Healthcare Cost Analysis

Swathi V Hebbar

22/12/2020

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.
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.

So, lets start by loading the data.

```
library(readxl)
Healthcare<- read_excel(file.choose())
View(Healthcare)
str(Healthcare)
```

```
## tibble [500 x 6] (S3: tbl_df/tbl/data.frame)
##  $ AGE   : num [1:500] 17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE: num [1:500] 1 0 1 1 1 0 1 1 1 1 ...
##  $ LOS   : num [1:500] 2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : num [1:500] 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num [1:500] 2660 1689 20060 736 1194 ...
##  $ APRDRG: num [1:500] 560 753 930 758 754 347 754 754 753 758 ...
```

Now as we can see, there are 6 variables. In that, there are several variables which needs to be categorised.

```
Healthcare$AGE <- as.factor(Healthcare$AGE)
Healthcare$FEMALE <- as.factor(Healthcare$FEMALE)
Healthcare$APRDRG <- as.factor(Healthcare$APRDRG)
str(Healthcare)
```

```
## tibble [500 x 6] (S3: tbl_df/tbl/data.frame)
##  $ AGE   : Factor w/ 18 levels "0","1","2","3",..: 18 18 18 18 18 18 18 17 17 18 ...
##  $ FEMALE: Factor w/ 2 levels "0","1": 2 1 2 2 2 1 2 2 2 2 ...
```

```
##  $ LOS   : num [1:500] 2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : num [1:500] 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num [1:500] 2660 1689 20060 736 1194 ...
##  $ APRDRG: Factor w/ 63 levels "21","23","49",..: 32 51 62 55 52 28 52 52 51 55 ...
```

Now we remove NA in the data set, if there are any.

```
Healthcare <- na.omit(Healthcare)
sum(is.na(Healthcare))
```

```
## [1] 0
```

Now we check for outliers and try to do the capping.

```
 table(Healthcare$AGE)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
## 306  10   1   3   2   2   2   3   2   2   4   8  15  18  25  29  29  38
```

```
 table(Healthcare$FEMALE)
```

```
##
##   0   1
## 244 255
```

```
 table(Healthcare$LOS)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  12  15  17  18  23  24  39  41
##  15  79 223  98  38  14   8  11   1   1   1   2   1   1   2   1   1   1   1
```

```
 Healthcare$LOS[Healthcare$LOS>4] <- 4 #Since there are less no. of people in LOS>4, we merge
 table(Healthcare$LOS)
```

```
##
##   0   1   2   3   4
##  15  79 223  98  84
```

```
 table(Healthcare$RACE)
```

```
##
##   1   2   3   4   5   6
## 484   6   1   3   3   2
```

```
 Healthcare$RACE[Healthcare$RACE>2] <- 2 #since there are less no. of people in Race>2, we merge
 table(Healthcare$RACE)
```

```
##
##   1   2
## 484  15
```

```
 table(Healthcare$APRDRG)
```

```
##
##  21  23  49  50  51  53  54  57  58  92  97 114 115 137 138 139 141 143 204 206
##   1   1   1   1   1  10   1   2   1   1   1   1   2   1   4   5   1   1   1   1
## 225 249 254 308 313 317 344 347 420 421 422 560 561 566 580 581 602 614 626 633
##   2   6   1   1   1   1   2   3   2   1   3   2   1   1   1   3   1   3   6   4
## 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811 812 863
##   2   3   4 266   1   1   2   1   1  14  36  37  13   2  20   2   1   2   3   1
## 911 930 952
```

```
##   1   2   1
```

```
summary(Healthcare$TOTCHG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     532    1218    1538    2778    2530   48388
```

Now that our data is free of outliers, we arrange the data set in the order of age.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Healthcare<-arrange(Healthcare,AGE)
```

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.

```
table(Age=Healthcare$AGE,Length_of_stay=Healthcare$LOS)
```

```
##     Length_of_stay
## Age    0   1   2   3   4
##   0    3  15 169  77  42
##   1    0   5   3   1   1
##   2    0   0   1   0   0
##   3    0   1   0   0   2
##   4    0   1   0   1   0
##   5    0   0   1   1   0
##   6    0   0   1   1   0
##   7    2   1   0   0   0
##   8    0   1   1   0   0
##   9    0   1   0   1   0
##   10   1   0   2   0   1
##   11   1   3   4   0   0
##   12   0   4   6   3   2
##   13   1   7   6   1   3
##   14   0   9   6   1   9
##   15   3  11   2   2  11
##   16   3  10  10   3   3
##   17   1  10  11   6  10
```

Here, from this table, we get to know the number of people who were hospitalized by their age. To get a overall view we can even try the below command. And from both the results, we can see that, infants whose age is less than 1 has more hospitalization.

