Hospital Costs Analysis - US Agency for Healthcare - Wisconsin

Import Libraries

In [2]:
```
library(readxl)
library(readxl)
```

Import dataset

In [3]:
```
hospitalCost = read_excel("Downloads/Hospital Costs.xlsx", sheet = 1, col_name = TRUE)
```

In [4]:
```
head(hospitalCost)
```

AGE	FEMALE	LOS	RACE	TOTCHG	APRDRG
17	1	2	1	2660	560
17	0	2	1	1689	753
17	1	7	1	20060	930
17	1	1	1	736	758
17	1	1	1	1194	754
17	0	0	1	3305	347

In [5]:
```
colnames(hospitalCost)
```

1. 'AGE'
2. 'FEMALE'
3. 'LOS'
4. 'RACE'
5. 'TOTCHG'
6. 'APRDRG'

1. Recorded patient statistics

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.

In [6]:
```
summary(hospitalCost)
```

```
      AGE              FEMALE            LOS              RACE      
 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000  
 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000  
 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000  
 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078  
 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000  
 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000  
                                                   NA's   :1      
     TOTCHG          APRDRG     
 Min.   :  532   Min.   : 21.0  
 1st Qu.: 1216   1st Qu.:640.0  
 Median : 1536   Median :640.0  
 Mean   : 2774   Mean   :616.4  
 3rd Qu.: 2530   3rd Qu.:751.0  
 Max.   :48388   Max.   :952.0  
```

Number of hospical visits based on age

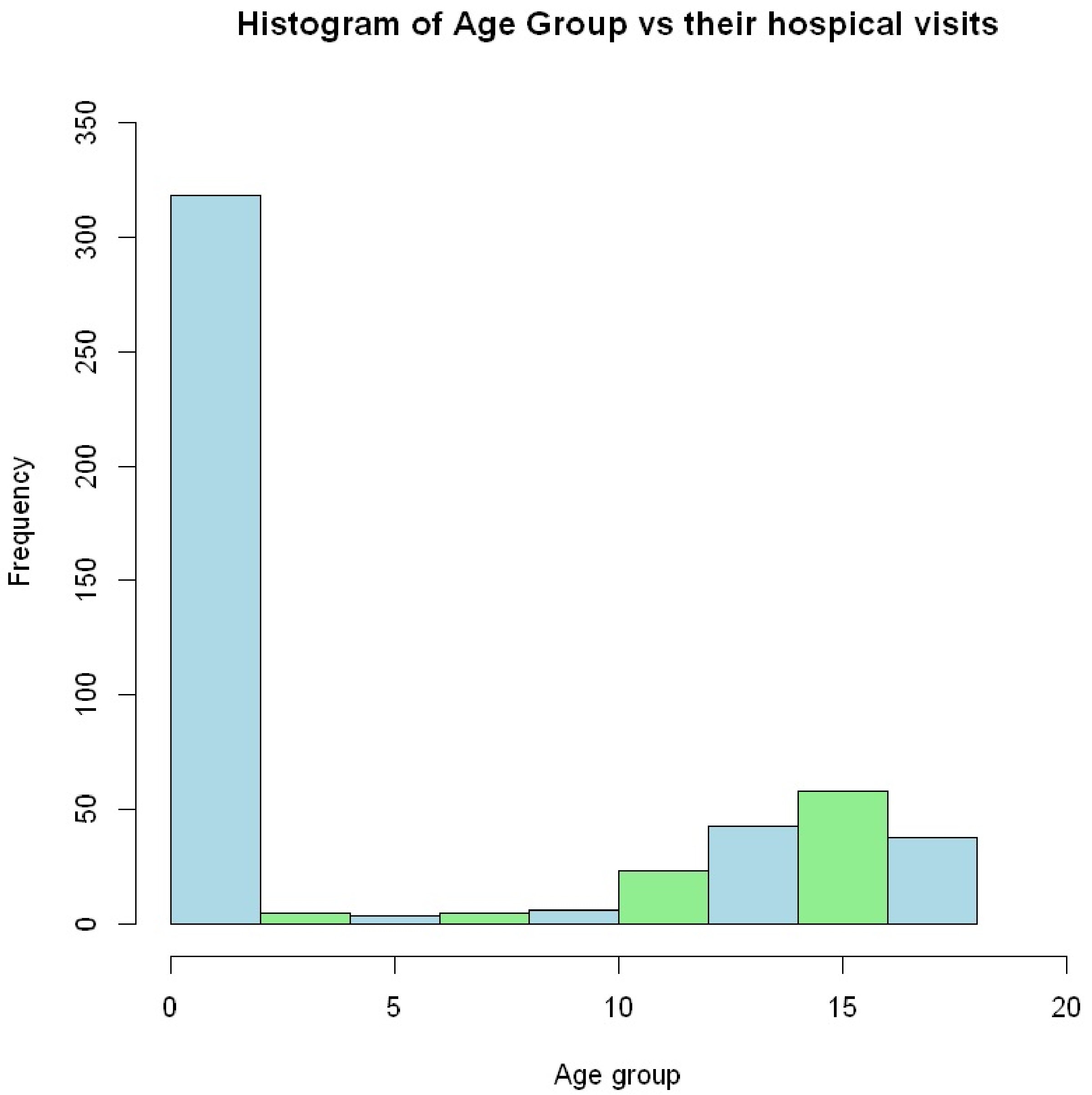
In [7]:
```
summary(as.factor(hospitalCost$AGE))
```

