#############------Pinku Pandey------########

##-- 28 April 2021--##

############------Assessment Project-7---#####

Healthcare cost analysis

#Project 7
#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 frequently visit the hospital and has the maximum expenditure.

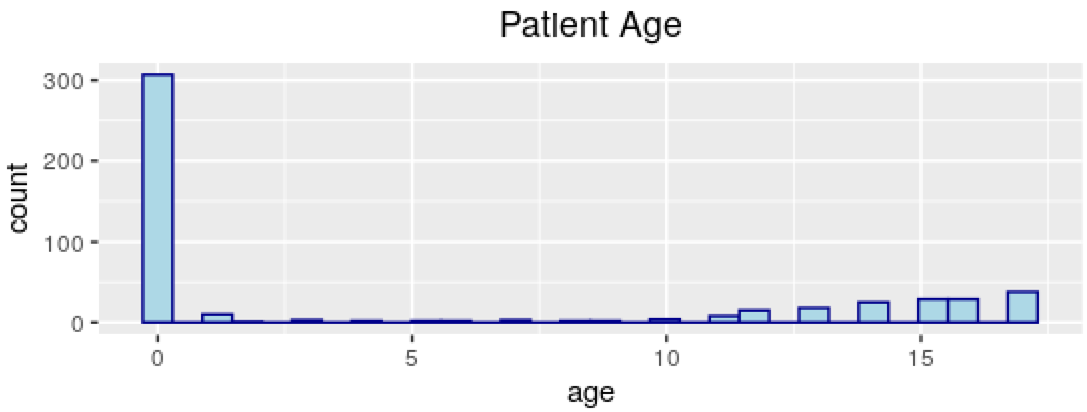
#2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

#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.
#Disclaimer: In Business Analytics, there are different ways of solving the same set of problems. Feel free to explore other ways of answering these questions.
##########---END-------#################

```
############----Main Entry----################

library("readxl")
library(ggplot2)
hospital_cost<-read_excel("1555054100_hospitalcosts.xlsx")
#View(hospital_cost)
head(hospital_cost)
summary(hospital_cost)
```

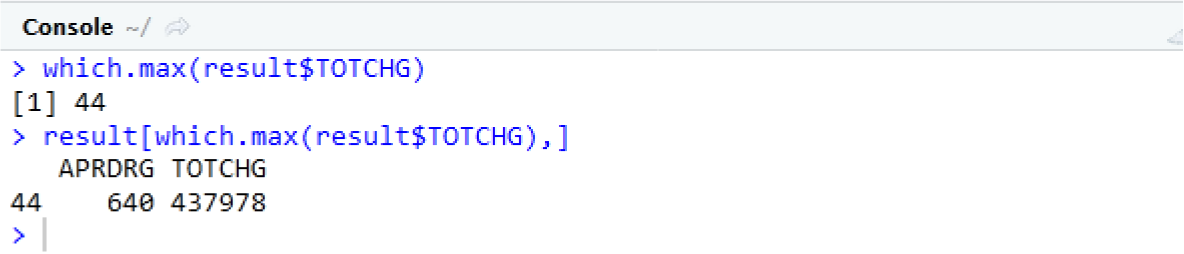
```
##########-----Analysis Part -1 ----##########
```

```
# --- Select Age of patient ---##
age <- hospital_cost$AGE
head(age)
summary(age)
table(age)
##--- Plot Histogram to show age of patient---###
#hist(age)
ggplot(hospital_cost, aes(x=age)) +
  geom_histogram(color="darkblue", fill="lightblue")+
  ggtitle("Patient Age") +
  theme(plot.title = element_text(hjust = 0.5))

summary(as.data.frame(age))
max(table(age))
max(summary(as.factor(age)))
which.max(table(age))
aggregate_age <- aggregate(TOTCHG ~ AGE, data = hospital_cost, sum)
max(aggregate_age)
```



```
#########------#Analysis part -2----######
treatment  <- table(hospital_cost$APRDRG)
```