```
library(magrittr)
Los_report <- Healthcare %>% group_by(Age=Healthcare$AGE) %>% summarise(Total_lOS=sum(LOS))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

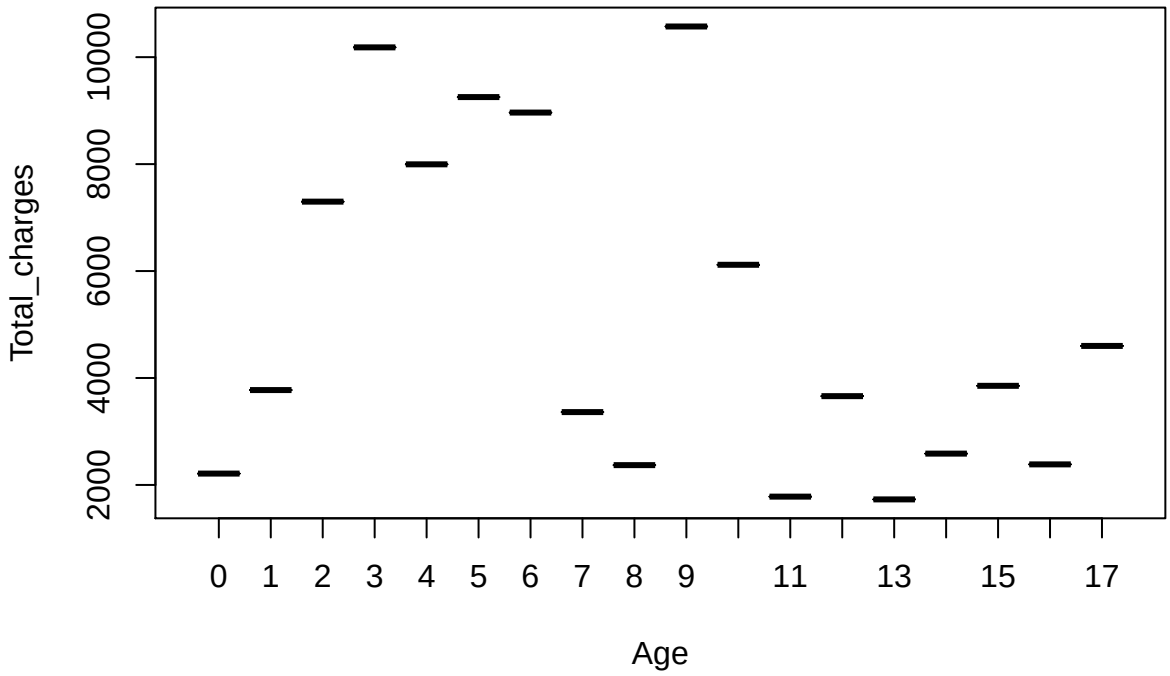
```
View(Los_report)
```

In order to get the Maximum expenditure, you can analyse the data in 2 views. 1) If we take mean of hospitalization cost for each age category, we get an idea of every person's average hospitalization cost in respective age.

```
Age_Max_report_mean <- Healthcare %>% group_by(Age=Healthcare$AGE) %>% summarise(Total_charges=mean(TOT
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Age_Max_report_mean <- Age_Max_report_mean %>% arrange(desc(Total_charges))
View(Age_Max_report_mean)
plot(Age_Max_report_mean)
```



