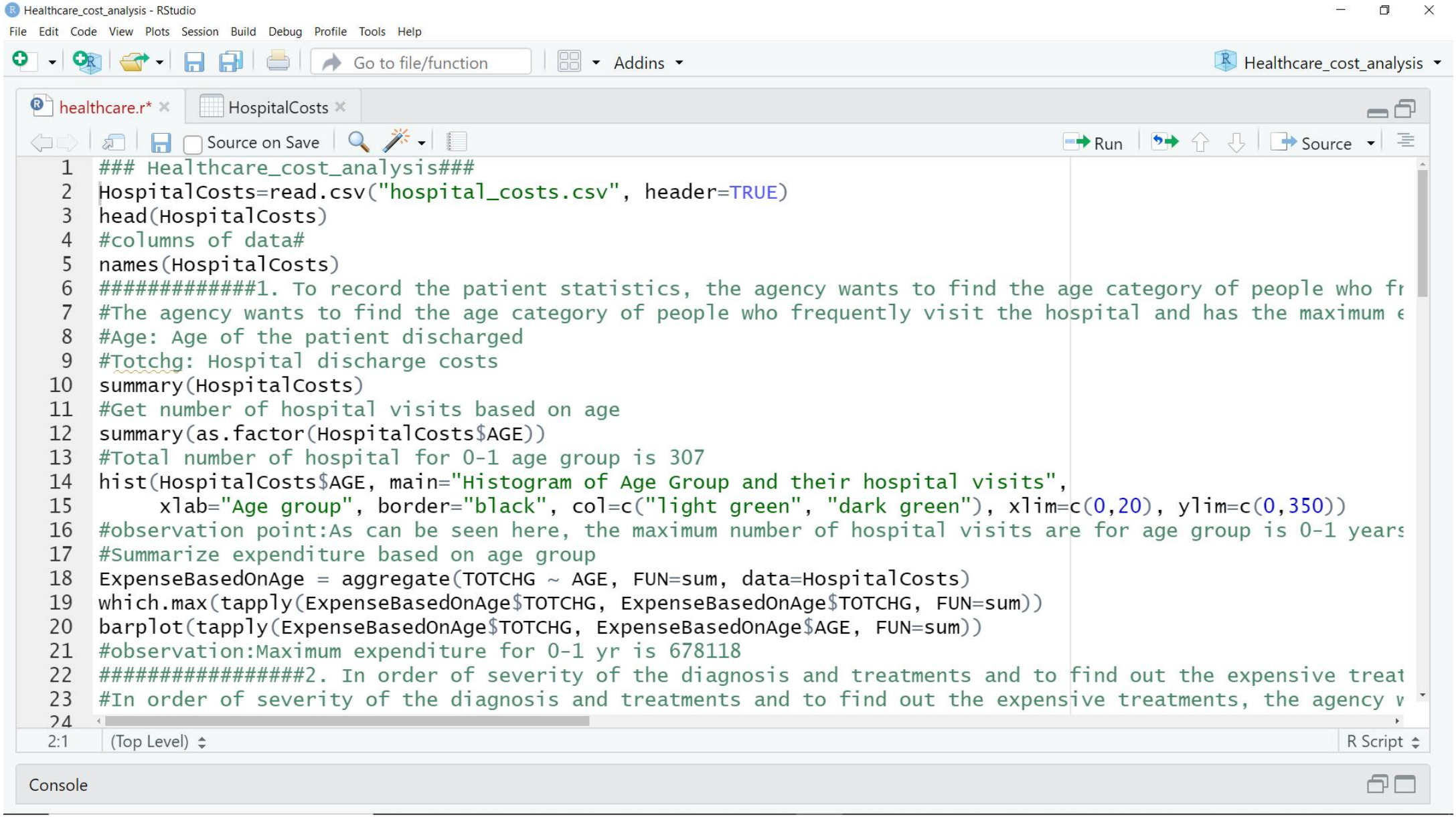
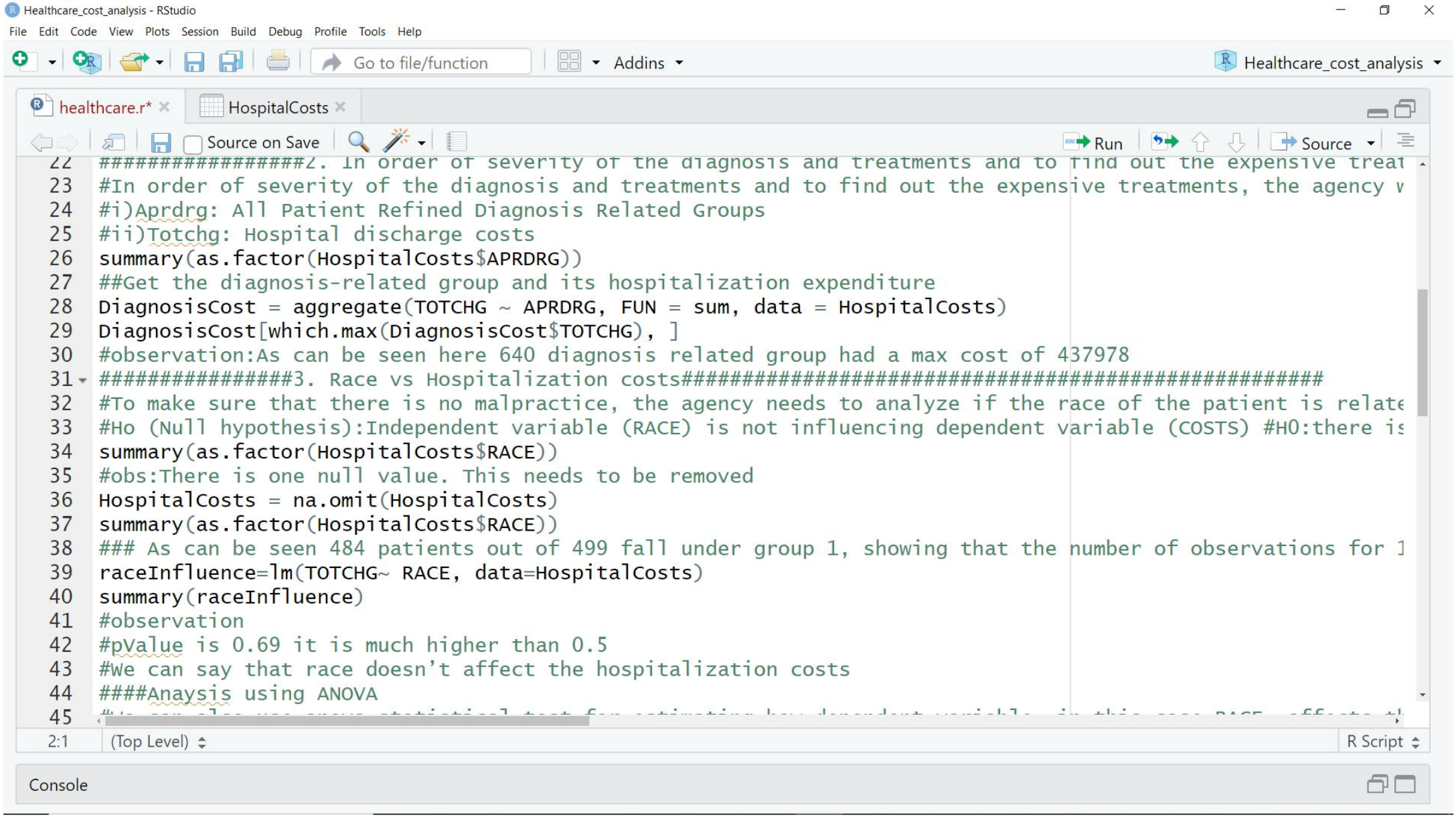
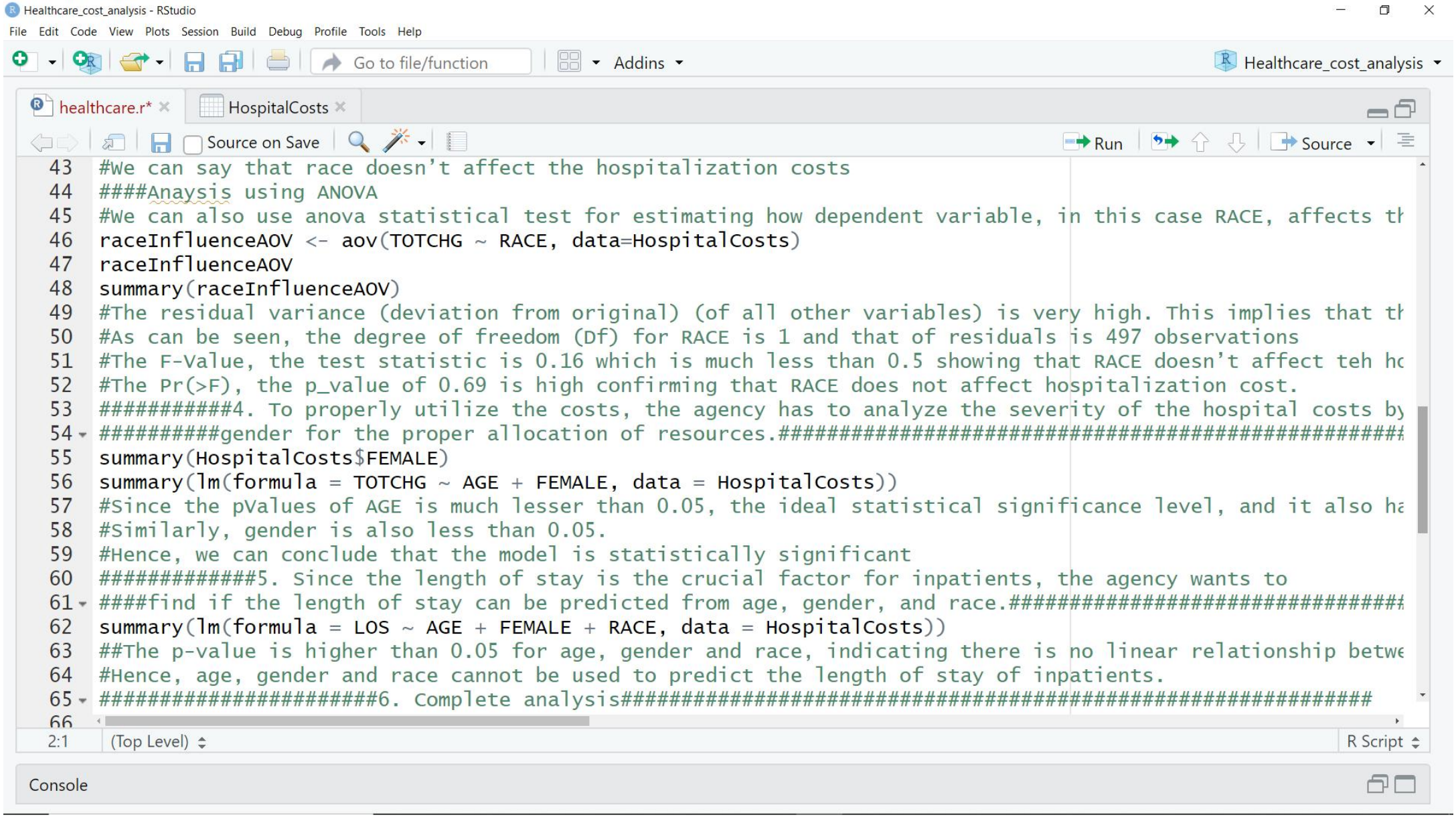
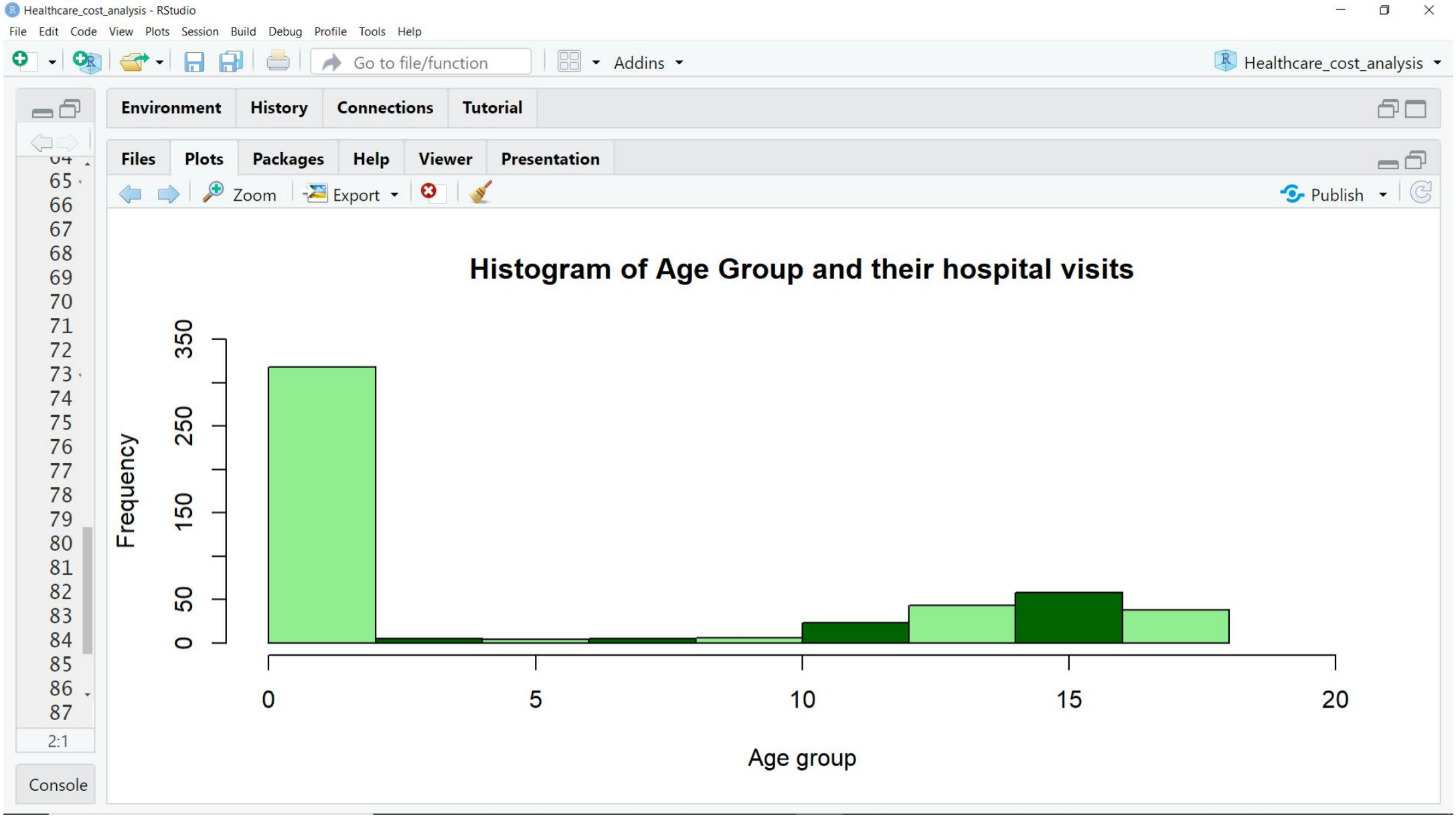
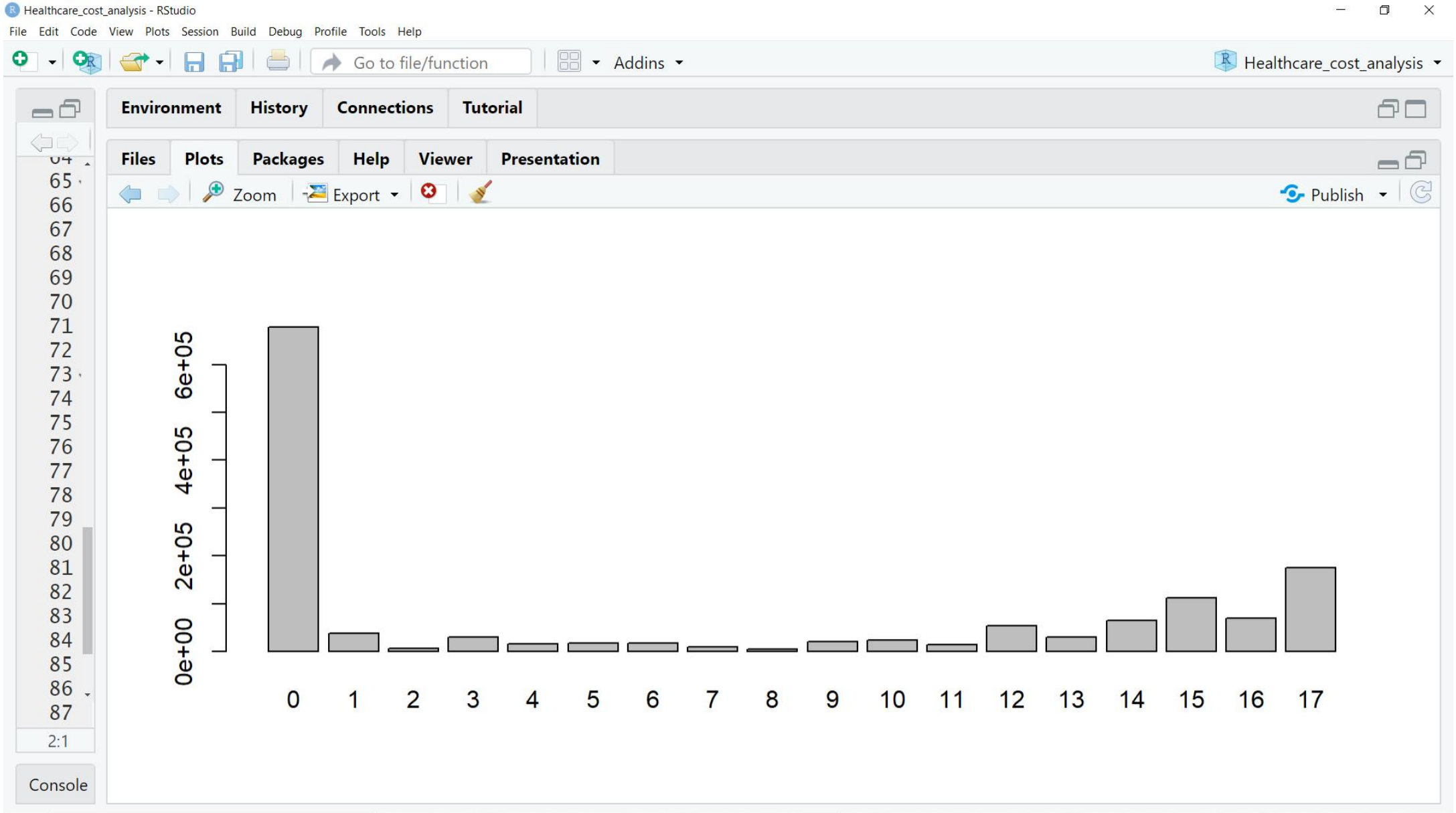

Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
healthcare.r*
HospitalCosts
Source on Save
Run
Source
Healthcare_cost_analysis###
HospitalCosts=read.csv("hospital_costs.csv", header=TRUE)
head(HospitalCosts)
#columns of data#
names(HospitalCosts)
#############1. To record the patient statistics, the agency wants to find the age category of people who fr
#The agency wants to find the age category of people who frequently visit the hospital and has the maximum e
#Age: Age of the patient discharged
#Totchg: Hospital discharge costs
summary(HospitalCosts)
#Get number of hospital visits based on age
summary(as.factor(HospitalCosts$AGE))
#Total number of hospital for 0-1 age group is 307
hist(HospitalCosts$AGE, main="Histogram of Age Group and their hospital visits",
xlab="Age group", border="black", col=c("light green", "dark green"), xlim=c(0,20), ylim=c(0,350))
#observation point:As can be seen here, the maximum number of hospital visits are for age group is 0-1 years
#Summarize expenditure based on age group
ExpenseBasedOnAge = aggregate(TOTCHG ~ AGE, FUN=sum, data=HospitalCosts)
which.max(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$TOTCHG, FUN=sum))
barplot(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$AGE, FUN=sum))
#observation:Maximum expenditure for 0-1 yr is 678118
#################2. In order of severity of the diagnosis and treatments and to find out the expensive treat
#In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency w
2:1
(Top Level)
R Script
Console