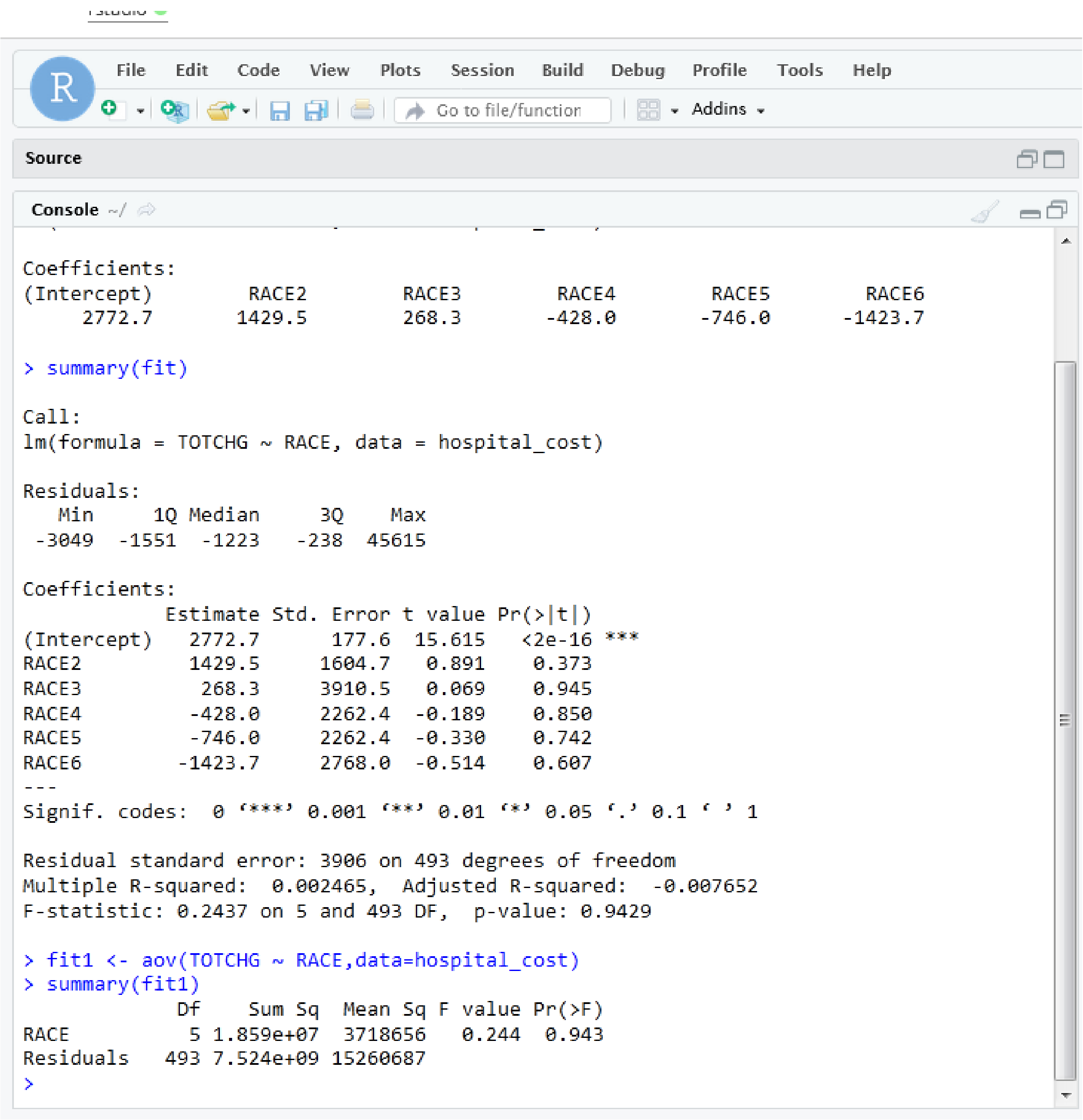
```
treatment
diagnosis <- as.data.frame(treatment)
names(diagnosis)[1] = 'Diagnosis Group'
diagnosis
which.max(table(hospital_cost$APRDRG))
which.max(treatment)
#which.max(diagnosis)
result <- aggregate(TOTCHG ~ APRDRG, data = hospital_cost, sum)
result
which.max(result$TOTCHG)
result[which.max(result$TOTCHG),]
```



```
> which.max(result$TOTCHG)
[1] 44
> result[which.max(result$TOTCHG),]
   APRDRG TOTCHG
44    640 437978
> 
```

