**Healthcare Cost Analysis**

**Introduction**

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.

**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.**

**R Code & Output:**

#Importing and viewing the dataset##

library(readxl)

hosp <- read\_excel("C:/Business Analytics/Data Science with R Programing/R programming/Project/Projects for Submission/1555054100\_hospitalcosts.xlsx")

head(hosp)

Output:

# A tibble: 6 x 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

#Basic stats of the variables in the dataset##

summary(hosp)

Output:

AGE FEMALE LOS RACE TOTCHG

Min. : 0.000 Min. :0.000 Min. : 0.000 Min. :1.000 Min. : 532

1st Qu.: 0.000 1st Qu.:0.000 1st Qu.: 2.000 1st Qu.:1.000 1st Qu.: 1216

Median : 0.000 Median :1.000 Median : 2.000 Median :1.000 Median : 1536

Mean : 5.086 Mean :0.512 Mean : 2.828 Mean :1.078 Mean : 2774

3rd Qu.:13.000 3rd Qu.:1.000 3rd Qu.: 3.000 3rd Qu.:1.000 3rd Qu.: 2530

Max. :17.000 Max. :1.000 Max. :41.000 Max. :6.000 Max. :48388

NA's :1

#Solution

attach(hosp)

#to see the value of category of infants

ag<-as.factor(AGE)

summary(ag)

Output:

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

#Inpatients Age Distribution

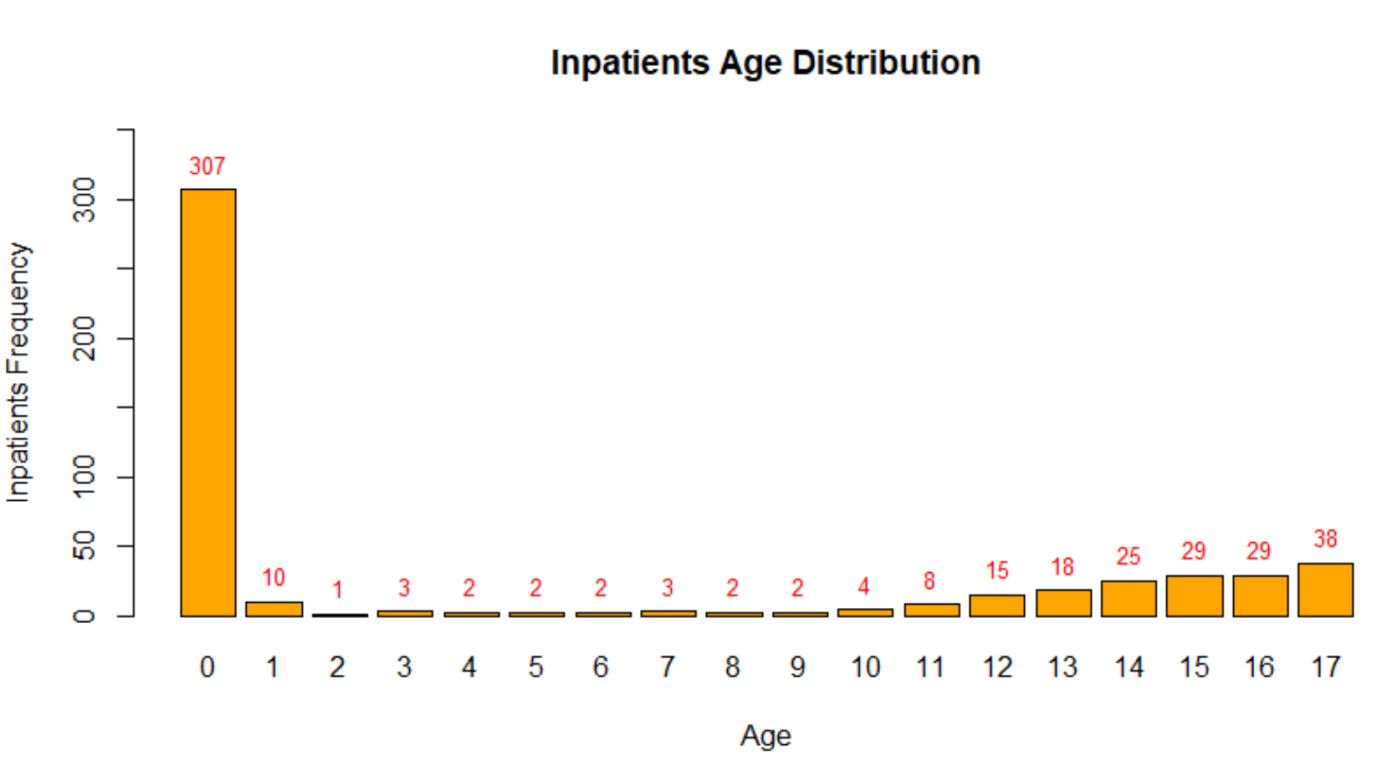
agebar = summary(ag)

agebarplt = barplot(agebar,xlab="Age",ylab="Inpatients Frequency",col="orange",main ='Inpatients Age Distribution',ylim = c(0,350))