0	307
1	10
2	1
3	3
4	2
5	2
6	2
7	3
8	2
9	2
10	4
11	8
12	15
13	18
14	25
15	29
16	29
17	38

- Total number of patients from 0-1 age group is 307

In [8]:

```
hist(hospitalCost$AGE,
     main = "Histogram of Age Group vs their hospical visits",
     xlab = "Age group",
     border = "black",
     xlim = c(0,20),
     ylim = c(0, 350),
     col = c("light blue", "light green"))
```



Summarize expenditure based on age group

```
In [9]: expenseBasedOnAge = aggregate(TOTCHG ~ AGE, FUN = sum, data = hospitalCost)
        expenseBasedOnAge
```

AGE	TOTCHG
0	678118
1	37744
2	7298
3	30550
4	15992
5	18507
6	17928
7	10087
8	4741
9	21147
10	24469
11	14250
12	54912
13	31135
14	64643
15	111747
16	69149
17	174777

Maximum total expense

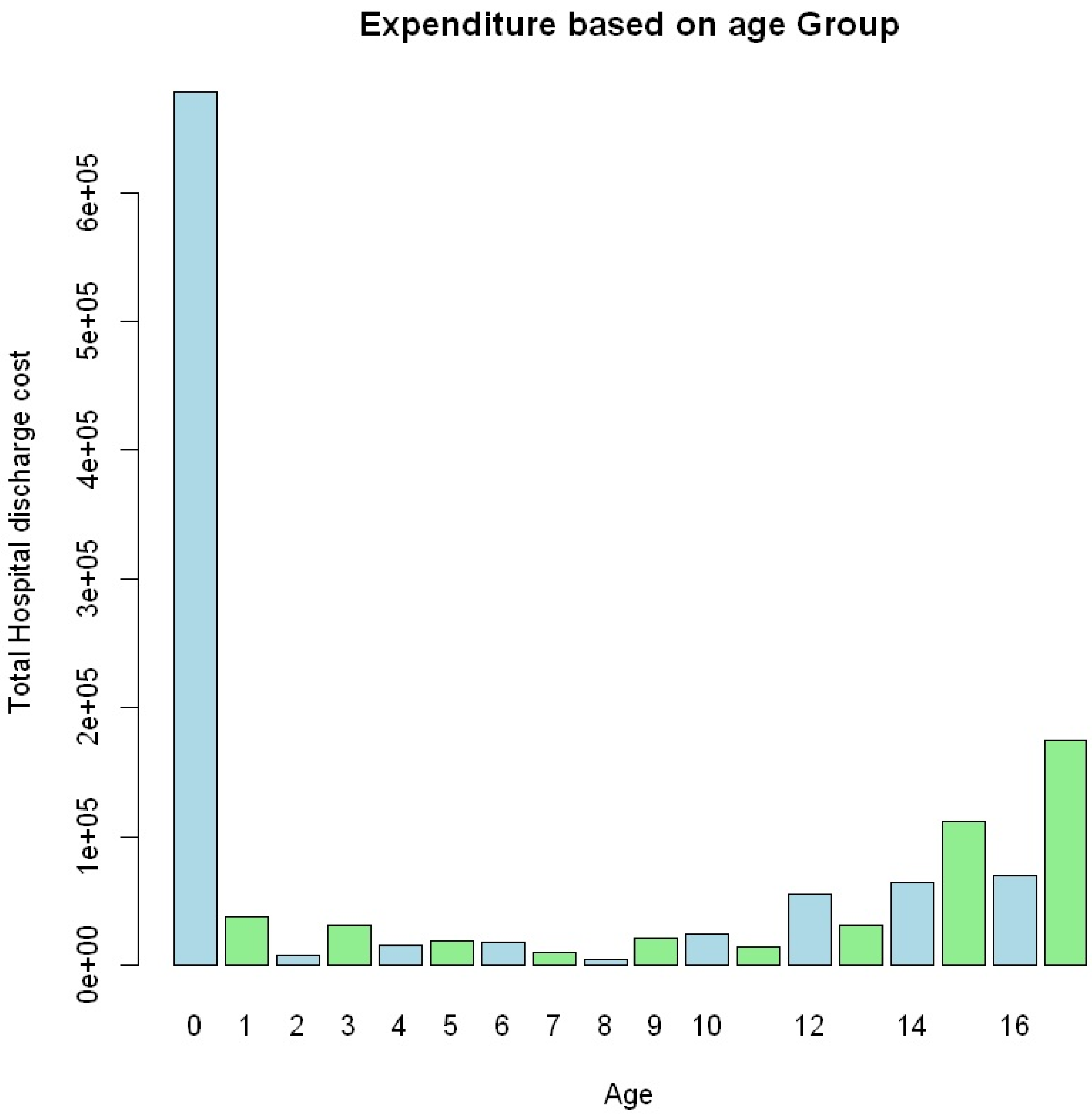
In [10]:

```
expenseBasedOnAge[which.max(expenseBasedOnAge$TOTCHG), ]
```

AGE	TOTCHG
0	678118

In [11]:

```
barplot(tapply(expenseBasedOnAge$TOTCHG, expenseBasedOnAge$AGE, FUN = sum),
        main = "Expenditure based on age Group",
        col = c("light blue", "light green"),
        xlab = "Age",
        ylab = "Total Hospital discharge cost")
```



2. Diagnosis-related group that has maximum hospitalization and expenditure

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.

In [12]:
```
summary(as.factor(hospitalCost$APRDRG))
