Data Science With R-Project Healthcare - Simplilearn

Dheeraj Bharat Sethi

8/1/2020

Project Statement

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.

First, lets import the required csv file, and get the feel of data, columns etc We will be using various R basic functions to do that

```
hospital_data<-read.csv(file.choose(),header = T)
names(hospital_data) #Get the columns
```

```
## [1] "AGE"    "FEMALE" "LOS"    "RACE"   "TOTCHG" "APRDRG"
```

```
head(hospital_data) #First 6 rows
```

	AGE <int>	FEMALE <int>	LOS <int>	RACE <int>	TOTCHG <int>	APRDRG <int>
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

6 rows

```
str(hospital_data) #Structure, datatype of each column
```

```
## 'data.frame':    500 obs. of  6 variables:
##  $ AGE   : int  17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE: int  1 0 1 1 1 0 1 1 1 1 ...
##  $ LOS   : int  2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: int  2660 1689 20060 736 1194 3305 2205 1167 532 1363 ...
##  $ APRDRG: int  560 753 930 758 754 347 754 754 753 758 ...
```

Task 1

To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.

So, here we need to find out the number of records grouped by AGE As AGE is int variable, its better to convert that to a factor, and then use table, summary functions to get the count of rows based on AGE values

```
#hospital_data$AGE<-as.factor(hospital_data$AGE)
levels(as.factor(hospital_data$AGE))
```

```
##  [1] "0"  "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
## [16] "15" "16" "17"
```

```
table(as.factor(hospital_data$AGE)) #here AGE 0 has max number, getting an idea
```

```
## 
##   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
```

```
summary(hospital_data) #Summary of each variable in the dataset
```

```
##       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  
## 
```

```
#Now the max cost analysis, we need the aggregated values of TOTCHRG based on AGE, we can use aggregate function
cost_aggregate<-aggregate(TOTCHG~AGE,data = hospital_data,FUN = sum)
cost_aggregate
```

AGE <int>	TOTCHG <int>
0	678118
1	37744
2	7298
3	30550
4	15992
5	18507
6	17928
7	10087
8	4741
9	21147

1-10 of 18 rows   Previous 1 2 Next

```
max(cost_aggregate) #678118
```

```
## [1] 678118
```

```
cost_aggregate[which.max(cost_aggregate$TOTCHG),] #Tells AGE=0 as maximum entry in aggregate
```

	AGE <int>	TOTCHG <int>
1	0	678118

1 row

Conclusion 1:

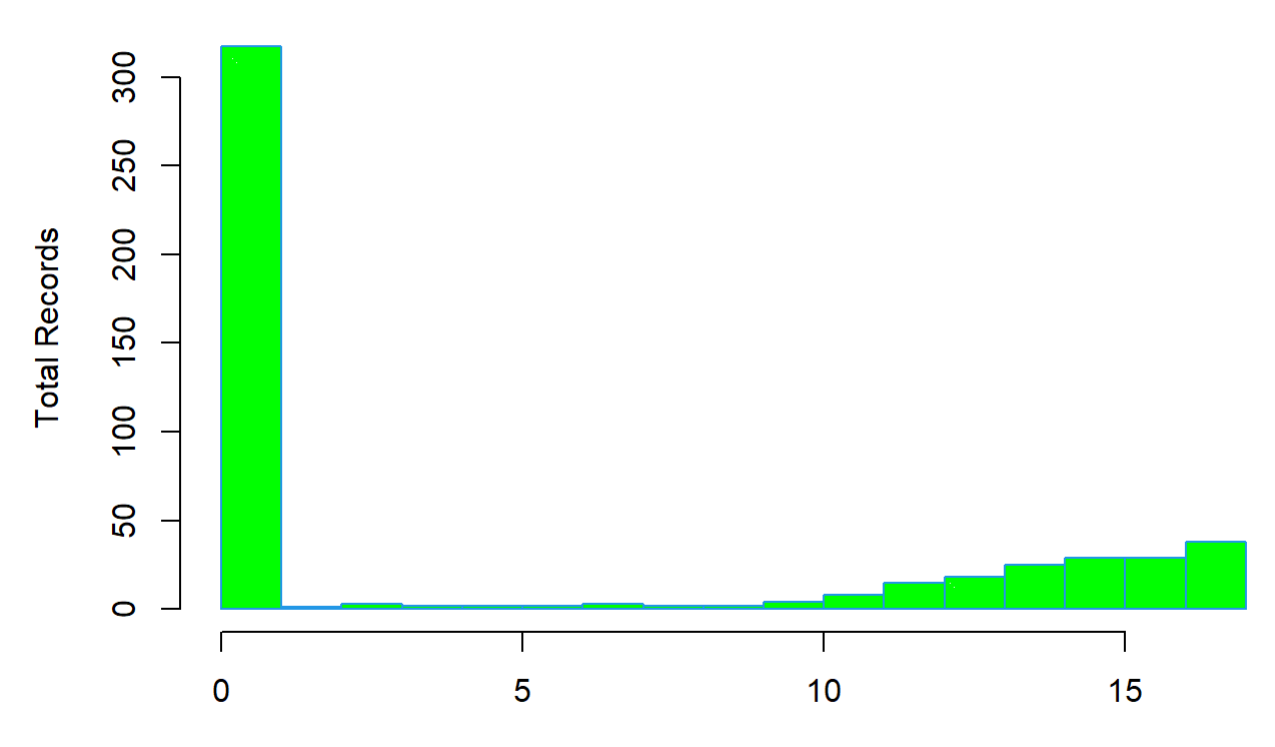
Above output clearly shows the AGE=0 has 307 records, meaning, patients with AGE=0 are the most frequent visitors to the hospital_data.

Conclusion 2:

Aggregated expenditure of AGE=0 is the maximum, meaning, AGE=0 has the maximum costs.

Its time to PLOT these to visualize the observation We can draw a Histogram, or a BarPlot

```
hist(hospital_data$AGE,breaks = nlevels(as.factor(hospital_data$AGE)),xlab = "Age",ylab = "Total Records",col = "green",freq = T,density = 100,border = 4)
```




```
barplot(table(hospital_data$LOS,hospital_data$AGE),xlab = "Age",ylab = "Total rows, Days Stayed stacked",col = "green",density = 100,border = 4)
```

Task 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.

This is similar problem to Task 1, here we need to aggregate the Cost based on the diagnosis codes

```
summary(as.factor(hospital_data$APRDRG)) #we can see code 640 has 267 records out of 500
```

```
##  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   4 267   1   1   2   1   1  14  36  37  13   2  20   2   1   2   3   1 
## 911 930 952 
##   1   2   1
```

```
cost_diag_aggregate<-aggregate(TOTCHG~APRDRG,data = hospital_data,FUN = sum)
cost_diag_aggregate
```

APRDRG <int>	TOTCHG <int>
21	10002
23	14174
49	20195
50	3908
51	3023
53	82271
54	851
57	14509
58	2117
92	12024

1-10 of 63 rows   Previous 1 2 3 4 5 6 7 Next

```
max(cost_diag_aggregate) #678118
```

```
## [1] 437978
```