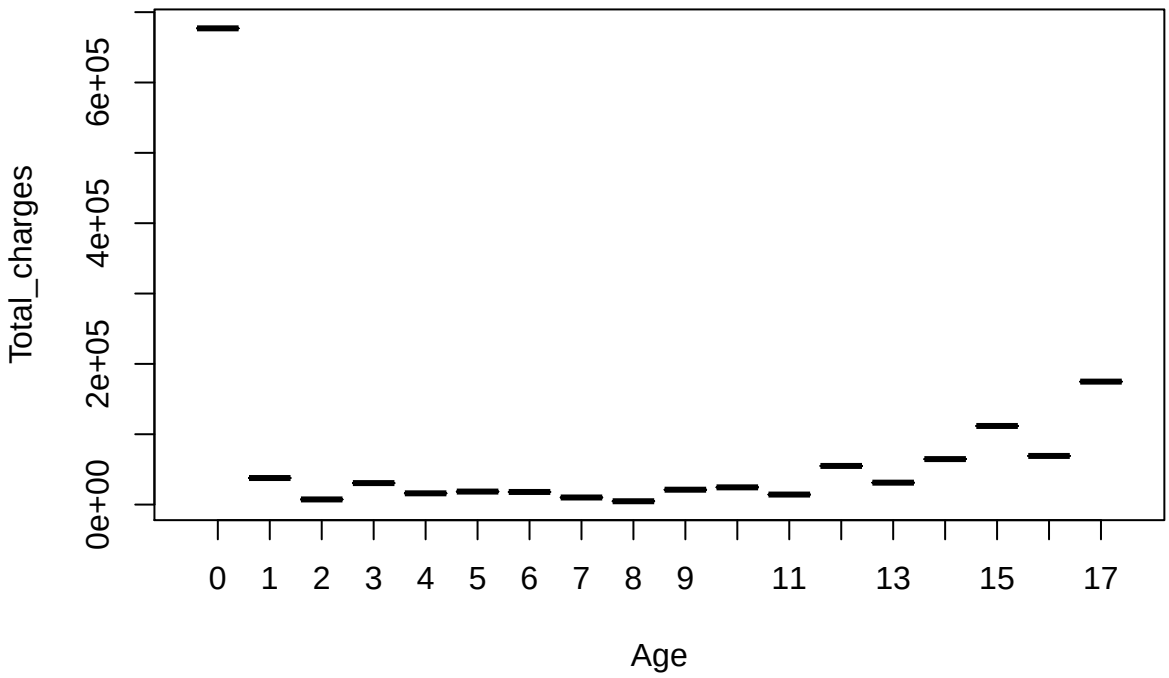
So, from the above data we can see that, average hospitilization cost of age category 9 is maximum.

2) If we take sum of hospitalization cost for each age category, we get an idea of the overall sum of hospitalization cost by each age group(Which just adds up the cost of all person's in age group)

```
Age_Max_report_sum <- Healthcare %>% group_by(Age=Healthcare$AGE) %>% summarise(Total_charges=sum(TOTCH
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Age_Max_report_sum <- Age_Max_report_sum %>% arrange(desc(Total_charges))
View(Age_Max_report_sum)
plot(Age_Max_report_sum)
```



From the above data, we can see that, the total hospitalization cost of infants(ie age category 0) is more. (Since the infants ie age category 0 has more hospitalization, even their sum of hospitalization cost will be more.) So, if we want to estimate the hospital cost for every person in the age group , then first option will give a good view.

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.

```
table(Healthcare$APRDRG,Healthcare$LOS)
```

```
##
##         0   1   2   3   4
##   21    0   0   1   0   0
##   23    0   0   1   0   0
##   49    0   0   0   0   1
##   50    0   0   1   0   0
##   51    0   0   0   1   0
##   53    0   0   4   3   3
##   54    0   1   0   0   0
##   57    0   2   0   0   0
##   58    0   1   0   0   0
##   92    0   1   0   0   0
##   97    0   0   0   1   0
##   114   0   0   0   1   0
##   115   0   0   1   0   1
```

```
##   137   0   0   0   0   1
##   138   0   2   2   0   0
##   139   1   2   1   1   0
##   141   0   0   1   0   0
##   143   0   0   1   0   0
##   204   0   0   0   1   0
##   206   0   0   0   1   0
##   225   0   0   1   0   1
##   249   0   4   2   0   0
##   254   1   0   0   0   0
##   308   0   1   0   0   0
##   313   0   1   0   0   0
##   317   0   0   0   0   1
##   344   1   0   0   0   1
##   347   2   1   0   0   0
##   420   0   1   1   0   0
##   421   0   0   0   0   1
##   422   0   2   0   1   0
##   560   0   0   2   0   0
##   561   0   0   0   0   1
##   566   0   0   1   0   0
##   580   0   1   0   0   0
##   581   2   1   0   0   0
##   602   0   0   0   0   1
##   614   0   0   0   0   3
##   626   0   0   0   3   3
##   633   0   0   1   2   1
##   634   0   0   1   0   1
##   636   0   0   0   0   3
##   639   0   0   0   1   3
##   640   0  11 162  69  24
##   710   0   0   0   0   1
##   720   0   0   0   0   1
##   723   0   2   0   0   0
##   740   0   0   0   0   1
##   750   0   0   0   1   0
##   751   0   5   4   2   3
##   753   1   9  13   4   9
##   754   5   9  12   3   8
##   755   1  11   1   0   0
##   756   0   2   0   0   0
##   758   0   6   5   2   7
##   760   0   0   1   0   1
##   776   0   0   1   0   0
##   811   1   1   0   0   0
##   812   0   1   2   0   0
##   863   0   0   0   0   1
##   911   0   0   0   0   1
##   930   0   0   0   1   1
##   952   0   1   0   0   0
```