## Add text at top of bars

text(x = agebarplt, y = agebar, label = agebar, pos = 3, cex = 0.8, col = "red")

Output Plot:



**Insights: From the above graph we can say that the among all the inpatients the infants (Age=0) visit the hospital most frequently, followed by the teenagers aging from 13 to 17.**

#Highest expenditure among all patients wrt age

tapply(TOTCHG,AGE,max)

maxexp= tapply(TOTCHG,AGE,max)

Output:

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

29188 9606 7298 14243 9230 10584 9530 6425 3588 10585 17524 3908 17434 5615 10756 20195 10002 48388

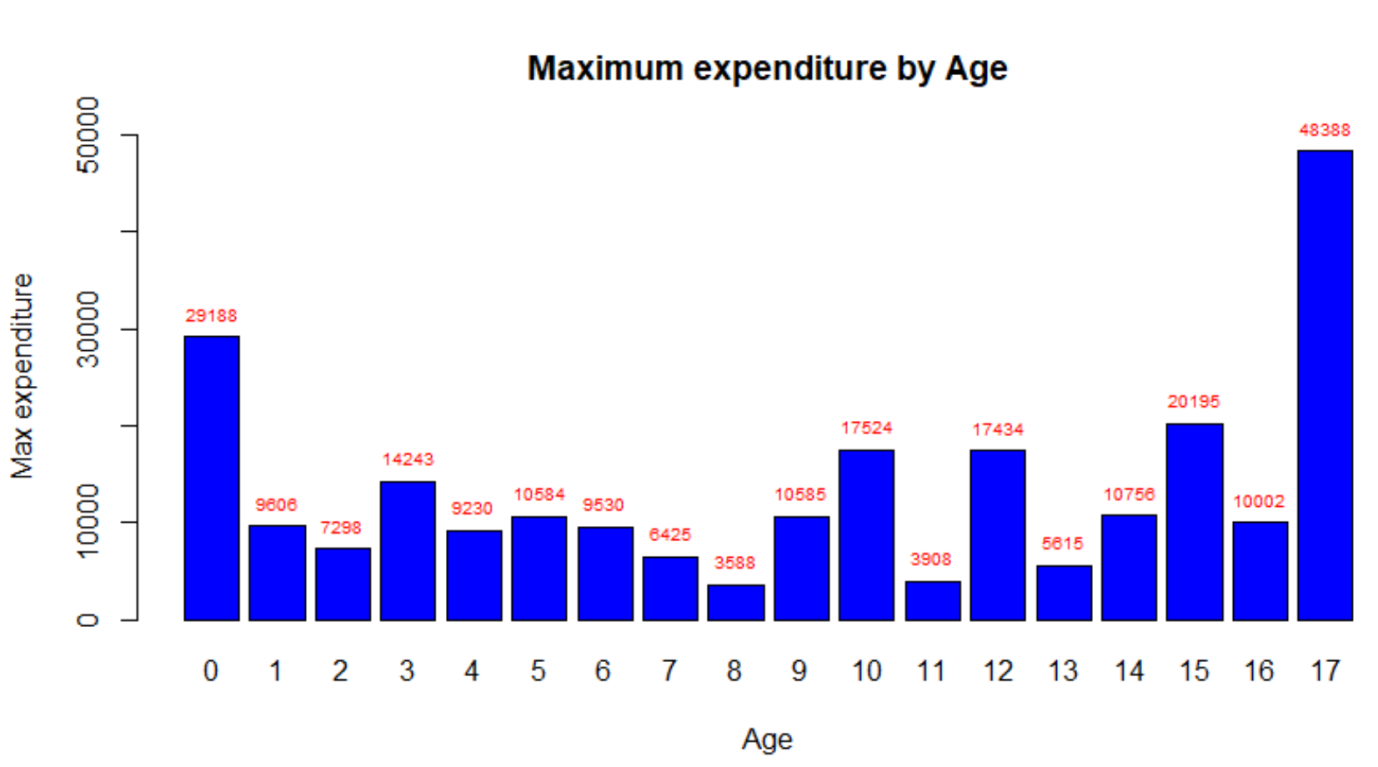
#Max expenditure distribution by age

maxexpplt=barplot(maxexp,xlab="Age",ylab="Max expenditure",col="blue",main = 'Maximum expenditure by Age',ylim = c(0,50000))