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

	APRDRG <int>	TOTCHG <int>
44	640	437978

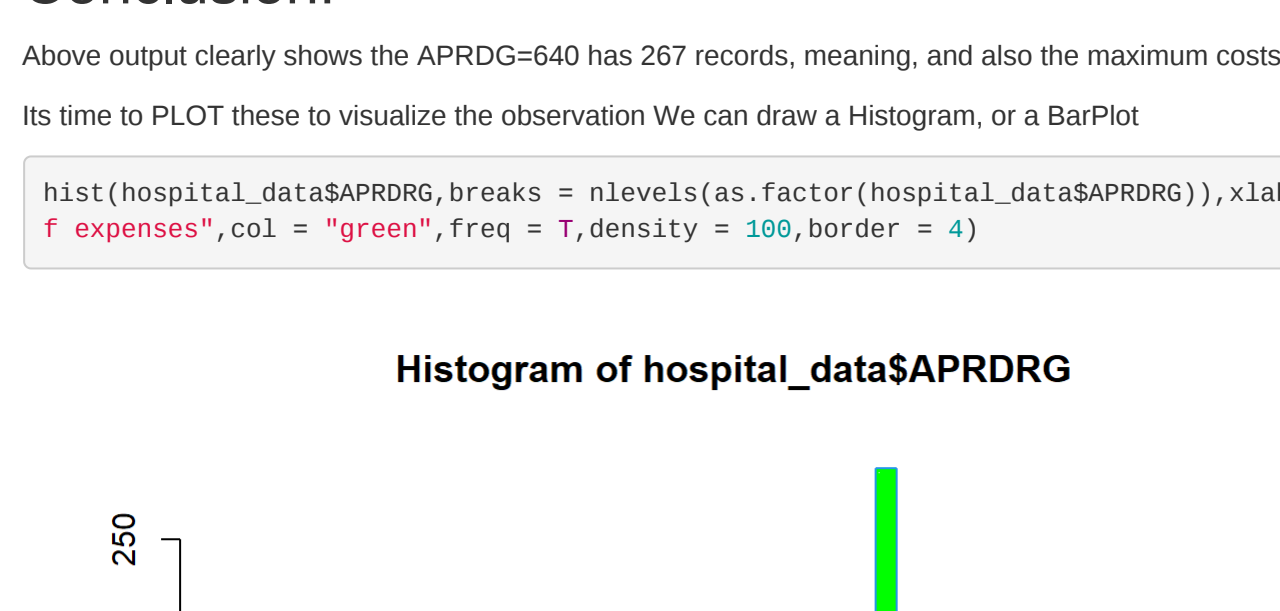
1 row

Conclusion:

Above output clearly shows the APRDG=640 has 267 records, meaning, and also the maximum costs value.

Its time to PLOT these to visualize the observation We can draw a Histogram, or a BarPlot

```
hist(hospital_data$APRDRG,breaks = nlevels(as.factor(hospital_data$APRDRG)),xlab = "Diagnosis Code",ylab = "Sum of expenses",col = "green",freq = T,density = 100,border = 4)
```




```
#Histo chart shows the same results,
#Barplot won't be practical here, as it would take a lot of space to draw x bars. As can be seen with unique number of values of diagnosis codes
nlevels(as.factor(hospital_data$APRDRG))
```

```
## [1] 63
```

```
#or
unique(hospital_data$APRDRG)
```

```
##  [1] 560 753 930 758 754 347 751 812 566 249 422  50 139 141 420  97 811 755 720
## [20]  53 760 710 776 115 602 138 137 640 639 143 254 581 633 626 636  23  57 421
## [39] 580 750  49  51 313 614 634 952  21  92 756 317 344 114 206 723 911  54 225
## [58]  58 740 308 204 561 863
```

```
#We have 63 unique codes
```

Task 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.

Here, we have RACE - Categorical variable

Cost - Continuous variable

To find the relation between such combination, we can use ANOVA test

And we are doing the Hypothesis Testing

Null Hypothesis H0: RACE has No Effect on Costs

Alternate Hypothesis Ha : RACE has Effect on Costs

To REJECT the H0, we need enough evidences, means, a striking difference between the both.

