Project 7 Healthcare cost analysis

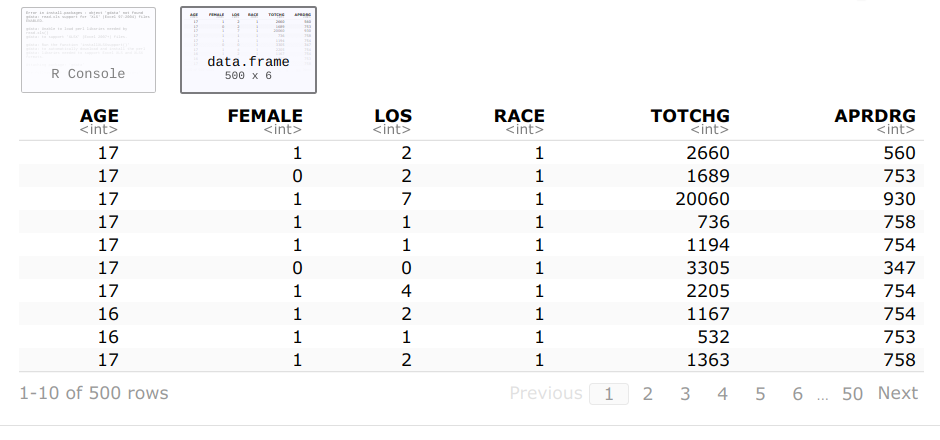
``{r}

install.packages(gdata)

library (gdata)

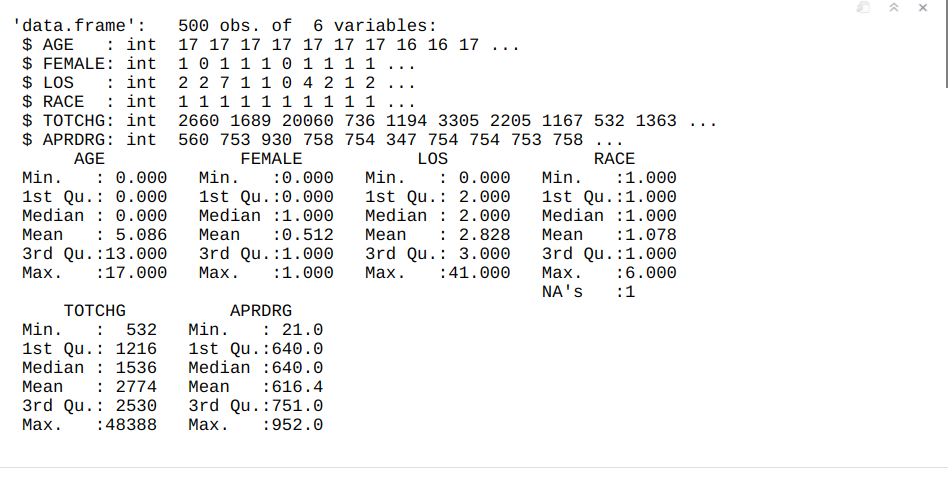
health=read.csv("1555054100\_hospitalcosts - HospitalCosts.csv")

health



str(health)

summary(health)



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.

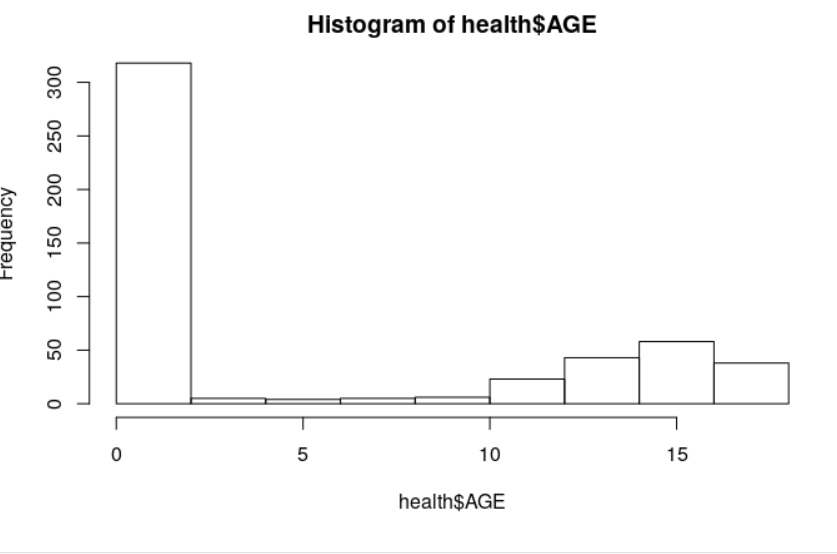
hist(health$AGE)