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
healthcare.r*
HospitalCosts
Source on Save
Run
Source
22 ##################2. In order of severity of the diagnosis and treatments and to find out the expensive treat
23 #In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency w
24 #i)Aprdrg: All Patient Refined Diagnosis Related Groups
25 #ii)Totchg: Hospital discharge costs
26 summary(as.factor(HospitalCosts$APRDRG))
27 ##Get the diagnosis-related group and its hospitalization expenditure
28 DiagnosisCost = aggregate(TOTCHG ~ APRDRG, FUN = sum, data = HospitalCosts)
29 DiagnosisCost[which.max(DiagnosisCost$TOTCHG),]
30 #observation:As can be seen here 640 diagnosis related group had a max cost of 437978
31 ##################3. Race vs Hospitalization costs##
32 #To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is relate
33 #Ho (Null hypothesis):Independent variable (RACE) is not influencing dependent variable (COSTS) #H0:there is
34 summary(as.factor(HospitalCosts$RACE))
35 #obs:There is one null value. This needs to be removed
36 HospitalCosts = na.omit(HospitalCosts)
37 summary(as.factor(HospitalCosts$RACE))
38 ### As can be seen 484 patients out of 499 fall under group 1, showing that the number of observations for 1
39 raceInfluence=lm(TOTCHG~ RACE, data=HospitalCosts)
40 summary(raceInfluence)
41 #observation
42 #pValue is 0.69 it is much higher than 0.5
43 #We can say that race doesn't affect the hospitalization costs
44 ####Anaysis using ANOVA
45
2:1
(Top Level)
R Script
Console