## Add text at top of bars

text(x = maxexpplt, y = maxexp, label = maxexp, pos = 3, cex = 0.6, col = "red")

Output Plot:



**Insights:** **From the above graph we can say that the inpatients with age group of 17 and 0(infants) have maximum expenditure.**

1. **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.**

**R Code & Output:**

#to see count of distinct APRDRG : All Patient Refined Diagnosis Related Groups

aprdg<-as.factor(APRDRG)

summary(aprdg)

Output:

21 23 49 50 51 53 54 57 58 92 97 114 115 137 138 139 141 143 204 206 225 249 254 308

1 1 1 1 1 10 1 2 1 1 1 1 2 1 4 5 1 1 1 1 2 6 1 1

313 317 344 347 420 421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740

1 1 2 3 2 1 3 2 1 1 1 3 1 3 6 4 2 3 4 267 1 1 2 1

750 751 753 754 755 756 758 760 776 811 812 863 911 930 952

1 14 36 37 13 2 20 2 1 2 3 1 1 2 1

#Finding the group with max frequency

which.max(summary(aprdg))

Output: 640

44

#APRDRG Distribution

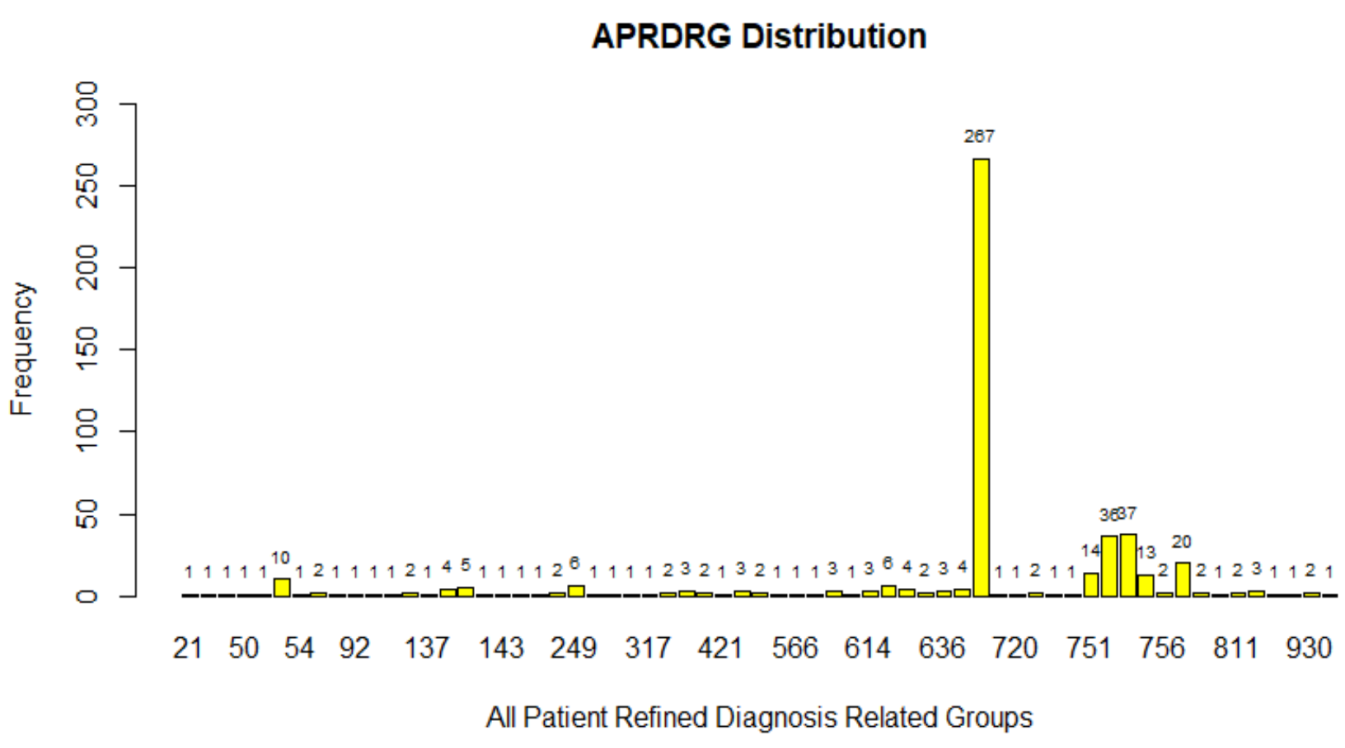
aprdgs = summary(aprdg)

aprdgplt=barplot(aprdgs,xlab="All Patient Refined Diagnosis Related Groups",ylab="Frequency",col="yellow",main ='APRDRG Distribution',ylim = c(0,300))