```

21	1
23	1
49	1
50	1
51	1
53	10
54	1
57	2
58	1
92	1
97	1
114	1
115	2
137	1
138	4
139	5
141	1
143	1
204	1
206	1
225	2
249	6
254	1
308	1
313	1
317	1
344	2
347	3
420	2
421	1
422	3
560	2
561	1
566	1
580	1
581	3
602	1
614	3
626	6
633	4
634	2
636	3
639	4
640	267
710	1
720	1
723	2
740	1
750	1
751	14
753	36
754	37
755	13
756	2
758	20
760	2
776	1
811	2
812	3
863	1
911	1
930	2
952	1

In [13]:

```
diagnosisCost = aggregate(TOTCHG ~ APRDRG, FUN = sum, data = hospitalCost)
```

```
diagnosisCost
```

APRDRG	TOTCHG
21	10002
23	14174
49	20195
50	3908
51	3023
53	82271
54	851
57	14509
58	2117
92	12024
97	9530
114	10562
115	25832
137	15129
138	13622
139	17766
141	2860
143	1393
204	8439
206	9230
225	25649
249	16642
254	615
308	10585
313	8159
317	17524
344	14802
347	12597
420	6357
421	26356
...	...
566	2129
580	2825
581	7453
602	29188
614	27531
626	23289
633	17591
634	9952
636	23224
639	12612
640	437978
710	8223
720	14243
723	5289
740	11125
750	1753
751	21666
753	79542

754	59150
755	11168
756	1494
758	34953
760	8273
776	1193
811	3838
812	9524
863	13040
911	48388
930	26654
952	4833

Maximum Diagnostic Cost

In [14]:
```
diagnosisCost[which.max(diagnosisCost$TOTCHG),]
```

	APRDRG	TOTCHG
44	640	437978

3. Race vs Hospitalization Costs

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.