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
healthcare.r*
HospitalCosts
Source on Save
Run
Source
#We can say that race doesn't affect the hospitalization costs
####Anaysis using ANOVA
#We can also use anova statistical test for estimating how dependent variable, in this case RACE, affects th
raceInfluenceAOV <- aov(TOTCHG ~ RACE, data=HospitalCosts)
raceInfluenceAOV
summary(raceInfluenceAOV)
#The residual variance (deviation from original) (of all other variables) is very high. This implies that th
#As can be seen, the degree of freedom (Df) for RACE is 1 and that of residuals is 497 observations
#The F-Value, the test statistic is 0.16 which is much less than 0.5 showing that RACE doesn't affect teh hc
#The Pr(>F), the p_value of 0.69 is high confirming that RACE does not affect hospitalization cost.
############4. To properly utilize the costs, the agency has to analyze the severity of the hospital costs by
##########gender for the proper allocation of resources.##########
summary(HospitalCosts$FEMALE)
summary(lm(formula = TOTCHG ~ AGE + FEMALE, data = HospitalCosts))
#Since the pValues of AGE is much lesser than 0.05, the ideal statistical significance level, and it also ha
#Similarly, gender is also less than 0.05.
#Hence, we can conclude that the model is statistically significant
##############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.##########
summary(lm(formula = LOS ~ AGE + FEMALE + RACE, data = HospitalCosts))
##The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship betwe
#Hence, age, gender and race cannot be used to predict the length of stay of inpatients.
########################6. Complete analysis##########
2:1
(Top Level)
R Script
Console


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Zoom
Export
Publish
Histogram of Age Group and their hospital visits
Frequency
0 50 150 250 350
0 5 10 15 20
Age group
Console


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Zoom
Export
Publish
0e+00 2e+05 4e+05 6e+05
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Console

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

healthcare.r* | HospitalCosts

Source on Save | Run | Source

```
#hence, age, gender and race cannot be used to predict the length of stay of inpatients.
########################6. Complete analysis#########################################################
#The agency wants to find the variable that mainly affects hospital costs.
#Significance method - build a model using all independent variables vs dependent variable
summary(lm(formula = TOTCHG ~ ., data = HospitalCosts))
summary(lm(formula = TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = HospitalCosts))
summary(lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = HospitalCosts))
###########Since APRDRG has -ve t-value, dropping it.
summary(lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCosts))
############################Analysis Conclusion:#########################################################
#As is evident in the multiple models above, health care costs is dependent on age, length of stay and the c
#Healthcare cost is the most for patients in the 0-1 yrs age group category
#Maximum expenditure for 0-1 yr is 678118
#Length of Stay increases the hospital cost

#All Patient Refined Diagnosis Related Groups also affects healthcare costs

#640 diagnosis related group had a max cost of 437978
#Race or gender doesn't have that much impact on hospital cost
```