## Add text at top of bars

text(x = aprdgplt, y = aprdgs, label = aprdgs, pos = 3, cex = 0.6, col = "black")

Output Plot:



**Insights:** **From the above graph and R code we can say that the; Diagnosis Related Group no. 640 has the maximum (count =240) entries of hospitalization.**

#Diagnosis Group with maximum avg hospt expendiure

tapply(TOTCHG,APRDRG,mean)

maxexpdigs= tapply(TOTCHG,APRDRG,mean)

Output:

21 23 49 50 51 53 54 57

10002.0000 14174.0000 20195.0000 3908.0000 3023.0000 8227.1000 851.0000 7254.5000

58 92 97 114 115 137 138 139

2117.0000 12024.0000 9530.0000 10562.0000 12916.0000 15129.0000 3405.5000 3553.2000

141 143 204 206 225 249 254 308

2860.0000 1393.0000 8439.0000 9230.0000 12824.5000 2773.6667 615.0000 10585.0000

313 317 344 347 420 421 422 560

8159.0000 17524.0000 7401.0000 4199.0000 3178.5000 26356.0000 1725.6667 2438.5000

561 566 580 581 602 614 626 633

2296.0000 2129.0000 2825.0000 2484.3333 29188.0000 9177.0000 3881.5000 4397.7500

634 636 639 640 710 720 723 740

4976.0000 7741.3333 3153.0000 1640.3670 8223.0000 14243.0000 2644.5000 11125.0000

750 751 753 754 755 756 758 760

1753.0000 1547.5714 2209.5000 1598.6486 859.0769 747.0000 1747.6500 4136.5000

776 811 812 863 911 930 952

1193.0000 1919.0000 3174.6667 13040.0000 48388.0000 13327.0000 4833.0000

#Diagnosis Group with maximum avg hospt expenditure & its Value

which.max(maxexpdigs) max(maxexpdigs)

