, desc(TotChgAPRDRG))

arrange(df\_APRDRG, desc(TotChgAPRDRG))[1,] # 1st row

# OUTPUT

# Result: From the result we can see that the DIAGNOSIS-GROUP 640 has

# maximum hospital costs as well

# (in accordance with the number or frequency of visit 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.

# If there is any effect of RACE on TOTCHG

# Then, to verify if the races made an impact on the costs, perform an ANOVA with the

# following variables:

# ANOVA dependent variable: TOTCHG

# Categorical/grouping variable: RACE Missing values: 1 NA value, use na.omit

# to remove the NA value

#

# Code:

str(hc\_cost$RACE)

str(hc\_cost$TOTCHG)

model3 <- aov(TOTCHG ~ RACE, data = hc\_cost) # numerical/int ~ categorical varibale

# dependent variable ~ independent variable

summary(model3)

alpha = 0.05

pvalue = 0.943

pvalue < alpha

# OUTPUT

# We get FALSE So, There is no malpractice, the agency analyzed

# race of the patient, which is not related to the hospitalization costs.

# if this is true = whenever p\_value is less than alpha; we reject the null hypothesis

# 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.

#-----------------------------------------------------------------------------------------------------

# ANOVA is used to test relationship between one quantitative and one or

# more qualitative variable.

#-----------------------------------------------------------------------------------------------------

str(hc\_cost$AGE)

str(hc\_cost$FEMALE)

model4 <- aov(TOTCHG ~ AGE + FEMALE, data = hc\_cost)

summary(model4)

alpha = 0.05

pvalue\_Age = 0.00323

pvalue\_Gender = 0.03638

pvalue\_Age < alpha

pvalue\_Gender < alpha

# OUTPUT

# 1. pvalue\_Age is very less this means it is a important factor in the

# hospital costs as seen by the significance levels

# 2. pvalue\_Gender is also very less this means it is a important factor in

# the hospital costs as seen by the significance levels and same

# with allocation of resource.

# 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.

# Since the length of stay is a continuous variable, we use linear regression to

# predict the variable.

# Dependent variable: LOS ;; Independent variables: AGE, FEMALE, RACE

# length of stay can be predicted from age, female, and race.

Model5 <- lm(LOS ~ AGE + FEMALE + RACE, data = hc\_cost)

summary(Model5)

#OUTPUT

# The significance codes are almost null for all the variables, except for the

# intercept. The p-value high which signifies that there is no linear

# relationship between the given variables.Hence we cannot predict the length of

# stay of the patients based on the age, gender, and race.

# 6. To perform a complete analysis, the agency wants to find the variable that

# mainly affects hospital costs.

model6 <- lm(TOTCHG ~ ., data = hc\_cost)

summary(model6)

# Based on the output we can see that the Age and Length of stay affects the

# total Hospital cost. Cost is directly proportional to the Length

# i.e. higher the Length of stay of patients will result to higher hospital cost.

# As per the output we can see that with an increase of 1 day stay,

# the hospital cost will increase by 742.