PROJECT NO. 07

DECEMBER 2020

HEALTHCARE COST ANALYSIS

PREPARED BY
JAY SHEMBEKAR

ShembekarJay@gmail.com
+91 - 7898497385

PRIMARY GOAL

ANALYZE HEALTHCARE COST AND UTILIZATION IN WISCONSIN HOSPITALS

Background

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

Data-set 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)
TOTCHA	Hospital discharge costs
APRDG	All Patient Refined Diagnosis Related Groups

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 must 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.

To make this analysis informative, more than the requirement of this project, I have put some extra efforts & went out of the box, trying my level best to make it interesting and insightful.

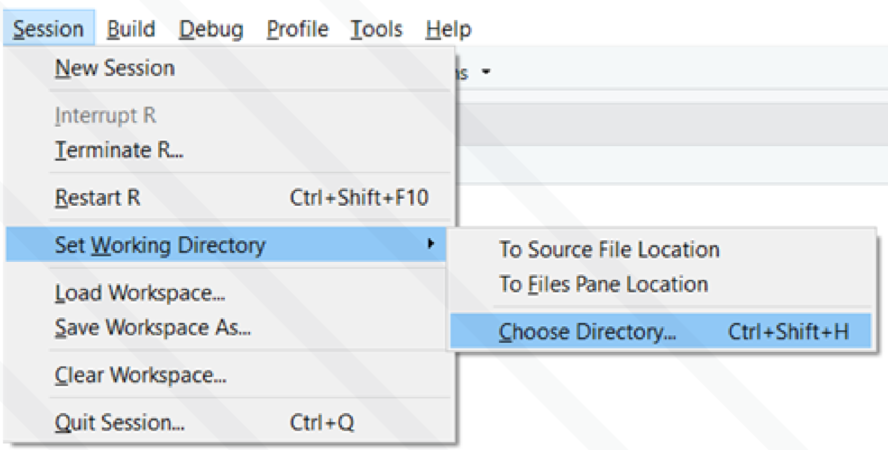
First and foremost, let us understand the dataset. We have been asked to analyze the data of Wisconsin Hospitals based on some attributes and how they are affecting the total costs involved in a treatment, length of stay of the patients and, we will investigate the practices followed in Wisconsin Hospitals based on Race.

The tool we will be using for this analysis is **rStudio.**

To start working on the dataset, first we will check the dataset extension which we will be working on. To do that we will use **NCmisc** library (Miscellaneous Functions) and **get.ext()** function along with full dataset name ("1555054100_hospitalcosts.xlsx"), as below:

```
> library(NCmisc)
> get.ext("1555054100_hospitalcosts.xlsx")
1555054100_hospitalcosts.xlsx
                       "xlsx"
```

From above output we can see that the extension is xlsx. Now to start working on it, first we need to set working directory to the path where it is stored on machine. To do so, navigate to
Tab - **Session > Set Working Directory > Choose Directory >**
and select the location where the dataset is present.



Or, we can press "**Ctrl + Shift + H**" together and then select the location where the dataset is present.

As we saw earlier, dataset is present in XLSX format, which is an Excel file, we will use **readxl** library from CRAN repository to import it into rStudio as "**Healthcare**". Use **head()** to see its first 6 rows.

```
> library(readxl)
> Healthcare <- read_excel("1555054100_hospitalcosts.xlsx")
> head(Healthcare)
# A tibble: 6 x 6
    AGE FEMALE   LOS  RACE TOTCHG APRDRG
  <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
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
```

This confirms that the dataset is successfully loaded into rStudio. Now we will check internal structure of the dataset using **str()** function.

```
> 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 ...
```

As we can see that all the columns are of Numerical Values, **num**. We will consider converting some of them to **factor** as it is helpful in categorizing data and storing it on multiple levels.

Using **summary()** we can get an overview of minimum and maximum values in every column, their 1st and 3rd Quartiles as well as Means & Medians.

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

Note: The RACE column has a **NULL** value, **NA's : 1,** this will come in picture at a later point.

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.

- First, we will check which Age category has frequently visited the hospital.

To do so, first we will convert AGE column values to factors by using **as.factor()** function and save it in a new column **AGE_New**.

Then we will create a new data frame using AGE_New column (**Healthcare$AGE_New**) and we will use **data.frame(summary())** function which will directly sum up the Number of Patients for every Age.