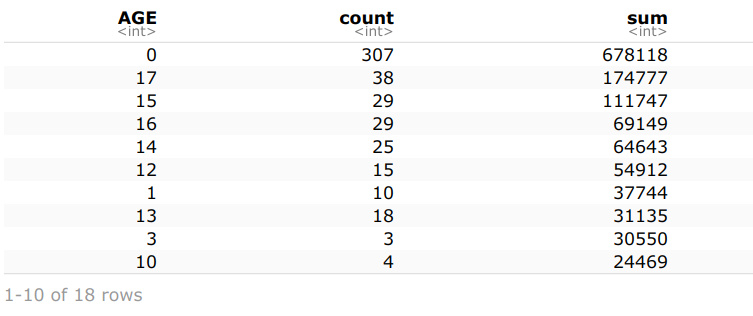
library(dplyr)

arrange(group\_by(health, AGE) %>%

summarise(count=n(),

sum=sum(TOTCHG)), desc(sum))





The age "0" (infant) category of children is a group who frequently visit the hospital (307 times).

Max expenditures are in the group 0 (infant)- $678.118.

healthage = as.factor(health$AGE)

which.max(summary(healthage))

max(summary(healthage))

which.max(tapply(health$TOTCHG,health$AGE,sum))

max(tapply(health$TOTCHG,health$AGE,sum))

0

1

[1] 307

0

1

[1] 678118

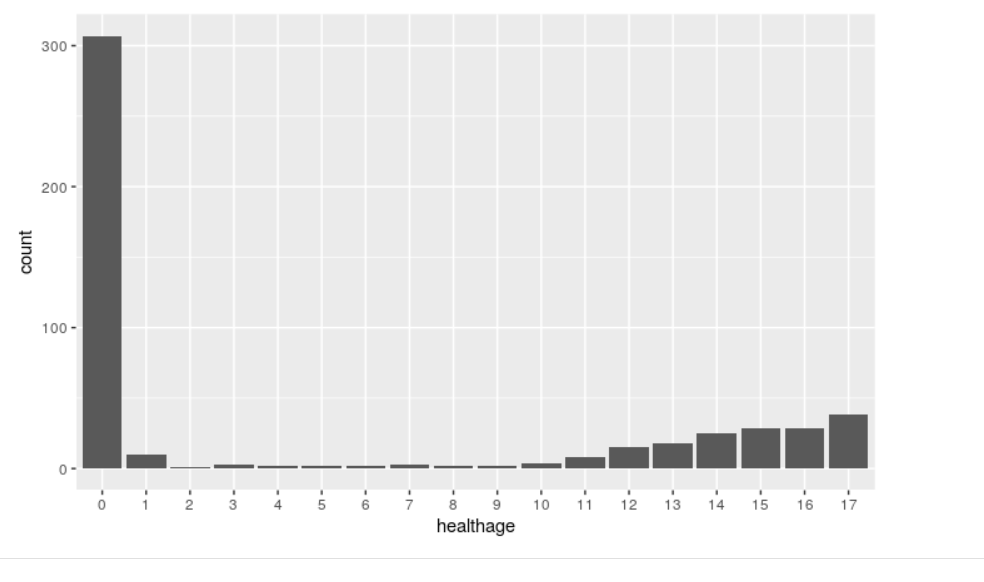
library(ggplot2)

ggplot(health, aes(x=healthage))+geom\_bar()

summary(healthage)