In ANOVA, we have F-statistic

Step 1

Data Wrangling/Cleaning

We need to clean up the entries with values NA in any column ## Step 2

Apply ANOVA test R function, on TOTCHG as Dependent, RACE as independent variable

```
anyNA(hospital_data) #this tells TRUE
```

```
## [1] TRUE
```

```
anyNA(hospital_data$RACE) #RACE has some NA entries
```

```
## [1] TRUE
```

```
summary(hospital_data$RACE) # 1 NA value, we can omit this, as we have no way to predict this as we do for numerical variables
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   1.000   1.000   1.078   1.000   6.000       1
```

```
hospital_data<-na.omit(hospital_data)
anyNA(hospital_data) #Now we have no NA values
```

```
## [1] FALSE
```

```
#ready to test ANOVA now
model_race_vost_aov<-aov(hospital_data$TOTCHG~hospital_data$RACE)
summary(model_race_vost_aov)
```

```
##                     Df    Sum Sq  Mean Sq F value Pr(>F)
## hospital_data$RACE   1 2.488e+06  2488459   0.164  0.686
## Residuals          497 7.540e+09 15170268
```

Conclusion

F-statistic is 0.164, and P-value (probability) is 0.686 > 0.05 (significance level), hence we CANNOT REJECT the NULL Hypothesis.

Means RACE HAS NO IMPACT ON COSTS

Task 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.

Hospital wants to understand, whether The Cost Increases/decreases with AGE or GENDER, so we need to find the linear relation among these.

Using Linear Regression model, with TOTCHG as dependent (predicted) variable, GENDER, AGE as independent (predictor) variables

```
hospital_data$FEMALE<-as.factor(hospital_data$FEMALE)
summary(hospital_data$FEMALE) #almost similar number of MALES and FEMALES
```

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

```
fit_1<-lm(formula =  TOTCHG~AGE+FEMALE,data = hospital_data)
fit_1
```

```
## 
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE, data = hospital_data)
## 
## Coefficients:
## (Intercept)          AGE      FEMALE1  
##     2719.45        86.04      -744.21
```

```
summary(fit_1)
```

```
## 
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE, data = hospital_data)
## 
## 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 ***
## 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
```

Conclusion

We see, p-values for AGE is much lower than significance level, but p-value for Gender is much closer to 0.05.

That means, AGE has bigger impact on costs, although GENDER also is significant for the model.

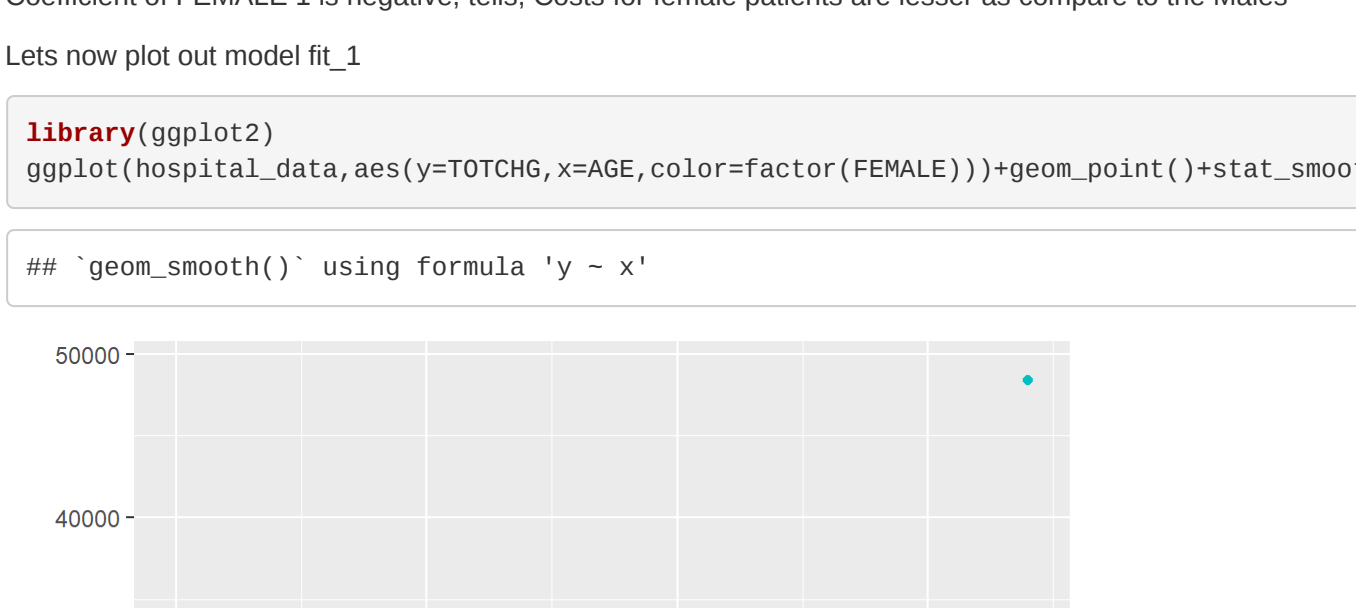
Coefficient of AGE is higher, means cost is positively increasing with age.