2:1 (Top Level) R Script

Console

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
# A tibble: 6 × 6
    AGE FEMALE   LOS  RACE TOTCHG APRDRG
  <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
1    17      1     2     1   2660    560
2    17      0     2     1   1689    753
3    17      1     7     1  20060    930
4    17      1     1     1    736    758
5    17      1     1     1   1194    754
6    17      0     0     1   3305    347
> #columns of data#
> names(HospitalCosts)
[1] "AGE"    "FEMALE" "LOS"    "RACE"   "TOTCHG" "APRDRG"
> #1. Record patient statistics:
> #The agency wants to find the age category of people who frequently visit the hospital and has the maximum expen
diture.
> #Age: Age of the patient discharged
> #Totchg: Hospital discharge costs
> summary(HospitalCosts)
      AGE             FEMALE            LOS              RACE            TOTCHG          APRDRG
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000   Min.   :  532   Min.   : 21.0
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1216   1st Qu.:640.0
 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000   Median : 1536   Median :640.0
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078   Mean   : 2774   Mean   :616.4
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.: 2530   3rd Qu.:751.0
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000   Max.   :48388   Max.   :952.0
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000   Max.   :48388   Max.   :952.0
                                                   NA's   :1
> #Get number of hospital visits based on age
> summary(as.factor(HospitalCosts$AGE))
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
307  10   1   3   2   2   2   3   2   2   4   8  15  18  25  29  29  38
> #Total number of hospital for 0-1 age group is 307
> hist(HospitalCosts$AGE, main="Histogram of Age Group and their hospital visits",
+      xlab="Age group", border="black", col=c("light green", "dark green"), xlim=c(0,20), ylim=c(0,350))
> #observation point:As can be seen here, the maximum number of hospital visits are for age group is 0-1 years
> #Summarize expenditure based on age group
> ExpenseBasedOnAge = aggregate(TOTCHG ~ AGE, FUN=sum, data=HospitalCosts)
> which.max(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$TOTCHG, FUN=sum))
678118
    18
> barplot(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$AGE, FUN=sum))
> ############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.
> #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.
> #i)Aprdrg: All Patient Refined Diagnosis Related Groups
> #ii)Totchg: Hospital discharge costs
> summary(as.factor(HospitalCosts$APRDRG))
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317 344 347 42
0 421
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
> #ii)Totchg: Hospital discharge costs
> summary(as.factor(HospitalCosts$APRDRG))
 21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317 344 347 42
0 421
  1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1   1   1   1   2   6   1   1   1   1   2   3
 2   1
422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811 81
2 863
  3   2   1   1   1   3   1   3   6   4   2   3   4 267   1   1   2   1   1  14  36  37  13   2  20   2   1   2
 3   1
911 930 952
  1   2   1
> ##Get the diagnosis-related group and its hospitalization expenditure
> DiagnosisCost = aggregate(TOTCHG ~ APRDRG, FUN = sum, data = HospitalCosts)
> DiagnosisCost[which.max(DiagnosisCost$TOTCHG), ]
   APRDRG TOTCHG
44    640 437978
> #observation:As can be seen here 640 diagnosis related group had a max cost of 437978
> ################3. Race vs Hospitalization costs
> #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.
> #Ho (Null hypothesis):Independent variable (RACE) is not influencing dependent variable (COSTS) #H0:there is no
 no correlation among residuals, # p-value = 0.7394 <== this is > 0.5 #i.e. they are independent #in case of regre
ssion, we need high p value so that we cannot reject the null
> summary(as.factor(HospitalCosts$RACE))
   1    2    3    4    5    6 NA's
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
ssion, we need high p value so that we cannot reject the null
> summary(as.factor(HospitalCosts$RACE))
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
> summary(as.factor(HospitalCosts$RACE))
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
> #obs:There is one null value. This needs to be removed
> HospitalCosts = na.omit(HospitalCosts)
> summary(as.factor(HospitalCosts$RACE))
  1   2   3   4   5   6
484   6   1   3   3   2
> ### As can be seen 484 patients out of 499 fall under group 1, showing that the number of observations for 1 cat
egory is way higher than others - hence data is skewed. This will only affect the results from linear regression o
r ANOVA analysis
> raceInfluence=lm(TOTCHG~ RACE, data=HospitalCosts)
> summary(raceInfluence)

Call:
lm(formula = TOTCHG ~ RACE, data = HospitalCosts)

Residuals:
   Min     1Q Median     3Q    Max
 -2256  -1560  -1227   -258  45600

Coefficients:
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
Residuals:
   Min     1Q Median     3Q    Max
 -2256  -1560  -1227   -258  45600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2925.7      405.0   7.224 1.92e-12 ***
RACE          -137.3      339.1  -0.405    0.686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3895 on 497 degrees of freedom
Multiple R-squared:  0.0003299, Adjusted R-squared:  -0.001681
F-statistic: 0.164 on 1 and 497 DF,  p-value: 0.6856

> #observation
> #pValue is 0.69 it is much higher than 0.5
> #We can say that race doesn’t affect the hospitalization costs
> ####Anaysis using ANOVA
> #We can also use anova statistical test for estimating how dependent variable, in this case RACE, affects the in
dependent variable, the hospitalization cost
> raceInfluenceAOV <- aov(TOTCHG ~ RACE, data=HospitalCosts)
> raceInfluenceAOV
Call:
   aov(formula = TOTCHG ~ RACE, data = HospitalCosts)
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
> raceInfluenceAOV
Call:
   aov(formula = TOTCHG ~ RACE, data = HospitalCosts)

Terms:
                     RACE  Residuals
Sum of Squares     2488459 7539623326
Deg. of Freedom          1        497

Residual standard error: 3894.903
Estimated effects may be unbalanced
> raceInfluenceAOV
Call:
   aov(formula = TOTCHG ~ RACE, data = HospitalCosts)

Terms:
                     RACE  Residuals
Sum of Squares     2488459 7539623326
Deg. of Freedom          1        497

Residual standard error: 3894.903
Estimated effects may be unbalanced
> summary(raceInfluenceAOV)
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          1 2.488e+06  2488459   0.164  0.686
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          1 2.488e+06  2488459   0.164  0.686
Residuals   497 7.540e+09 15170268
> #The residual variance (deviation from original) (of all other variables) is very high. This implies that there
 is very little influence from RACE on hospitalization costs
> #As can be seen, the degree of freedom (Df) for RACE is 1 and that of residuals is 497 observations
> #The F-Value, the test statistic is 0.16 which is much less than 0.5 showing that RACE doesn't affect teh hospit
alization cost.
> #The Pr(>F), the p_value of 0.69 is high confirming that RACE does not affect hospitalization cost.
> ############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.#############################################################
##
> summary(HospitalCosts$FEMALE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   0.511   1.000   1.000
> summary(lm(formula = TOTCHG ~ AGE + FEMALE, data = HospitalCosts)
+ )

Call:
lm(formula = TOTCHG ~ AGE + FEMALE, data = HospitalCosts)