From the table we to know the number of people's Length of stay for every category of APRDRG

```
los_report <- Healthcare %>% filter(Healthcare$LOS==max(Healthcare$LOS))
los_report <- los_report %>% group_by(aprdrg=los_report$APRDRG) %>% summarise(count=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
los_report %>% filter(count==max(los_report$count))
```

```
## # A tibble: 1 x 2
##   aprdrg count
##   <fct>  <int>
## 1 640       24
```

From the above data, we get to know the APRDRG category whose length of stay in the hospital was maximum. And we can see that, APRDRG category 640 had more number of people whose length of stay in the hospital is more. Now to find the expensive treatements and maximum expenditure cost,

```
Expensive <- Healthcare %>% filter(TOTCHG==max(Healthcare$TOTCHG))
Expensive[,5:6]
```

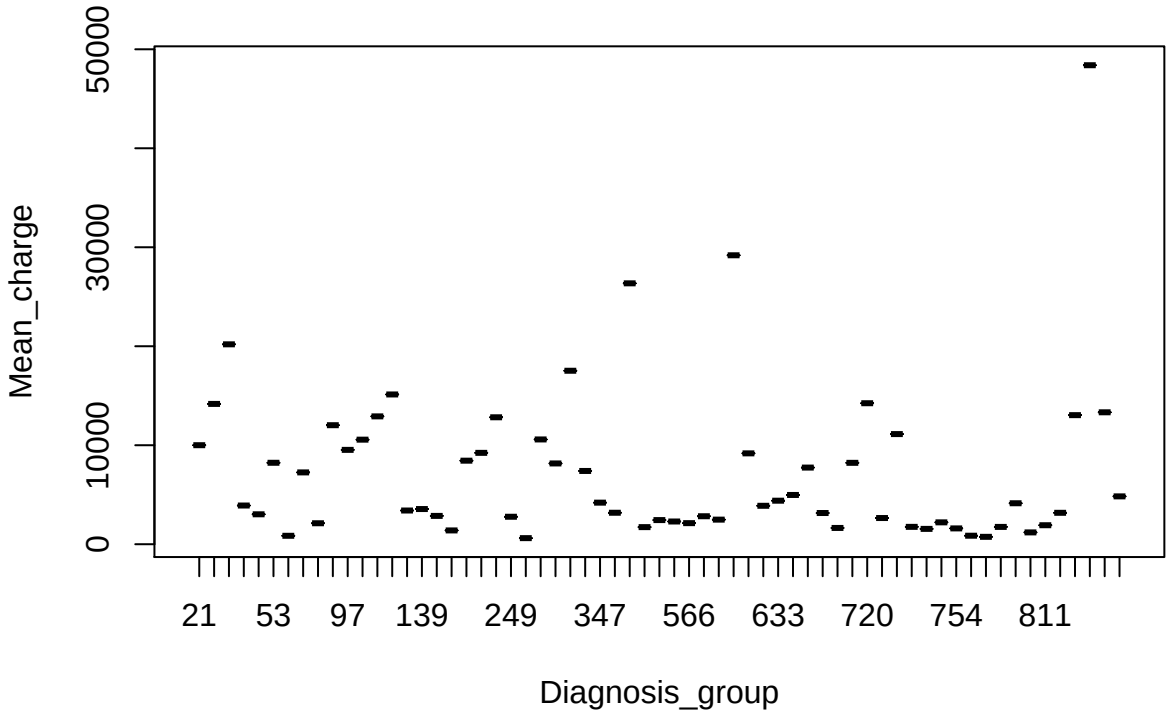
```
## # A tibble: 1 x 2
##   TOTCHG APRDRG
##    <dbl> <fct>
## 1  48388 911
```