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

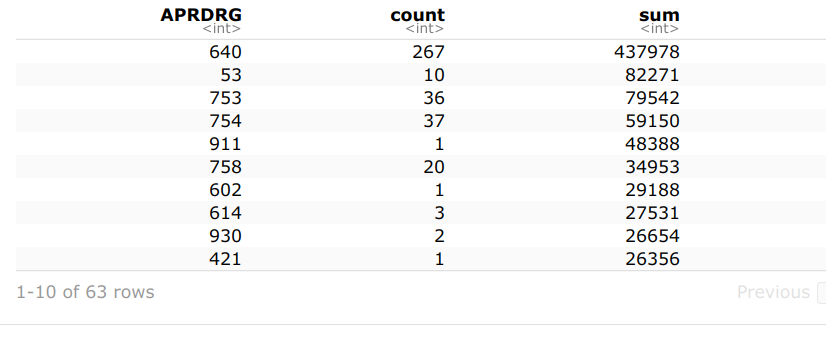


2. In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis-related group that has maximum hospitalization and expenditure.

arrange(group\_by(health, APRDRG) %>%

summarise(count=n(),

sum=sum(TOTCHG)), desc(sum))



From the table we see that the group 640 has maximum hospitalization and costs.

which.max(summary(as.factor(health$APRDRG)))

max(summary(as.factor(health$APRDRG)))

#summary(as.factor(health$APRDRG))

which.max(tapply(health$TOTCHG,health$APRDRG,sum))

max(tapply(health$TOTCHG,health$APRDRG,sum))

640

44

[1] 267

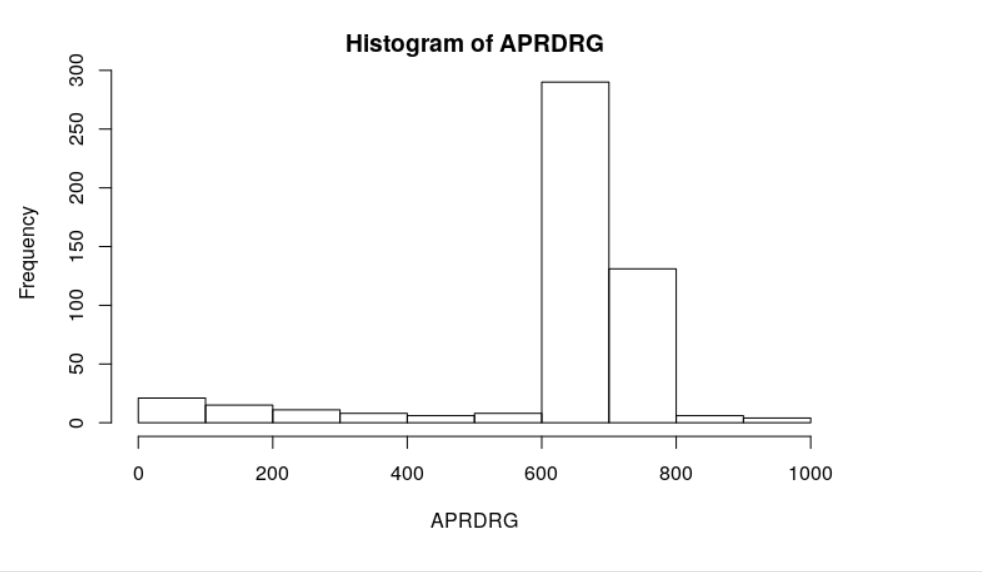
640

44

[1] 437978

attach(health)

hist(APRDRG)