#########------#Analysis part -3----######

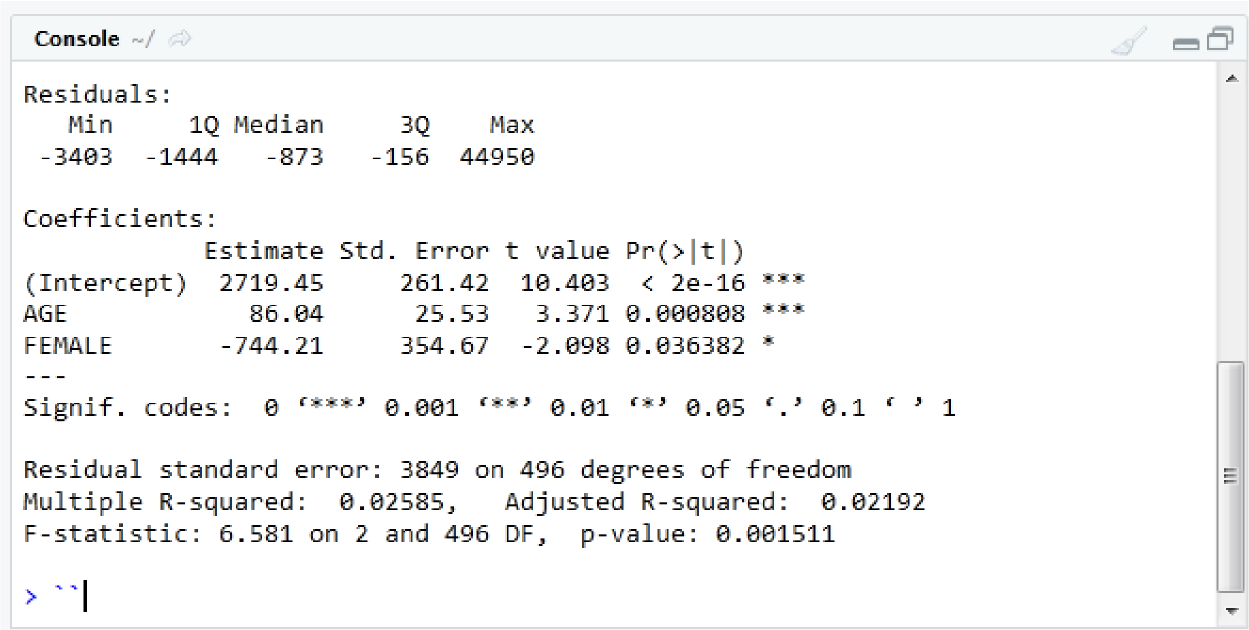
```
table(hospital_cost$RACE)
#class(hospital_cost)
# make factor....
hospital_cost$RACE <- as.factor(hospital_cost$RACE)
fit <- lm(TOTCHG ~ RACE,data=hospital_cost)
fit
summary(fit)
fit1 <- aov(TOTCHG ~ RACE,data=hospital_cost)
summary(fit1)
hospital_cost <- na.omit(hospital_cost)
```



```
Coefficients:
(Intercept)        RACE2        RACE3        RACE4        RACE5        RACE6
     2772.7       1429.5        268.3       -428.0       -746.0      -1423.7

> summary(fit)

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

Residuals:
   Min     1Q Median     3Q    Max
 -3049  -1551  -1223   -238  45615

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2772.7      177.6  15.615   <2e-16 ***
RACE2         1429.5     1604.7   0.891    0.373
RACE3          268.3     3910.5   0.069    0.945
RACE4         -428.0     2262.4  -0.189    0.850
RACE5         -746.0     2262.4  -0.330    0.742
RACE6        -1423.7     2768.0  -0.514    0.607
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3906 on 493 degrees of freedom
Multiple R-squared:  0.002465,  Adjusted R-squared:  -0.007652
F-statistic: 0.2437 on 5 and 493 DF,  p-value: 0.9429

> fit1 <- aov(TOTCHG ~ RACE,data=hospital_cost)
> summary(fit1)
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687
>
```

