1.Analysis: max expenditure also by infant of 0 age =678118, 15=111747 17=174777

```
> summary(hosp_cost)
      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
                                                   NA's   :1
> head(hosp_cost$AGE)
[1] 17 17 17 17 17 17
> summary(hosp_cost$AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   5.086  13.000  17.000
> table(hosp_cost$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
> hist(hosp_cost$AGE)
> summary(as.factor(hosp_cost$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
> max(table(hosp_cost$AGE))
[1] 307
> max(summary(as.factor(hosp_cost$AGE)))
[1] 307
> which.max(table(hosp_cost$AGE))
0
1
> tapply(hosp_cost$TOTCHG,hosp_cost$AGE,sum)
     0      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15
678118  37744   7298  30550  15992  18507  17928  10087   4741  21147  24469  14250  54912  31135  64643 111747
    16     17
 69149 174777
> max(tapply(hosp_cost$TOTCHG,hosp_cost$AGE,sum))
[1] 678118
>
```

2.Analysis:640 group has maximum hospitalization and expenditure

```
> diag<-as.factor(hosp_cost$APRDRG)
> str(hosp_cost$APRDRG)
 int [1:500] 560 753 930 758 754 347 754 754 753 758 ...
> summary(diag)
 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
  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
420 421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760
  2   1   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
776 811 812 863 911 930 952
  1   2   3   1   1   2   1
> which.max(summary(diag))
640
 44
> diag1<-tapply(hosp_cost$TOTCHG,diag,sum)
> which.max(diag1)
640
 44
>  max(diag1)
[1] 437978
>
```

3.Analysis: p value is high so we can reject the null hypothesis and conclude there is no relation between race and hospitalization cost.

```
- -
> hosp_cost$RACE1<-as.factor(hosp_cost$RACE)
>  summary(rc)
   1    2    3    4    5    6 NA's
 484    6    1    3    3    2    1
>  hosp_cost<-na.omit(hosp_cost)
>  #annova model
>  result<-aov(hosp_cost$TOTCHG~hosp_cost$RACE)
>  summary(result)
                Df    Sum Sq  Mean Sq F value Pr(>F)
hosp_cost$RACE   1 2.488e+06  2488459   0.164  0.686
Residuals      497 7.540e+09 15170268
>
```

4.Analysis :P-value for age is significantly low that means it is important factor which impact cost. Gender has also less P-value and it has impact on cost same with intercept.

```
> fit<-lm(hosp_cost$TOTCHG~hosp_cost$AGE+hosp_cost$FEMALE)
>  summary(fit)

Call:
lm(formula = hosp_cost$TOTCHG ~ hosp_cost$AGE + hosp_cost$FEMALE)

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 ***
hosp_cost$AGE       86.04      25.53   3.371 0.000808 ***
hosp_cost$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

>
```

5.Analysis: Except with intercept, P value is high and LOS can't be predicted on the basis of age, gender and race.

```
> los<-lm(hosp_cost$LOS~hosp_cost$AGE+hosp_cost$FEMALE+hosp_cost$RAC
>  summary(los)

Call:
lm(formula = hosp_cost$LOS ~ hosp_cost$AGE + hosp_cost$FEMALE +
    hosp_cost$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 ***
hosp_cost$AGE    -0.03960    0.02231  -1.775   0.0766 .
hosp_cost$FEMALE  0.37011    0.31024   1.193   0.2334
hosp_cost$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

>
```

6.Analysis: we can see that age, LOS AND APRDRG affects the cost.

```
> hosp_cost<-na.omit(hosp_cost)
>  var<-lm(hosp_cost$TOTCHG~.,data=hosp_cost)
>  summary(var)

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

Residuals:
   Min     1Q Median     3Q    Max
 -6367   -691   -186    121  43412

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5143.8194   574.3867   8.955  < 2e-16 ***
AGE           133.2207    17.6662   7.541 2.29e-13 ***
FEMALE       -392.5778   249.2981  -1.575    0.116
LOS           742.9637    35.0464  21.199  < 2e-16 ***
RACE         -118.8584   371.8254  -0.320    0.749
APRDRG         -7.8175     0.6881 -11.361  < 2e-16 ***
RACE12        577.1012  1143.9938   0.504    0.614
RACE13        568.2353  2731.7186   0.208    0.835
RACE14       -142.8065  1879.9689  -0.076    0.939
RACE15      -1309.1439  2133.4023  -0.614    0.540
RACE16              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2622 on 489 degrees of freedom
Multiple R-squared:  0.5544,    Adjusted R-squared:  0.5462
F-statistic:  67.6 on 9 and 489 DF,  p-value: < 2.2e-16
```