Coefficient of FEMALE 1 is negative, tells, Costs for female patients are lesser as compare to the Males

Lets now plot out model fit_1

```
library(ggplot2)
ggplot(hospital_data,aes(y=TOTCHG,x=AGE,color=factor(FEMALE)))+geom_point()+stat_smooth(method="lm",se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#The plot clearly defines our regression model, female have lesser costs than males,
#and More the age, more the costs
```

Task 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.

This is similar problem as 3, here the predictor variables are AGE,GENDER,RACE Predicted variable is Length of stay.

Lets use linear regression to find the relation

```
hospital_data$RACE<-as.factor(hospital_data$RACE)
hospital_data<-na.omit(hospital_data)
fit_los<-lm(data = hospital_data,formula = LOS~AGE+FEMALE+RACE)
fit_los
```

```
## 
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospital_data)
## 
## Coefficients:
## (Intercept)          AGE      FEMALE1        RACE2        RACE3        RACE4  
##     2.85687     -0.03938      0.35391     -0.37501      0.78922      0.59493  
##       RACE5        RACE6  
##    -0.85687     -0.71879
```

```
summary(fit_los)
```

```
## 
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospital_data)
## 
## 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 .  
## FEMALE1      0.35391    0.31292   1.131   0.2586    
## 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    
## ---
## 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
```

Conclusion

None of the predictor variables show significance to the model,

All the p-values are high, thus, Accepting the Null Hypothesis

Concluding that Length of stay cannot be predicted by AGE, RACE, or GENDER

Task 6

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

Predict Costs, with all other independent variables.

```
fit_costs_all<-lm(data = hospital_data,formula = TOTCHG~AGE+FEMALE+RACE+APRDRG+LOS)
fit_costs_all
```

```
## 
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE + RACE + APRDRG + LOS, data = hospital_data)
## 
## Coefficients:
## (Intercept)          AGE      FEMALE1        RACE2        RACE3        RACE4  
##    5024.961      133.221     -392.578      458.243      330.518     -499.382  
##       RACE5        RACE6       APRDRG          LOS  
##   -1784.578     -594.292       -7.818      742.964
```

```
summary(fit_costs_all)
```

```
## 
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE + RACE + APRDRG + LOS, data = hospital_data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -6367   -691   -186    121  43412 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5024.9610   440.1366  11.417  < 2e-16 ***
## AGE           133.2207    17.6662   7.541 2.29e-13 ***
## FEMALE1      -392.5778   249.2981  -1.575    0.116    
## RACE2         458.2427  1085.2320   0.422    0.673    
## RACE3         330.5184  2629.5121   0.126    0.900    
## RACE4        -499.3818  1520.9293  -0.328    0.743    
## RACE5       -1784.5776  1532.0048  -1.165    0.245    
## RACE6        -594.2921  1859.1271  -0.320    0.749    
## APRDRG         -7.8175     0.6881 -11.361  < 2e-16 ***
## LOS           742.9637    35.0464  21.199  < 2e-16 ***
## ---
## 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
```

Conclusion

AGE, LOS, APRDRG affect the costs for hospitals, none other does.

LOS has positive relation with Costs.

Here, LOS is the continuous variable, we can compare this with costs as=>

With each increment in length of days stayed, the Cost/Charges increases by 742.97 units.