So we can observe that, APRDRG category 911 had the expensive treatments.

```
Diag_max_report <- Healthcare %>% group_by(Diagnosis_group=Healthcare$APRDRG ) %>% summarise(Mean_charg
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Diag_max_report <- Diag_max_report %>% arrange(desc(Mean_charge))
View(Diag_max_report)
plot(Diag_max_report)
```



Now this gives the information about the average hospital cost for each APRDRG category. And we can see that apart from category 911 who had the expensive treatements, 602 is the category whose average hospitalization cost is high.

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.

```
lm(TOTCHG~RACE, Healthcare)->mod1
summary(mod1)
```

```
##
## Call:
## lm(formula = TOTCHG ~ RACE, data = Healthcare)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2323  -1557  -1235   -242  45615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2607.6     1066.3   2.445   0.0148 *
## RACE           165.1     1021.3   0.162   0.8717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3895 on 497 degrees of freedom
## Multiple R-squared:  5.256e-05,  Adjusted R-squared:  -0.001959
## F-statistic: 0.02612 on 1 and 497 DF,  p-value: 0.8717
```

From the above model, we can observe that p value is very high i.e., 0.8717. Here we take risk and we reject the null hypothesis that there exists a relationship betweeen RACE and Hospitalization cost. So race of the patient is not related to hospital 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.

```
Healthcare$AGE <- as.numeric(Healthcare$AGE)
lm(TOTCHG~AGE + FEMALE,Healthcare) -> mod2
summary(mod2)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE, data = Healthcare)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3403  -1444   -873   -156  44950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2633.41     271.01   9.717  < 2e-16 ***
## AGE            86.04      25.53   3.371 0.000808 ***
## FEMALE1      -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
```

From the above model, we can see that for Age, p value is very less. So Age contributes to the hospitalization cost. And we can even see that, Gender has some credits to hospital costs as its p value is also less. So, both Age and gender contributes to the hospitalization cost, but age contributes more.

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.

```
lm(LOS ~ AGE + FEMALE + RACE, Healthcare) -> mod3
summary(mod3)
```

```
##
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE, data = Healthcare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5367 -0.5367 -0.3277  0.6723  2.0889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.185439   0.283321   7.714 6.78e-14 ***
## AGE         -0.024504   0.006732  -3.640 0.000302 ***
## FEMALE1      0.208970   0.093476   2.236 0.025827 *
## RACE         0.166769   0.266341   0.626 0.531506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 495 degrees of freedom
## Multiple R-squared:  0.03006,    Adjusted R-squared:  0.02418
## F-statistic: 5.113 on 3 and 495 DF,  p-value: 0.001715
```

From the above model, we can see that both and gender are contributing to the length of stay. But Age's p value is even less when compared to the p value of gender. So both from Age,gender we will be able to predict the length of stay at the hospital. But Race is not related to length of stay at the hospital since p value is more.

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

```
Healthcare$APRDRG <- as.numeric(Healthcare$APRDRG)
lm(TOTCHG ~ . , Healthcare) -> mod4
summary(mod4)
```

```
##
## Call:
## lm(formula = TOTCHG ~ ., data = Healthcare)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7560  -1230   -183    209  43601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4723.81    1084.87   4.354 1.62e-05 ***
## AGE           155.72      21.53   7.232 1.83e-12 ***
## FEMALE1      -448.82     300.48  -1.494    0.136
## LOS          1654.83     140.59  11.771  < 2e-16 ***
## RACE         -793.71     834.32  -0.951    0.342
## APRDRG       -133.06      13.07 -10.177  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3168 on 493 degrees of freedom
## Multiple R-squared:  0.3439, Adjusted R-squared:  0.3373
## F-statistic: 51.69 on 5 and 493 DF,  p-value: < 2.2e-16
```

From the above model, we can see that, there are many 3 variables (Age, length of stay and APRDRG category) which are mainly contributing to the Hospital costs as they have very less p value. Even amongst those, length of stay and APRDRG category is more contributing to hospital costs. But Gender is not related to the hospital costs.