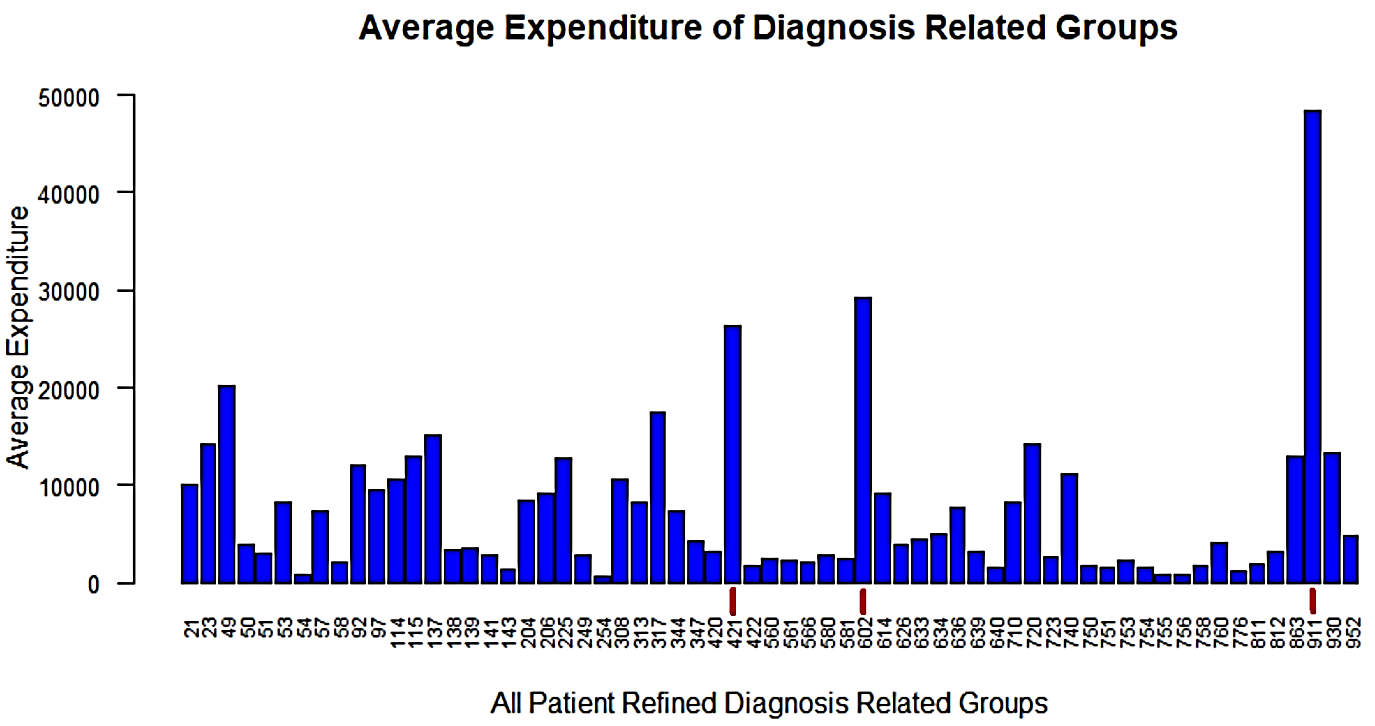
Output: 911 Output: 48388

61

#All Patient Refined Diagnosis Related Groups(APRDRG) Vs Avg hospt expenditure

maxexpdigsplt=barplot(maxexpdigs,xlab='All Patient Refined Diagnosis Related Groups',ylab="Average Expenditure",col="blue",main = 'Average Expenditure of Diagnosis Related Groups',ylim = c(0,50000),cex.axis=0.8, cex.names =.7, las=2)

Output Plot:



**Insights:** **From the above graph and R code we can say that the; Diagnosis Related Group no. 911 has the maximum hospitalization cost, the groups 602 & 421 also demand a relatively high treatment and hospitalization cost.**

1. **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.**

**R Code & Output:**

#h0:The race of the patient is related to the hospitalization costs.

#ha:no relation

rc<-as.factor(RACE)

summary(rc)

Output: 1 2 3 4 5 6 NA's

484 6 1 3 3 2 1

#now to omit na values from data set

hospna<-na.omit(hosp)

#Applying ANOVA

modelannova<-aov(TOTCHG~RACE)

summary(modelannova)

Output: 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

**Insights:** **pvalue comes out to be very high 68% this means we can take risk and reject the null hypothesis; this means there is no relation between the race of patient and the hospital cost.**

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

**R Code & Output:**

#Fitting Linear Model

modelm1<-lm(TOTCHG~AGE+FEMALE)

summary(modelm1)

Output:

Call:

lm(formula = TOTCHG ~ AGE + FEMALE)

Residuals:

Min 1Q Median 3Q Max

-3406 -1443 -869 -152 44951

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2718.63 261.14 10.411 < 2e-16 \*\*\*

AGE 86.28 25.48 3.387 0.000763 \*\*\*

FEMALE -748.19 353.83 -2.115 0.034967 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3845 on 497 degrees of freedom

Multiple R-squared: 0.0261, Adjusted R-squared: 0.02218

F-statistic: 6.66 on 2 and 497 DF, p-value: 0.001399

**Insights:** **pvalue for Age is very less this means it is an important factor in the hospital costs as seen by the significance levels and p-values and gender has also less p value means it is also having the impact on cost and same with intercept.**

1. **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.**

**R Code & Output:**

#Fitting Linear Model

modelm2<-lm(TOTCHG~AGE+FEMALE+RACE)

summary(modelm2)

Output: Call:

lm(formula = LOS ~ AGE + FEMALE + RACE)

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

(1 observation deleted due to missingness)

Multiple R-squared: 0.007898, Adjusted R-squared: 0.001886

F-statistic: 1.314 on 3 and 495 DF, p-value: 0.2692

**Insights:** **except for the intercept.**

**#The very high p-value signifies that there is no linear relationship between the given variables.**

**#That is, with just the age, gender, and race, it is not possible to predict the LOS of a patient.**

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

**R Code & Output:**

#Fitting Linear Model

modelm3<-lm(hospna$TOTCHG~ .,data=hospna)

summary(modelm3)

OutPut:

Call:

lm(formula = hospna$TOTCHG ~ ., data = hospna)

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

**Insights:** **except for the Gender and Race of thre patients**

**#The less p-value signifies that there is linear relationship between the given variables.**

**#That is, with the Age,Length of Stay and Diagnosis Group, it is possible to predict the hospital costs of a patient.**