PROJECT REPORT

Healthcare Cost Analysis and Utilization in Wisconsin Hospitals

Name: V. Aishwarya Laxmi

Course: PGP Data Science with R

CONTENT

BUSINESS SCENARIO

DESCRIPTION:

Background and Objective:

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.

Domain: Healthcare

Dataset Description: Here is a detailed description of the given dataset:

ATTRIBUTE	DESCRIPTION
AGE	Age of the patient discharged
FEMALE	A binary variable that indicates if the patient is female
LOS	Length of stay in days
RACE	Race of the patient (specified numerically)
TOTCHG	Hospital discharge costs
APRDRG	All Patient Refined Diagnosis Related Groups

EXPECTED OUTCOMES

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.

CODE WRITEUP:

Tool used: R-Studio

First, we need to import the data/records into the R- environment. For this we need to follow 2 important steps:

1. Check the working directory. If it is not set to desired folder/directory, set the directory.
2. Importing the dataset required and storing it in a variable.

Code:

```
# Import Dataset

```{r}

#setting working directory

setwd("D:/Tech_Code/Simplilearn_materials/Simplilearn_R/Simplilearn_R_Projects/R_Simplilearn_project7/Simplilearn_R_projects")

getwd()

```
```

[1]"D:/Tech_Code/Simplilearn_materials/Simplilearn_R/Simplilearn_R_Projects/R_Simplilearn_project7/Simplilearn_R_projects"

To import an excel sheet into R-studio, we need use an external package, since the function is not inbuilt in the tool.

Code:

```
```{r}

#install.packages("readxl")

library(readxl)

Hospital_costs <- read_excel("hospitalCosts.xlsx")

head(Hospital_costs) //To check the first 5 columns of the imported dataset.

```
```

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

6 rows

STEP-1: To record the patient statistics, the agency wants to find the age category of people who frequents the hospital and has the maximum expenditure.

Required attributes:

AGE: Age of the patient discharged

TOTCHG: Hospital discharge costs.

Before jumping into the analysis, we need to check the datatype of attribute, to make sure it is characterized into desired type. If not, we can change the datatype into required form.

Code:

```{r}

summary(Hospital_costs)

```

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

Note-The following question has two conclusions

1. Here the age groups are discrete values. But in-order to categorize them into several groups i.e. 0-1, 1-2, …, 16-17, we need the age groups to be categorical variable.

Code:

Get the number hospital visits based on age

```{r}

summary(as.factor(Hospital_costs$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

- We can use data visualization to get an overview of the all the categories, in this case we will use a histogram for frequency analysis

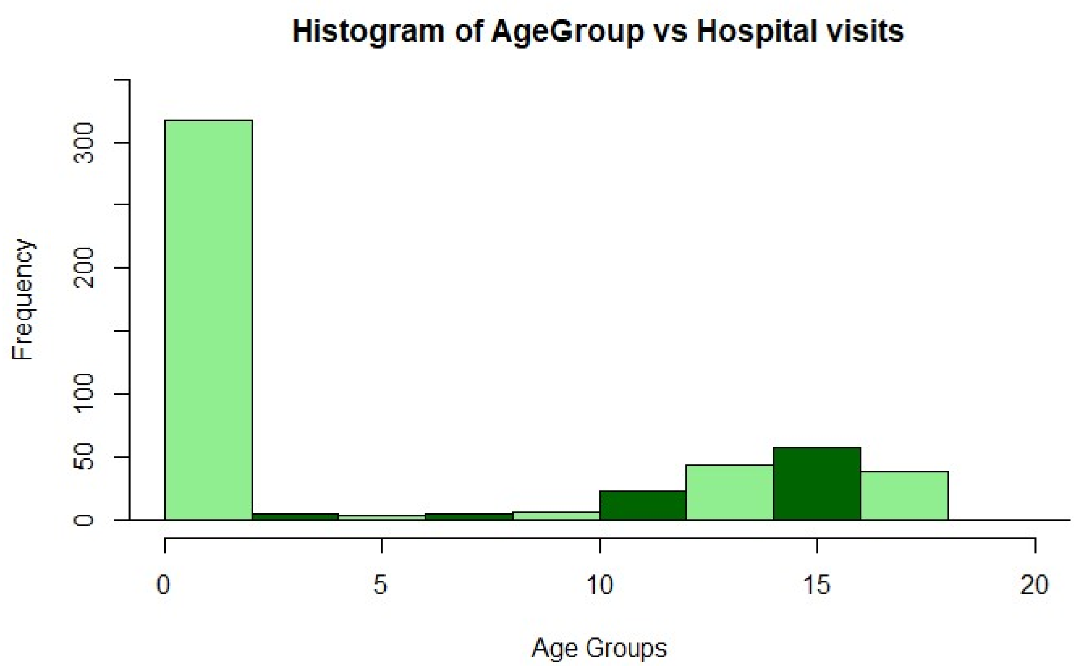
Code:

~~~
```{r}

hist(Hospital_costs$AGE, main = "Histogram of AgeGroup vs Hospital visits",

   xlab = "Age Groups",border = "black", col = c("light green","dark green"), xlim=c(0,20), ylim=c(0,350))

```
~~~



Conclusion 1: From the above results we conclude that infant category has the max hospital visits (above 300). The summary of Age gives us the exact numerical output showing that Age 0 patients have the max visits followed by Ages 15-17.

2.Use Aggregate function to add the expenditure from each age and then max function to find highest costs.

Code:

```
# Summarized expenditure based on age group
```

```{r}
Expense_Age <- aggregate(TOTCHG ~ AGE, FUN = sum, data = Hospital_costs)
```

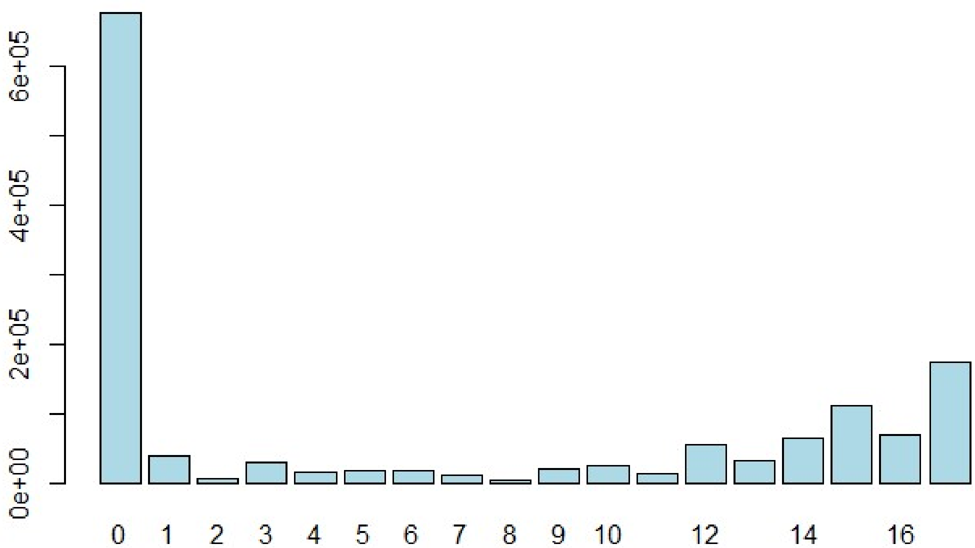
Get the maximum expense and its corresponding age group

```{r}
which.max(tapply(Expense_Age$TOTCHG, Expense_Age$TOTCHG, FUN=sum))
barplot(tapply(Expense_Age$TOTCHG, Expense_Age$AGE, FUN=sum),
    col = "light blue")
```

R Console

678118

18



Conclusion 2: Thus, we can conclude that the infants also have the maximum hospital costs followed by Age groups 15 to 17, additionally we can say confidently that number of hospital visits are proportional to hospital costs

STEP-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.

Required attributes:

APRDRG: All Patient Refined Diagnosis Related Groups

- is a classification system that classifies patients according to their reason of admission, severity of illness and risk of mortality.

TOTCHG: Hospital discharge costs

Make sure that category column("APRDRG") is numerical and then generate a summary along with the which.max to generate the max index of the category data frame, this will be done after using aggregate function to sum TOTCHG.

Code:

```{r}
summary(as.factor(Hospital_costs$APRDRG))
```

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

```
# Get the diagnosis-related group and its hospitalization expenditure
```

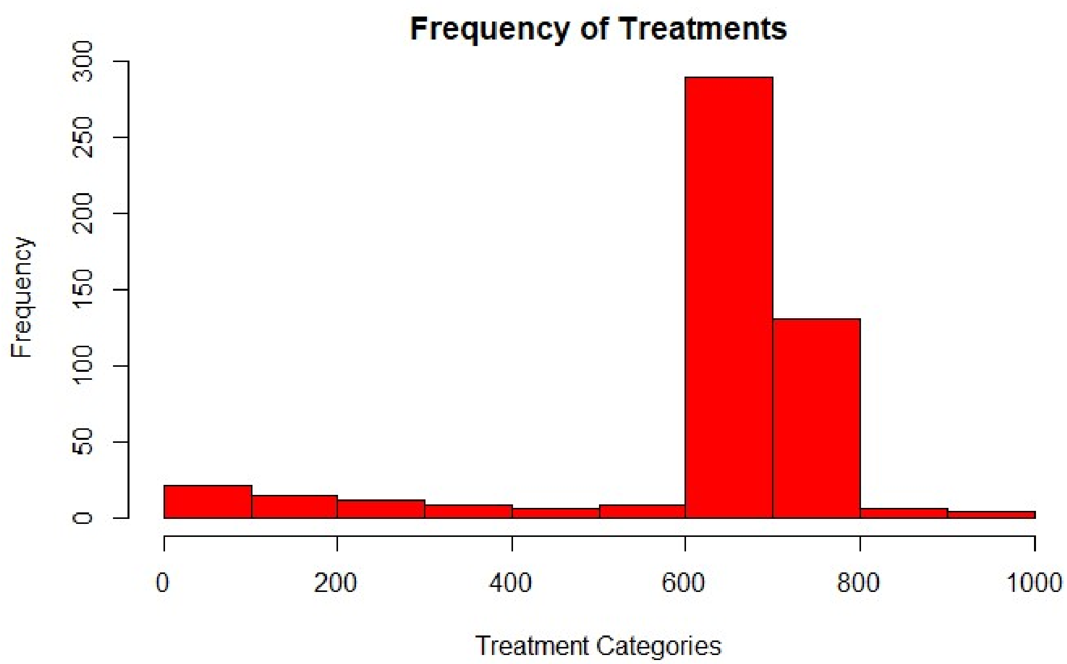
```{r}
DiagnosisCost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = Hospital_costs)
```

- Using a histogram to visualize the categories based on their frequencies.

Code:

```{r}
#Histogram depicting the categories with Diagnosis cost
```

```
hist(Hospital_costs$APRDRG,col = "cyan1",main = "Frequency of Treatments",xlab = "Treatment Categories")
```




```{r}
# Get the maximum diagnostic cost
DiagnosisCost[which.max(DiagnosisCost$TOTCHG),]
```

	APRDRG	TOTCHG
	<dbl>	<dbl>
44	640	437978

Conclusion: Hence we can conclude that category 640 has the maximum hospitalizations by a huge number (267 out of 500), along with this it also has the highest hospitalization cost i.e., 437978.

STEP-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.

* H0(Null Hypothesis): Independent variable (RACE) is not influencing dependent variable (COSTS)

* H0: There is no correlation among the residuals

* H1: There is a correlation among the residuals

* p-value < 0.5(significant value) ---> reject the null hypothesis

  - (we need high p-value so that we cannot reject the null hypothesis)

Code:

```{r}
summary(as.factor(Hospital_costs$RACE))  \\ To check for any null values.
```

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

Remove the "NA" values from the dataset, then factorize the Race variable to generate a summary, additionally to verify whether race made an impact on the hospital costs we will use ANOVA function with TOTCHG as dependent variable and RACE as grouping variable.

Code:

```{r}
Hospital_costs <- na.omit(Hospital_costs)
summary(as.factor(Hospital_costs$RACE))
```

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

- As it can be seen, 484 patients out of 499 fall under group 1, showing that the number of observations in group 1 is way higher than that of others.

- Hence the data is skewed. This will only affect the results from linear regression or ANOVA analysis.

Code:

```{r}
raceInfluenceAOV <- aov(TOTCHG ~ RACE, data = Hospital_costs)
```

raceInfluenceAOV

```

Call:

aov(formula = TOTCHG ~ RACE, data = Hospital_costs)

Terms:

| | RACE | Residuals |
|---|---|---|
| Sum of Squares | 2488459 | 7539623326 |
| Deg. of Freedom | 1 | 497 |

Residual standard error: 3894.903

Estimated effects may be unbalanced

Code:

```{r}

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 residual variance(deviation from original)(of all other variables) is very high.

- This implies that there is very little influence from RACE on hospitalization costs.

- Also, it can be seen that the degree of freedom(DF) for RACE is 1 and that of residuals is 497 obs.

- The F-value(test-statistic) is 0.16, which is much lesser than 0.5 showing that the race doesnt affect the hospitalization costs.

- The Pr(>F)(p-value) of 0.69 is high, conforming that RACE does not affect hospitalization costs.

Conclusion: F value is quite low, which means that variation between hospital costs among different races is much smaller than the variation of hospital costs within each race, and P value being quite high shows that there is no relationship between race and hospital costs, thereby accepting the Null hypothesis.
```

Additionally, we have more data for Race 1 in comparison to other races (484 out of 500 patients) which make the observations skewed and thus all we can say is that there isn't enough data to verify whether race of a patient affects hospital costs.

STEP-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.

Code:

```{r}
Hospital_costs$FEMALE <- as.factor(Hospital_costs$FEMALE)
summary(Hospital_costs$FEMALE)
```

```
  0   1
244 255
```

- There is an equal distribution of male and female in the group.

Now to analyze the severity of costs we will use linear regression with TOTCHG(Cost), a dependent variable along with AGE and Female as independent variables.

Code:

```{r}
ageGenderInflModel <- lm(TOTCHG ~ AGE+FEMALE, data=Hospital_costs)
summary(ageGenderInflModel)
```

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

- Since the p-values of AGE is much lesser than 0.05 and also lesser than 0.001, the ideal statistical significance level, it means AGE is the most statistically significant.

- Similarly, gender is also less than 0.05.

- Hence we can conclude that the model is statistically significant.

Conclusion: Age has more impact than gender according to the P-values and significant levels, also there are equal number of Females and Males and on an average (based on the negative coefficient values) females incur lesser hospital costs than males.

STEP-5: Since the length of stay is the crucial factor for in-patients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

Using linear Regression, we can show whether length of stay is dependent on age, gender or race. Here we LOS is the dependent variable and age, gender and race are independent variables

Code:

```
```{r}
Hospital_costs$RACE <- as.factor(Hospital_costs$RACE)
ageGenderRaceInflModel=lm(LOS ~ AGE+FEMALE+RACE, data=Hospital_costs)
summary(ageGenderRaceInflModel)
```
```

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

- The p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between these variables and length of stay.

Conclusion: p-values for all independent variables are quite high thus signifying that there is no linear relationship between the given variables, finally concluding the fact that we can’t predict length of stay of a patient based on age, gender and race.

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

Significance method- build a model using all independent variables vs dependent variable

Using linear Regression, we can show which variable affects the hospital costs the most, thus TOTCHG becomes dependent variable and rest all variables are taken as independent.