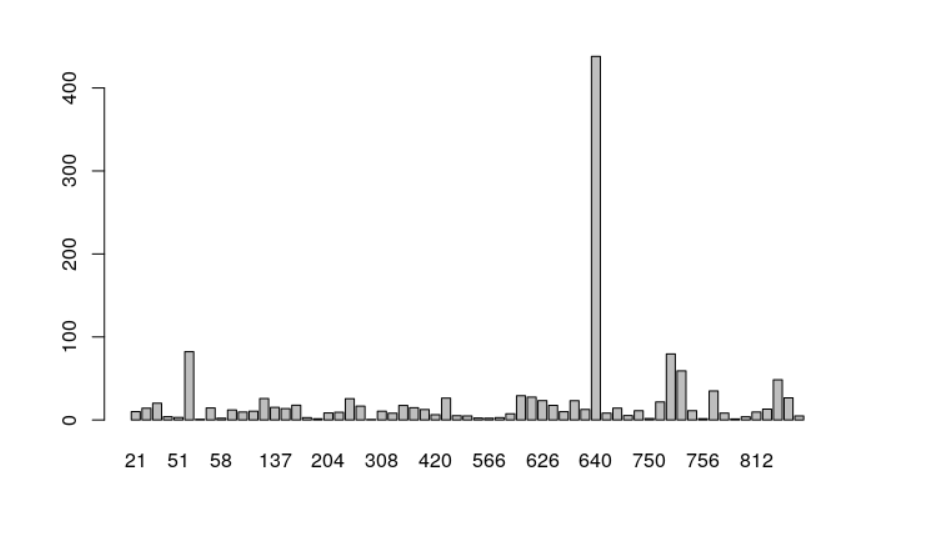
group1=as.factor(health$APRDRG)

health = mutate(health, TOTCHG1 = TOTCHG/1000)

table2 = tapply(health$TOTCHG1,group1,sum)