#########------#Analysis part -4----######

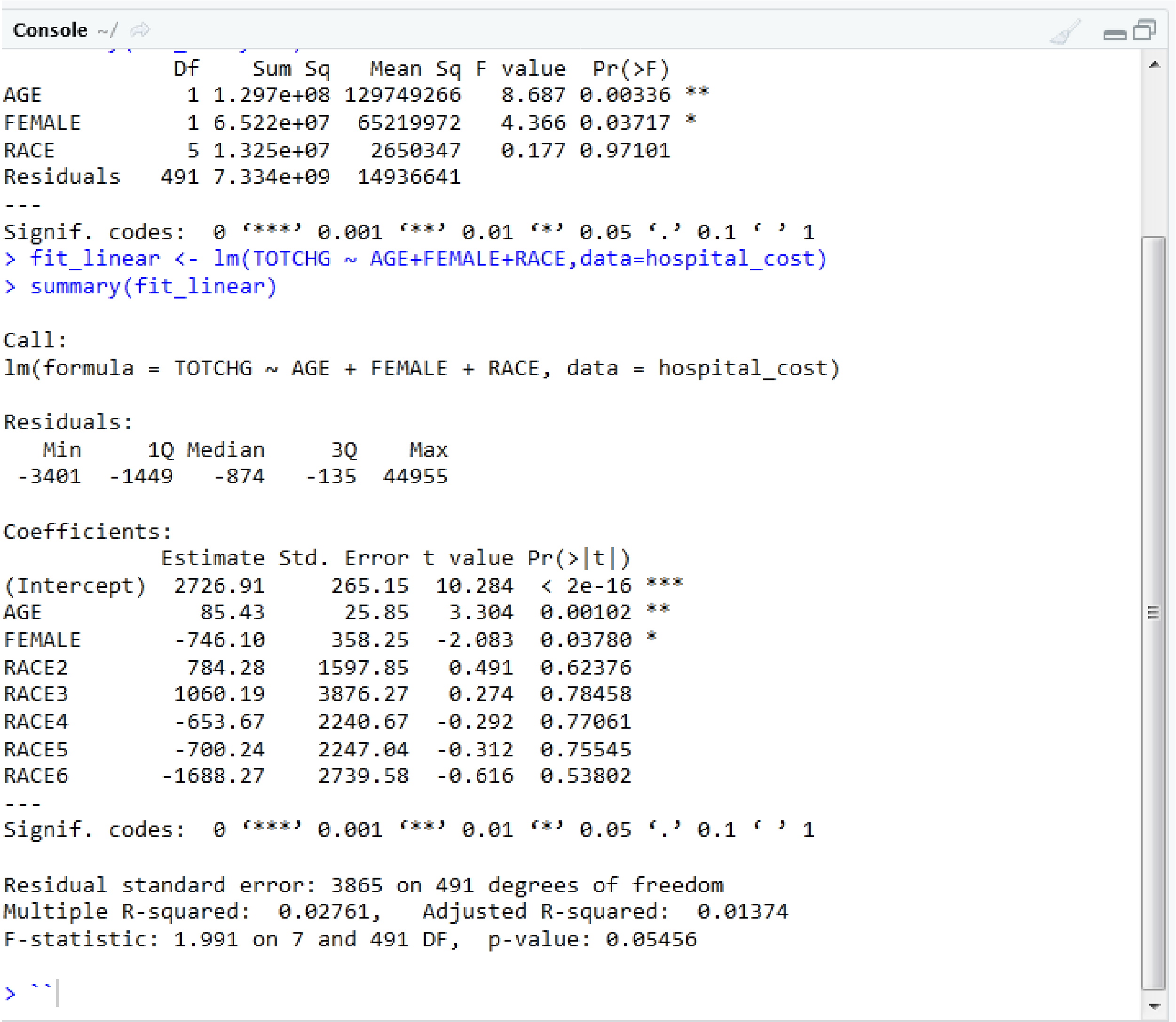
```
table(hospital_cost$FEMALE)
fit_analysis  <- aov(TOTCHG ~ AGE+FEMALE,data=hospital_cost)
summary(fit_analysis)
fit_linear <- lm(TOTCHG ~ AGE+FEMALE,data=hospital_cost)
summary(fit_linear)
```



```
Residuals:
   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

> ``
```