Code:

```{r}
HospitalcostModel <- lm(TOTCHG ~ ., data=Hospital_costs)
summary(HospitalcostModel)
```

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

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
LOS            742.9637   35.0464  21.199   < 2e-16 ***
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 ***
---
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
```

- As it is apparent that, from the coefficient values Age, Length of stay (LOS) and Patient Redefined diagnosis related groups (APRDRG) are the once with the statistical significance.

- Race is least significant.

- Build a model removing RACE

Code:

```
```{r}
hospcostM1 = lm(TOTCHG ~ .-RACE, data = Hospital_costs)
summary(hospcostM1)
```
```

```
Call:
lm(formula = TOTCHG ~ . - RACE, data = Hospital_costs)

Residuals:
  Min    1Q Median    3Q   Max
-6344  -687  -168   132 43387

Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 4971.980    433.116   11.480    < 2e-16 ***
AGE          134.241     17.462    7.688   8.16e-14 ***
FEMALE1     -383.082    247.571   -1.547      0.122
LOS          743.618     34.914   21.298    < 2e-16 ***
APRDRG        -7.767      0.681  -11.405    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2613 on 494 degrees of freedom
Multiple R-squared:  0.5528,  Adjusted R-squared:  0.5492
F-statistic: 152.7 on 4 and 494 DF,  p-value: < 2.2e-16
```

- Since FEMALE (Gender) is not having much significance in the model i.e., least significant, build the model removing FEMALE.

Code:

```{r}
hospcostM2 = lm(TOTCHG ~ .-(RACE+FEMALE), data = Hospital_costs)
summary(hospcostM2)
```

Call:

lm(formula = TOTCHG ~ . - (RACE + FEMALE), data = Hospital_costs)

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

- Since APRDRG has negative t-value, build the model dropping it (negative correlation).

Code:

```{r}
hospcostM3 = lm(TOTCHG ~ AGE+LOS, data = Hospital_costs)
summary(hospcostM3)
```

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

Comparing all the Models:

Code:

#Creating the required columns:

```{r}
Data <- rep("HospitalCosts",4)
Approach <- rep("Ap1:significance",4)
Model_Name <- c("HospitalcostModel", "hospcostM1", "hospcostM2", "hospcostM3")
Detail <- c("signif, all independent variables", "-RACE", "-(RACE+FEMALE(gender))", "AGE+LOS")
R2 <- c(0.554,0.553,0.551,0.419)
adj_R2 <- c(0.549,0.549,0.548,0.416)
std_err <- c(2610, 2610, 2620, 2970)
R2_sub_adjR2 <- c(0.005,0.004,0.003,0.003)
pvalue <- rep("<2e-16", 4)
```

```

#Structuring the columns of the Dataframe/ Table:

```{r}

ModelTable <- data.frame(Data, Approach, Model_Name, Detail, R2, adj_R2, std_err, R2_sub_adjR2, pvalue)

names(ModelTable)[c(3,6,7,8)] <- c("Model Name","adj R2","std error", "R2 - adj R2")

ModelTable

```

Data <chr>	**Approach** <chr>	**Model Name** <chr>	**Detail** <chr>	**R2** <dbl>	**adj R2** <dbl>	**std error** <dbl>	**R2-adj R2** <dbl>	**pvalue** <chr>
HospitalCosts	Ap1:significance	Hospitalcost Model	signif, all independent variables	0.554	0.549	2610	0.005	<2e-16
HospitalCosts	Ap1:significance	hospcostM1	-RACE	0.553	0.549	2610	0.004	<2e-16
HospitalCosts	Ap1:significance	hospcostM2	-(RACE+FEMALE(Gender))	0.551	0.548	2620	0.003	<2e-16
HospitalCosts	Ap1:significance	hospcospM3	AGE+LOS	0.419	0.416	2970	0.003	<2e-16

- Removing Race[RACE] and Gender[FEMALE] doesn't change the R2 value. It doesn't impact cost
- Removing APRDRG in model **hospcostM2** increases the standard error. Hence model **hospcostM2** seems to be better.

Conclusion: Age and length of stay affect the total hospital costs. Additionally, there is positive relationship between length of stay to the cost, so with an increase of 1 day there is an addition of a value of 742 to the cost.
```