barplot(table2)

```



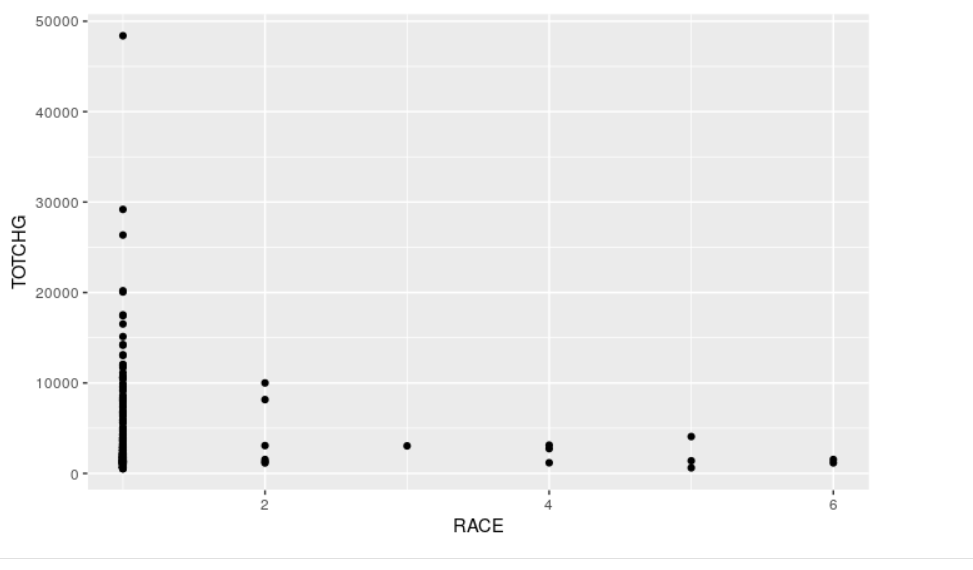
3. To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.

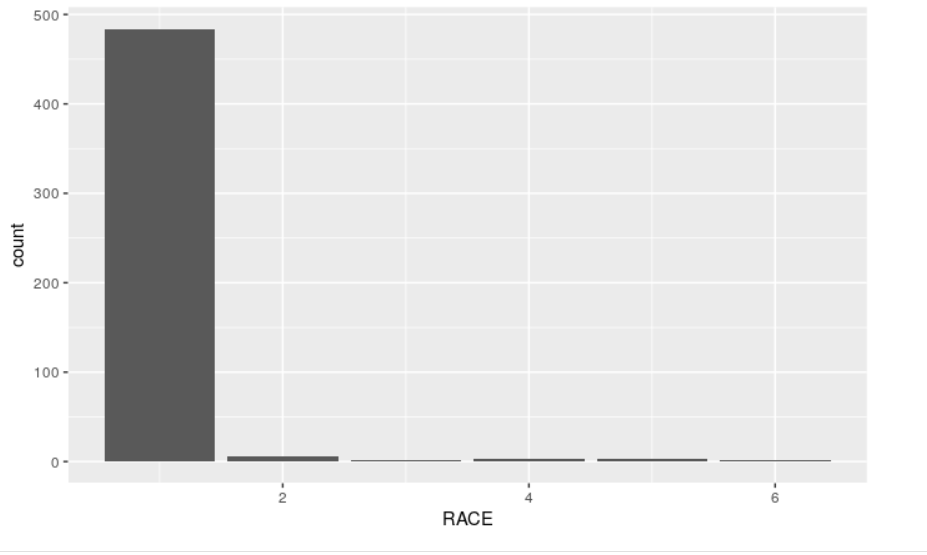
ggplot(health,aes(x=RACE, y=TOTCHG))+geom\_point()

ggplot(health,aes(x=RACE))+geom\_bar()

health$RACE1 = as.factor(health$RACE)

summary(health$RACE1)





1 2 3 4 5 6

484 6 1 3 3 2

health = na.omit(health)

healthrace = as.factor(health$RACE)

health\_anova=aov(TOTCHG~healthrace, data=health)

summary(health\_anova)

Df Sum Sq Mean Sq F value Pr(>F)

healthrace 5 1.859e+07 3718656 0.244 0.943

Residuals 493 7.524e+09 15260687

stripchart(TOTCHG~RACE, vertical = T, pch = 19, data = health)

analysis=lm(TOTCHG~RACE, data=health)

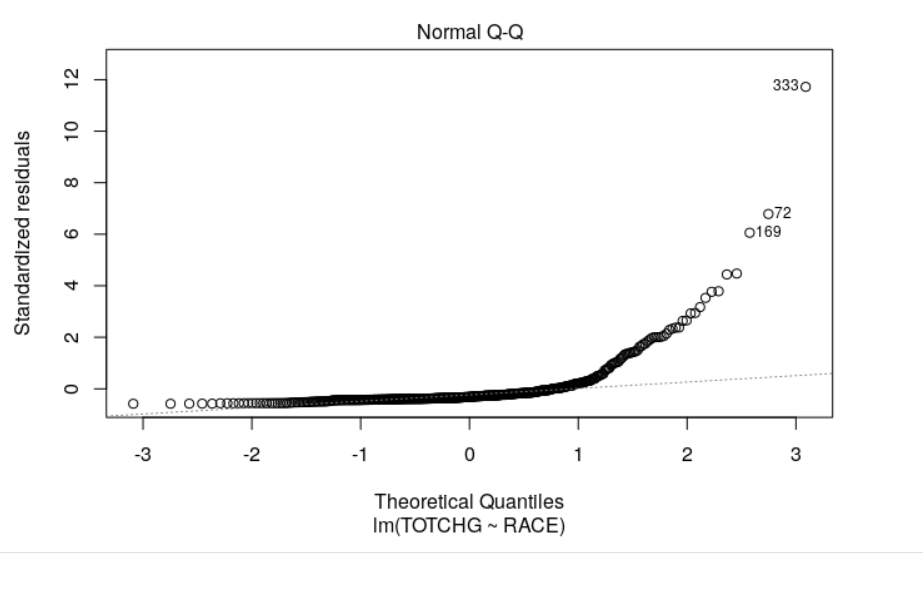
summary(analysis)

anova(analysis)

plot(analysis, which=2)

sresids=rstandard(analysis)

hist(sresids)



In H0 we assume that all patients from different race groups are treated equally and the same treatment and hospital costs are applied.

Therefore:

H0 - the means of the 6 race group are the equal.

H1 - not all of the means are equal

P-value 0.943 is greater than 0.05 significant level so H0 can not be rejected.

It means that there is no relations between race of patient and the hospital costs.