#########------#Analysis part -5----######

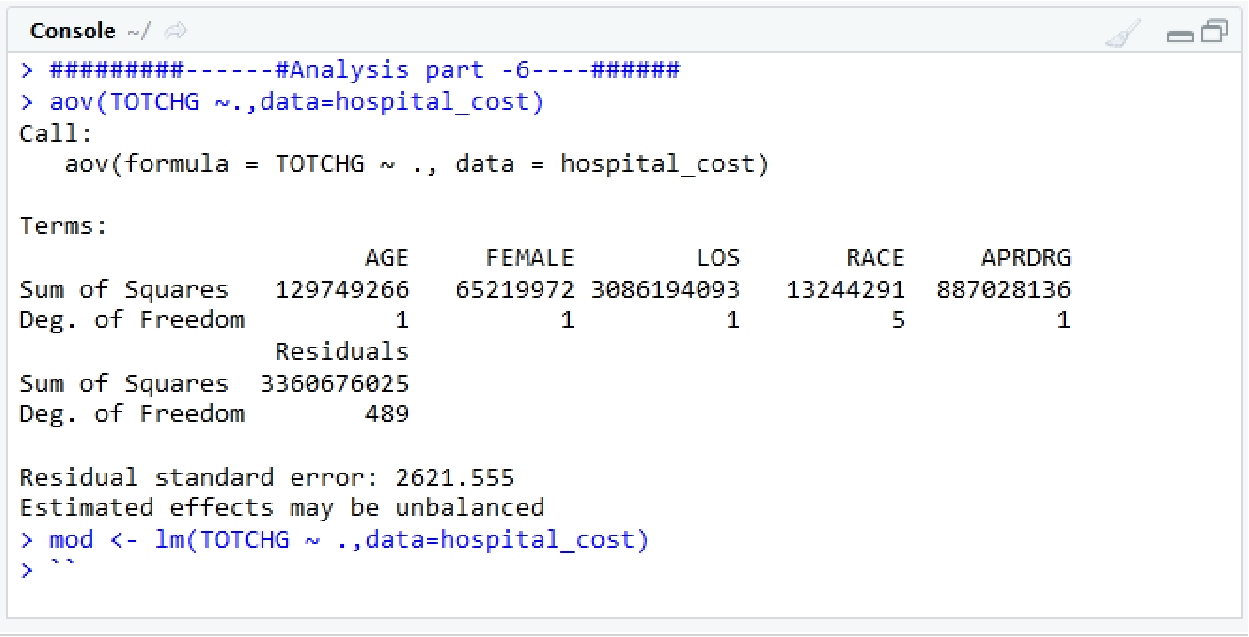
```
table(hospital_cost$LOS)
fit_analysis  <- aov(TOTCHG ~ AGE+FEMALE+RACE,data=hospital_cost)
summary(fit_analysis)
fit_linear <- lm(TOTCHG ~ AGE+FEMALE+RACE,data=hospital_cost)
summary(fit_linear)
```



```
               Df    Sum Sq   Mean Sq F value  Pr(>F)   
AGE             1 1.297e+08 129749266   8.687 0.00336 **
FEMALE          1 6.522e+07  65219972   4.366 0.03717 * 
RACE            5 1.325e+07   2650347   0.177 0.97101   
Residuals     491 7.334e+09  14936641                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> fit_linear <- lm(TOTCHG ~ AGE+FEMALE+RACE,data=hospital_cost)
> summary(fit_linear)

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

Residuals:
   Min     1Q Median     3Q    Max 
 -3401  -1449   -874   -135  44955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2726.91     265.15  10.284  < 2e-16 ***
AGE            85.43      25.85   3.304  0.00102 ** 
FEMALE       -746.10     358.25  -2.083  0.03780 *  
RACE2         784.28    1597.85   0.491  0.62376    
RACE3        1060.19    3876.27   0.274  0.78458    
RACE4        -653.67    2240.67  -0.292  0.77061    
RACE5        -700.24    2247.04  -0.312  0.75545    
RACE6       -1688.27    2739.58  -0.616  0.53802    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3865 on 491 degrees of freedom
Multiple R-squared:  0.02761,	Adjusted R-squared:  0.01374 
F-statistic: 1.991 on 7 and 491 DF,  p-value: 0.05456

> ``
```

#########------#Analysis part -6----######

```
aov(TOTCHG ~.,data=hospital_cost)
mod <- lm(TOTCHG ~ .,data=hospital_cost)
summary(mod)
```



```
> #########------#Analysis part -6----######
> aov(TOTCHG ~.,data=hospital_cost)
Call:
   aov(formula = TOTCHG ~ ., data = hospital_cost)

Terms:
                      AGE     FEMALE        LOS       RACE     APRDRG
Sum of Squares   129749266   65219972 3086194093   13244291  887028136
Deg. of Freedom          1          1          1          5          1
                 Residuals
Sum of Squares  3360676025
Deg. of Freedom        489

Residual standard error: 2621.555
Estimated effects may be unbalanced
> mod <- lm(TOTCHG ~ .,data=hospital_cost)
> ``
```