In [15]:
```
summary(as.factor(hospitalCost$RACE))
```

1	484
2	6
3	1
4	3
5	3
6	2
NA's	1

- There is one null value, we need to remove that record

In [16]:
```
hospitalCost = na.omit(hospitalCost)
```

In [17]:
```
summary(as.factor(hospitalCost$RACE))
```

1	484
2	6
3	1
4	3
5	3
6	2

- As evident from the above observation, 484 out of 499 patients belong to group 1, indicating a significant imbalance in the distribution of observations across categories.
- This skewness in the data may impact the results of linear regression or ANOVA analysis.

In [18]:
```
raceInfluenceModel = lm(TOTCHG ~ RACE, data = hospitalCost)
```

In [19]:
```
summary(raceInfluenceModel)
```

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

Residuals:
   Min     1Q Median     3Q    Max 
 -2256  -1560  -1227   -258  45600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2925.7      405.0   7.224 1.92e-12 ***
RACE          -137.3      339.1  -0.405    0.686    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3895 on 497 degrees of freedom
Multiple R-squared:  0.0003299,	Adjusted R-squared:  -0.001681 
F-statistic: 0.164 on 1 and 497 DF,  p-value: 0.6856
```

- pValue is 0.686 it is much higher that 0.05
- So, we can infer that race doesn't affect the hospitalization costs

Analysis using ANOVA

We can also use the ANOVA Statistical test for estimating how dependent variable (in this case RACE), affect the independent variables (in this case TOTCHG)

In [20]:

```
raceInfluenceAOV = aov(TOTCHG ~ RACE, data = hospitalCost)
raceInfluenceAOV
```

```
Call:
   aov(formula = TOTCHG ~ RACE, data = hospitalCost)

Terms:
                      RACE  Residuals
Sum of Squares     2488459 7539623326
Deg. of Freedom          1        497

Residual standard error: 3894.903
Estimated effects may be unbalanced
```

In [21]:

```
summary(raceInfluenceAOV)
```

```
             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               
```

- The Residuals variance is very high. This implies that there is very little influence from RACE on TOTCHG
- The Pr(>F), the pValue for 0.69 is higher than 0.05 which confirms that RACE doesn't affect hospitalization cost

4. Age and Gender vs Hospitalization costs

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.

In [22]:

```
summary(as.factor(hospitalCost$FEMALE))
```

0	244
1	255

In [23]:

```
ageGenderInfluenceModel = lm(TOTCHG ~ FEMALE + AGE, data = hospitalCost)
ageGenderInfluenceModel
```

```
Call:
lm(formula = TOTCHG ~ FEMALE + AGE, data = hospitalCost)

Coefficients:
(Intercept)       FEMALE          AGE  
    2719.45      -744.21        86.04  
```

In [24]:

```
summary(ageGenderInfluenceModel)
```

```
Call:
lm(formula = TOTCHG ~ FEMALE + AGE, data = hospitalCost)

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 ***
FEMALE       -744.21     354.67  -2.098 0.036382 *  
AGE            86.04      25.53   3.371 0.000808 ***
---
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
```

- pValue of AGE is much less that 0.05, means AGE has the most statitical significance
- Similarty, GENDER also have pValue less that 0.05
- Hence, we can conclude that the model is statitical significance

5. Can lenght of stay be predicted from age, gender, and race

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.

In [25]:

```
hospitalCost2 = hospitalCost
hospitalCost2$RACE = as.factor(hospitalCost$RACE)
```

In [26]:

```
ageGenderRaceInfluenceModel = lm (LOS ~ AGE + RACE + FEMALE, data = hospitalCost2)
ageGenderRaceInfluenceModel
```

```
Call:
lm(formula = LOS ~ AGE + RACE + FEMALE, data = hospitalCost2)

Coefficients:
(Intercept)          AGE        RACE2        RACE3        RACE4        RACE5  
    2.85687     -0.03938     -0.37501      0.78922      0.59493     -0.85687  
      RACE6       FEMALE  
   -0.71879      0.35391  
```

In [27]:

```
summary(ageGenderRaceInfluenceModel)
```

```
Call:
lm(formula = LOS ~ AGE + RACE + FEMALE, data = hospitalCost2)

Residuals:
   Min     1Q Median     3Q    Max 
-3.211 -1.211 -0.857  0.143 37.789 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.85687    0.23160  12.335   <2e-16 ***
AGE         -0.03938    0.02258  -1.744   0.0818 .  
RACE2       -0.37501    1.39568  -0.269   0.7883    
RACE3        0.78922    3.38581   0.233   0.8158    
RACE4        0.59493    1.95716   0.304   0.7613    
RACE5       -0.85687    1.96273  -0.437   0.6626    
RACE6       -0.71879    2.39295  -0.300   0.7640    
FEMALE       0.35391    0.31292   1.131   0.2586    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.376 on 491 degrees of freedom
Multiple R-squared:  0.008699,	Adjusted R-squared:  -0.005433 
F-statistic: 0.6156 on 7 and 491 DF,  p-value: 0.7432
```

- The pValue is greater than 0.05 for age, gender and race, indicating that there is no linear relationship between there variables and length of stay.
- Hence, age, gender and race can not be user to predict the length of stay of inpatients.

6. Complete Analysis

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

In [28]:

```
hospitalCostModel = lm(TOTCHG ~ AGE + FEMALE + LOS + RACE + APRDRG,
             data = hospitalCost)
```

In [29]:

```
summary(hospitalCostModel)
```

```
Call:
lm(formula = TOTCHG ~ AGE + FEMALE + LOS + RACE + APRDRG, data = hospitalCost)

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
```

- As AGE, LOS, and APRDRG have pValue less that 0.05, so they are the onces with statistical significance
- As pValue for variables FEMALE and RACE is greater than 0.05, so building another model after removing these variables.

In [30]:

```
hospitalCostModel2 = lm(TOTCHG ~ AGE + LOS + APRDRG,
                        data = hospitalCost)
```

In [31]:

```
summary(hospitalCostModel2)
```

```
Call:
lm(formula = TOTCHG ~ AGE + LOS + APRDRG, data = hospitalCost)

Residuals:
   Min     1Q Median     3Q    Max 
 -6603   -719   -169    124  43350 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4960.1705   433.6579   11.44  < 2e-16 ***
AGE          128.5519    17.0946    7.52 2.59e-13 ***
LOS          740.8057    34.9161   21.22  < 2e-16 ***
APRDRG        -8.0055     0.6643  -12.05  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2617 on 495 degrees of freedom
Multiple R-squared:  0.5506,	Adjusted R-squared:  0.5479 
F-statistic: 202.2 on 3 and 495 DF,  p-value: < 2.2e-16
```

In [32]:

```
hospitalCostModel3 = lm(TOTCHG ~ AGE + LOS,
                        data = hospitalCost)
```

In [33]:

```
summary(hospitalCostModel3)
```

```
Call:
lm(formula = TOTCHG ~ AGE + LOS, data = hospitalCost)

Residuals:
   Min     1Q Median     3Q    Max 
 -4783  -1103   -458   -133  41382 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   200.66     203.48   0.986    0.325    
AGE            97.96      19.21   5.101 4.83e-07 ***
LOS           734.27      39.66  18.512  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2973 on 496 degrees of freedom
Multiple R-squared:  0.4188,	Adjusted R-squared:  0.4164 
F-statistic: 178.7 on 2 and 496 DF,  p-value: < 2.2e-16
```

- Removing RACE and FEMALE doesn't change the R-square values. There variables doesn't impact the cost.
- Removal of APRDRG in the model hospitalCostModel3 incresess the residual standard error, Hence model hospitalCostModel2

seems to be BETTER.

Analysis Conclusion

- As evident from the above multiple models, health care cost is dependent on Age, Length of stay and the diagnosis type.

1. Healthcare cost is the most for patients in the 0-1 yrs age group category

- Maximum expenditure for 0-1 yr is 678118

2. Length of Stay increases the hospital cost
3. All Patient Refined Diagnosis Related Groups also affects healthcare costs

- 640 diagnosis related group had a max cost of 437978

4. Race or gender doesn’t have that much impact on hospital cost