Residuals:
   Min     1Q Median     3Q    Max
 -3403  -1444   -873   -156  44950
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
   Min     1Q Median     3Q    Max
 -3403  -1444   -873   -156  44950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2719.45     261.42  10.403  < 2e-16 ***
AGE            86.04      25.53   3.371 0.000808 ***
FEMALE       -744.21     354.67  -2.098 0.036382 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3849 on 496 degrees of freedom
Multiple R-squared:  0.02585,   Adjusted R-squared:  0.02192
F-statistic: 6.581 on 2 and 496 DF,  p-value: 0.001511

> #Since the pValues of AGE is much lesser than 0.05, the ideal statistical significance level, and it also has th
ree stars (***) next to it, it means AGE has the most statitical significance
> #Similarly, gender is also less than 0.05.
> #Hence, we can conclude that the model is statistically significant
> #############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.#################################
######
> summary(lm(formula = LOS ~ AGE + FEMALE + RACE, data = HospitalCostsNA)
+ )
Error in is.data.frame(data) : object 'HospitalCostsNA' not found
> #Since the pValues of AGE is much lesser than 0.05, the ideal statistical significance level, and it also has th
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
######
> summary(lm(formula = LOS ~ AGE + FEMALE + RACE, data = HospitalCosts))

Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = HospitalCosts)

Residuals:
   Min     1Q Median     3Q    Max
 -3.22  -1.22  -0.85   0.15  37.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94377    0.39318   7.487 3.25e-13 ***
AGE         -0.03960    0.02231  -1.775   0.0766 .
FEMALE       0.37011    0.31024   1.193   0.2334
RACE        -0.09408    0.29312  -0.321   0.7484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.363 on 495 degrees of freedom
Multiple R-squared:  0.007898,  Adjusted R-squared:  0.001886
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692

> ##The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between t
hese variables and length of stay.
> #Hence, age, gender and race cannot be used to predict the length of stay of inpatients
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
> ##The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between t
hese variables and length of stay.
> #Hence, age, gender and race cannot be used to predict the length of stay of inpatients.
> ########################6. Complete analysis#########################################################
> #The agency wants to find the variable that mainly affects hospital costs.
> #Significance method - build a model using all independent variables vs dependent variable
> summary(lm(formula = TOTCHG ~ ., data = HospitalCosts)
+ )

Call:
lm(formula = TOTCHG ~ ., data = HospitalCosts)

Residuals:
   Min     1Q Median     3Q    Max
 -6377   -700   -174    122  43378

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5218.6769   507.6475  10.280  < 2e-16 ***
AGE          134.6949    17.4711   7.710 7.02e-14 ***
FEMALE      -390.6924   247.7390  -1.577    0.115
LOS          743.1521    34.9225  21.280  < 2e-16 ***
RACE        -212.4291   227.9326  -0.932    0.352
APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
---
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5218.6769   507.6475  10.280  < 2e-16 ***
AGE          134.6949    17.4711   7.710 7.02e-14 ***
FEMALE      -390.6924   247.7390  -1.577    0.115
LOS          743.1521    34.9225  21.280  < 2e-16 ***
RACE        -212.4291   227.9326  -0.932    0.352
APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2613 on 493 degrees of freedom
Multiple R-squared:  0.5536,    Adjusted R-squared:  0.5491
F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16

> ##The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between t
hese variables and length of stay.
> #Hence, age, gender and race cannot be used to predict the length of stay of inpatients.
> ########################6. Complete analysis#########################################################
> #The agency wants to find the variable that mainly affects hospital costs.
> #Significance method - build a model using all independent variables vs dependent variable
> summary(lm(formula = TOTCHG ~ ., data = HospitalCosts))