From the summary we can see that RACE 1 has the maximum patients 484 out 500. Also Normal Q-Q chart shows that our data has normal distribution on the beginning and at the end is very much skewed. It effects also Anova results.

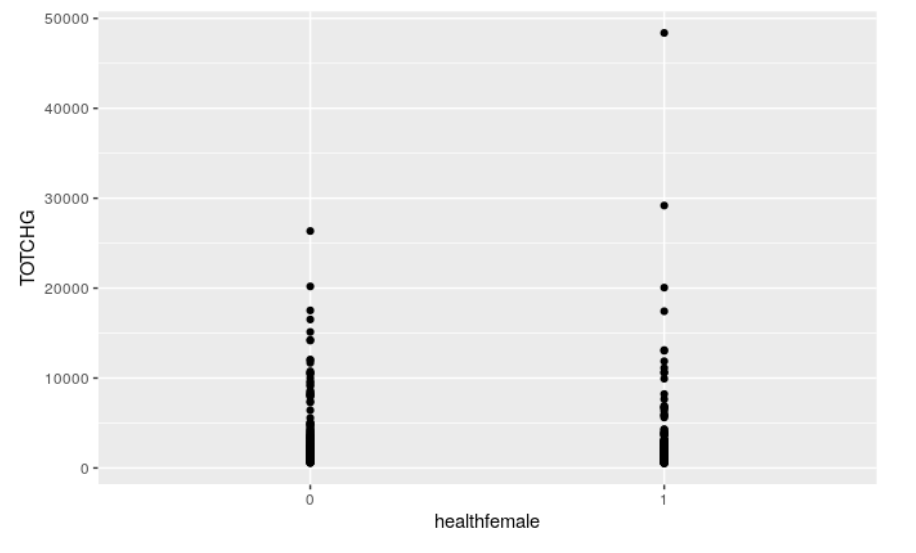
4. To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for the proper allocation of resources.

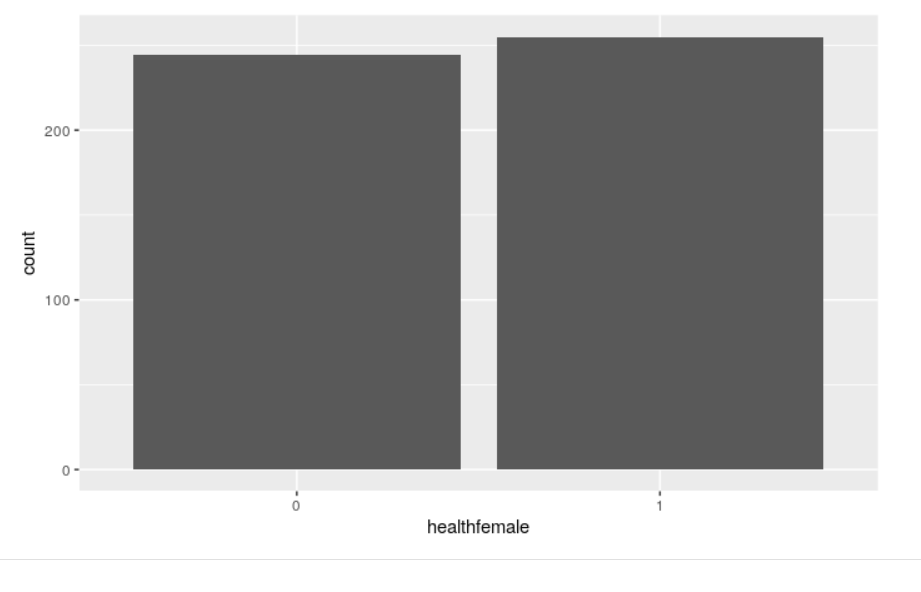
attach(health)

healthfemale = as.factor(health$FEMALE)

ggplot(health,aes(x=healthfemale, y=TOTCHG))+geom\_point()

ggplot(health,aes(x=healthfemale))+geom\_bar()

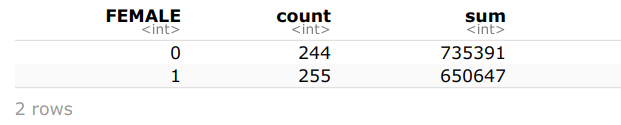




arrange(group\_by(health, FEMALE) %>%

summarise(count=n(),

sum=sum(TOTCHG)), desc(sum))



model1 = lm(TOTCHG~AGE + FEMALE, data=health)

summary(model1)

Call:

lm(formula = TOTCHG ~ AGE + FEMALE, data = health)

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

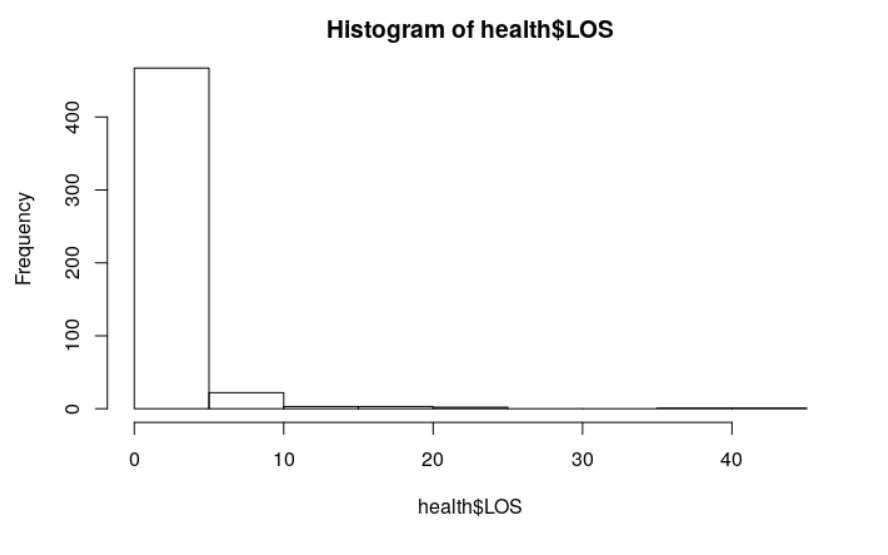
F-statistic: 6.581 on 2 and 496 DF, p-value: 0.001511

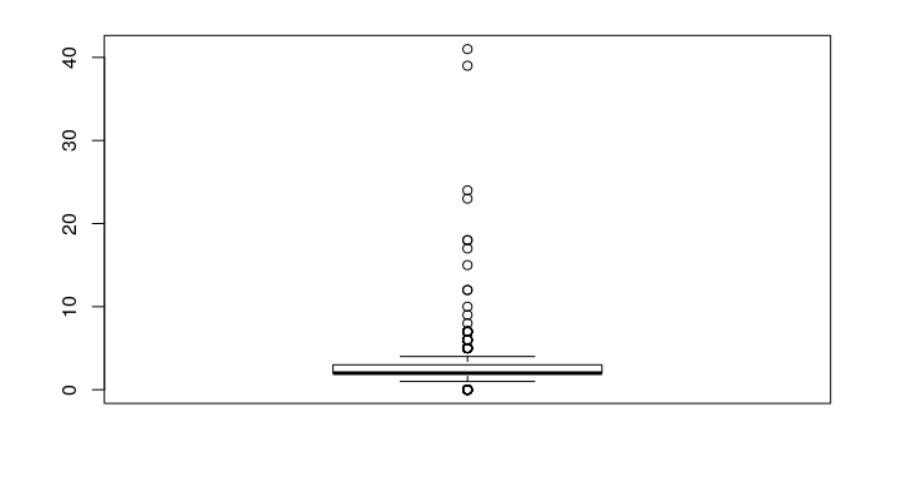
Age has a significant level and lowest p-value that means that Age has affect on hospital costs. It seems there is a relationship between gender and costs too as Female has less p-value. Negative co-efficient in Female shows that female patients uncur less less costs than male. We can prove it also from the summary table.

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.

hist(health$LOS)

boxplot(health$LOS)





Call:

lm(formula = LOS ~ AGE + FEMALE + RACE, data = health)

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 very high p-value means that there is no relationship between independent and dependent variables. Therefore we cannot predict the length of stay from age, gender and race.

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

model3 = lm(TOTCHG~ ., data=health1)

summary(model3)

Call:

lm(formula = TOTCHG ~ ., data = health1)

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 \*\*\*

---

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

Based on our linear regression analysis we can see that AGE, LOS and APRDRG have lower p-value and below 0.05 significant level. That means that age, length of stay and diagnosisand treatments have significant impact on hospital costs. Increasing in the length of stay will increase the hospital cost (1 additional day will cost 743$). It is directly proportional cost.