Call:
```

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

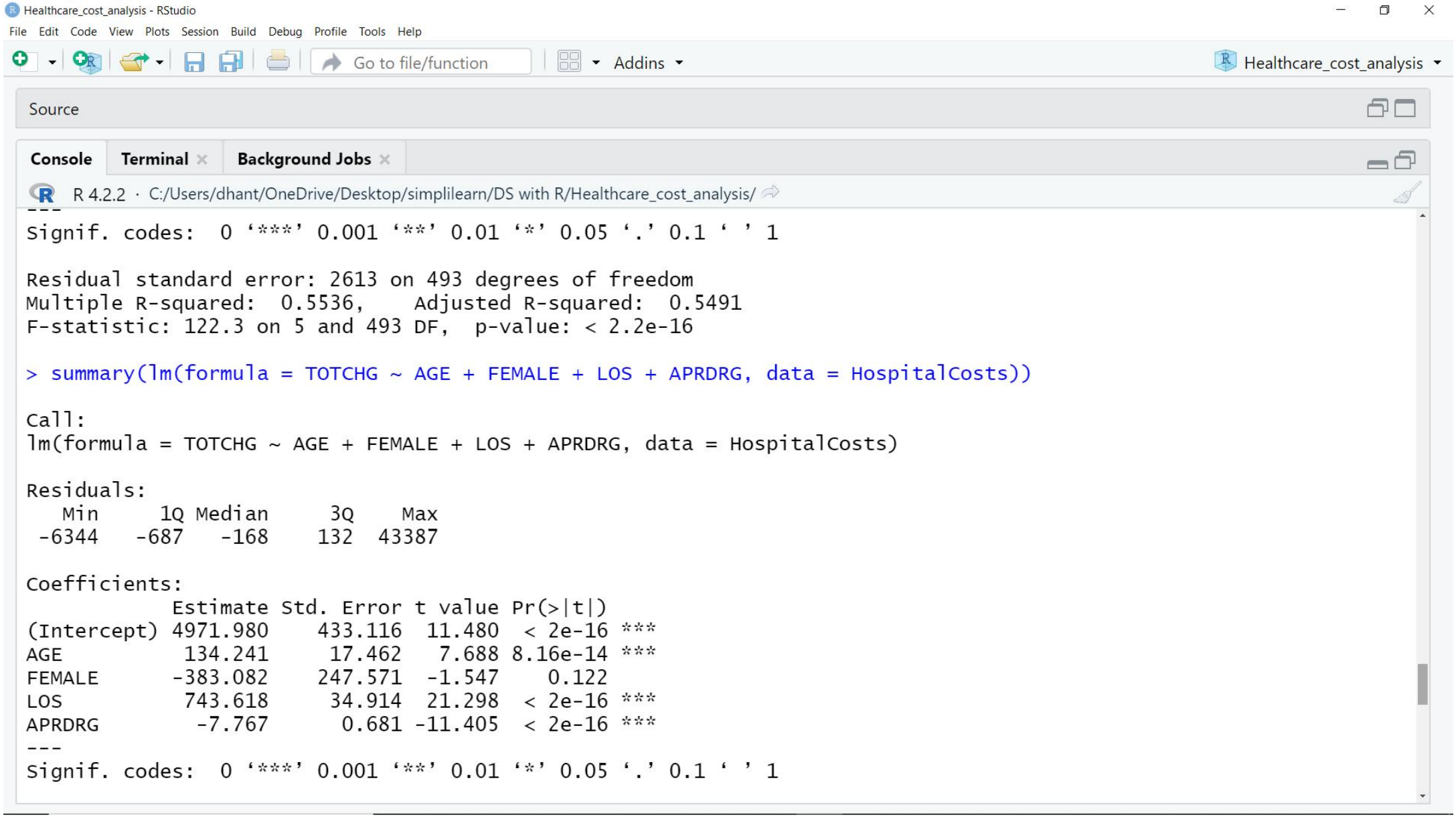
Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

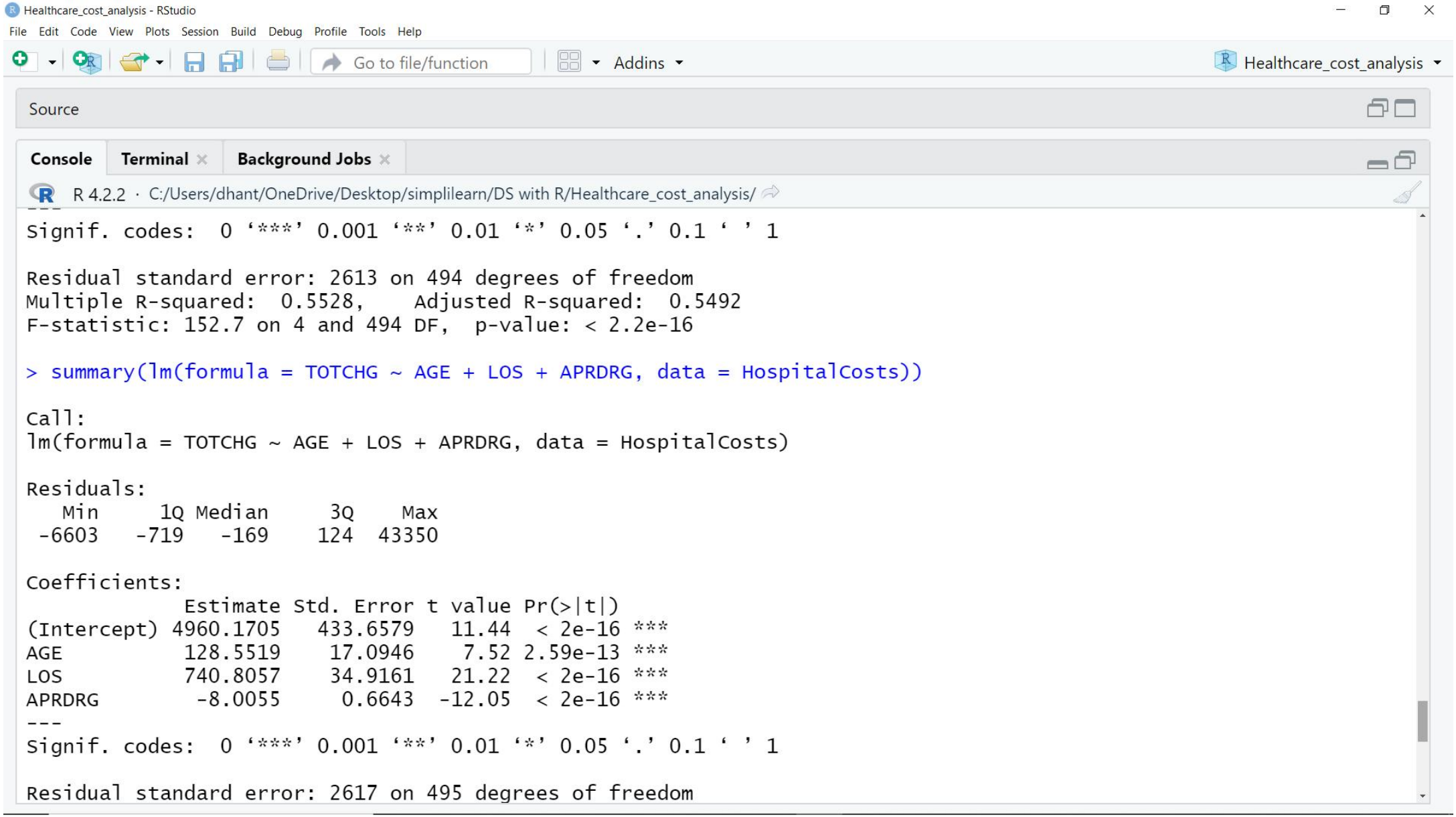
R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
> ##The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between t
hese variables and length of stay.
> #Hence, age, gender and race cannot be used to predict the length of stay of inpatients.
> ########################6. Complete analysis#################################################################
> #The agency wants to find the variable that mainly affects hospital costs.
> #Significance method - build a model using all independent variables vs dependent variable
> summary(lm(formula = TOTCHG ~ ., data = HospitalCosts))

Call:
lm(formula = TOTCHG ~ ., data = HospitalCosts)

Residuals:
   Min     1Q Median     3Q    Max
 -6377   -700   -174    122  43378

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5218.6769   507.6475  10.280  < 2e-16 ***
AGE          134.6949    17.4711   7.710 7.02e-14 ***
FEMALE      -390.6924   247.7390  -1.577    0.115
LOS          743.1521    34.9225  21.280  < 2e-16 ***
RACE        -212.4291   227.9326  -0.932    0.352
APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
---
```


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
Source
Console
Terminal
Background Jobs
R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2613 on 493 degrees of freedom
Multiple R-squared: 0.5536, Adjusted R-squared: 0.5491
F-statistic: 122.3 on 5 and 493 DF, p-value: < 2.2e-16
> summary(lm(formula = TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = HospitalCosts))
Call:
lm(formula = TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = HospitalCosts)
Residuals:
Min 1Q Median 3Q Max
-6344 -687 -168 132 43387
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4971.980 433.116 11.480 < 2e-16 ***
AGE 134.241 17.462 7.688 8.16e-14 ***
FEMALE -383.082 247.571 -1.547 0.122
LOS 743.618 34.914 21.298 < 2e-16 ***
APRDRG -7.767 0.681 -11.405 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
Source
Console
Terminal
Background Jobs
R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2613 on 494 degrees of freedom
Multiple R-squared: 0.5528, Adjusted R-squared: 0.5492
F-statistic: 152.7 on 4 and 494 DF, p-value: < 2.2e-16
> summary(lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = HospitalCosts))
Call:
lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = HospitalCosts)
Residuals:
Min 1Q Median 3Q Max
-6603 -719 -169 124 43350
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4960.1705 433.6579 11.44 < 2e-16 ***
AGE 128.5519 17.0946 7.52 2.59e-13 ***
LOS 740.8057 34.9161 21.22 < 2e-16 ***
APRDRG -8.0055 0.6643 -12.05 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2617 on 495 degrees of freedom

Healthcare_cost_analysis - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

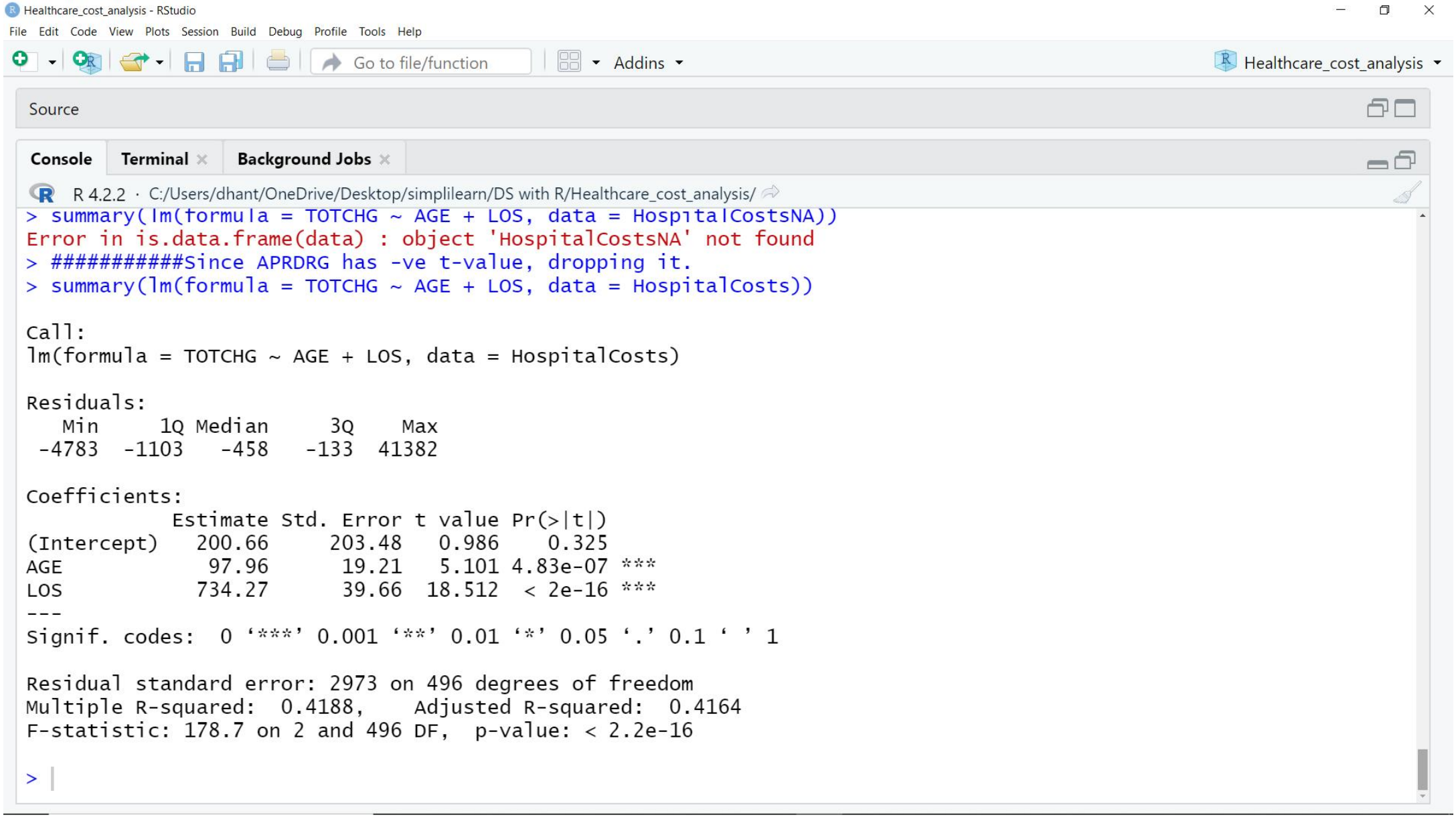
Go to file/function | Addins | Healthcare_cost_analysis

Source

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/

```
APRDRG         -8.0055     0.6643  -12.05  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2617 on 495 degrees of freedom
Multiple R-squared:  0.5506,    Adjusted R-squared:  0.5479
F-statistic: 202.2 on 3 and 495 DF,  p-value: < 2.2e-16

> ###########Since APRDRG has -ve t-value, dropping it.
> summary(lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCostsNA))
Error in is.data.frame(data) : object 'HospitalCostsNA' not found
> ###########Since APRDRG has -ve t-value, dropping it.
> summary(lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCosts))

Call:
lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCosts)

Residuals:
   Min     1Q Median     3Q    Max
 -4783  -1103   -458   -133  41382

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   200.66     203.48   0.986    0.325
AGE            97.96      19.21   5.101 4.83e-07 ***
LOS           734.27      39.66  18.512  < 2e-16 ***
```


Healthcare_cost_analysis - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Healthcare_cost_analysis
Source
Console
Terminal
Background Jobs
R 4.2.2 · C:/Users/dhant/OneDrive/Desktop/simplilearn/DS with R/Healthcare_cost_analysis/
> summary(lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCostsNA))
Error in is.data.frame(data) : object 'HospitalCostsNA' not found
> ##########Since APRDRG has -ve t-value, dropping it.
> summary(lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCosts))
Call:
lm(formula = TOTCHG ~ AGE + LOS, data = HospitalCosts)
Residuals:
Min 1Q Median 3Q Max
-4783 -1103 -458 -133 41382
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 200.66 203.48 0.986 0.325
AGE 97.96 19.21 5.101 4.83e-07 ***
LOS 734.27 39.66 18.512 < 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2973 on 496 degrees of freedom
Multiple R-squared: 0.4188, Adjusted R-squared: 0.4164
F-statistic: 178.7 on 2 and 496 DF, p-value: < 2.2e-16
>