The name our data frame is **AGE_Dataframe**.

```
> Healthcare$AGE_New <- as.factor(Healthcare$AGE)
> AGE_Dataframe <- data.frame(summary(Healthcare$AGE_New))
> AGE_Dataframe
   summary.Healthcare.AGE_New.
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
```

We can see from the above data frame that the maximum frequency is for AGE group, "0" or **Infants**. To print the maximum value from above data frame we can use **max()** function as:

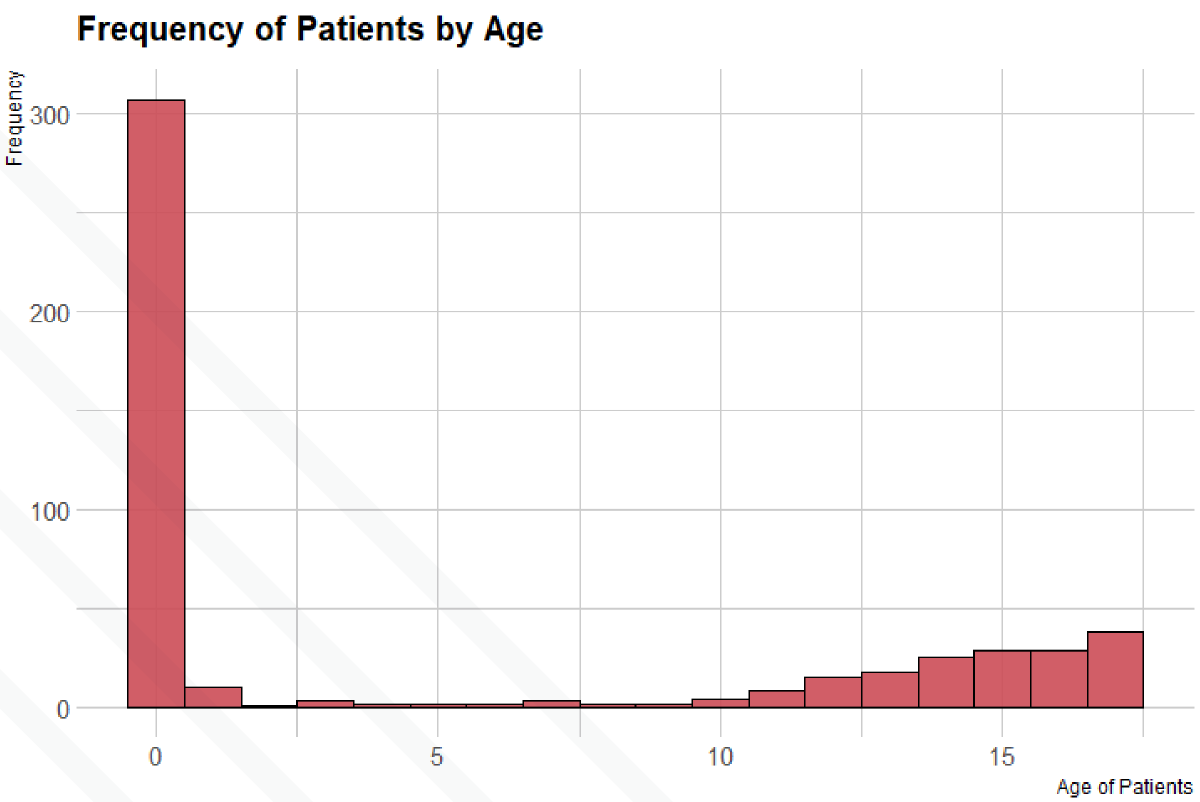
```
> paste("Max number of Patients for AGE - 0 are : ",
+       max(AGE_Dataframe))
[1] "Max number of Patients for AGE - 0 are :  307"
```

To view the same result in a graphical representation, it is relatively straightforward to build a histogram with **ggplot2**, thanks to the **geom_histogram()** function.

Here, we are using **hrbrthemes** library along with **tidyverse** which will attach **ggplot2** as well.

I've named this plot as **Plot_1**.

```
> library(tidyverse)
> library(hrbrthemes)
> Plot_1 <- Healthcare %>%
+
+   filter( AGE < 18 ) %>%
+
+   ggplot( aes(x=AGE)) +
+
+   geom_histogram( binwidth=1,
+                   fill="#cc4d56",
+                   color="black",
+                   alpha=0.9) +
+
+   ggtitle("Frequency of Patients by Age") +
+
+   theme_ipsum() +
+
+   labs(y= "Frequency",
+        x = "Age of Patients") +
+
+   theme(plot.title = element_text(size=15))
> Plot_1
```



Here, **fill = "#cc4d56"** is nothing but HEX code of the color representing the bars. The reason being so particular about selecting this color is that it's the favorite color of one my close friends.

- Secondly, to figure out which Age group has the maximum expenditure we will use **aggregate()** function along side with **which.max()** function.

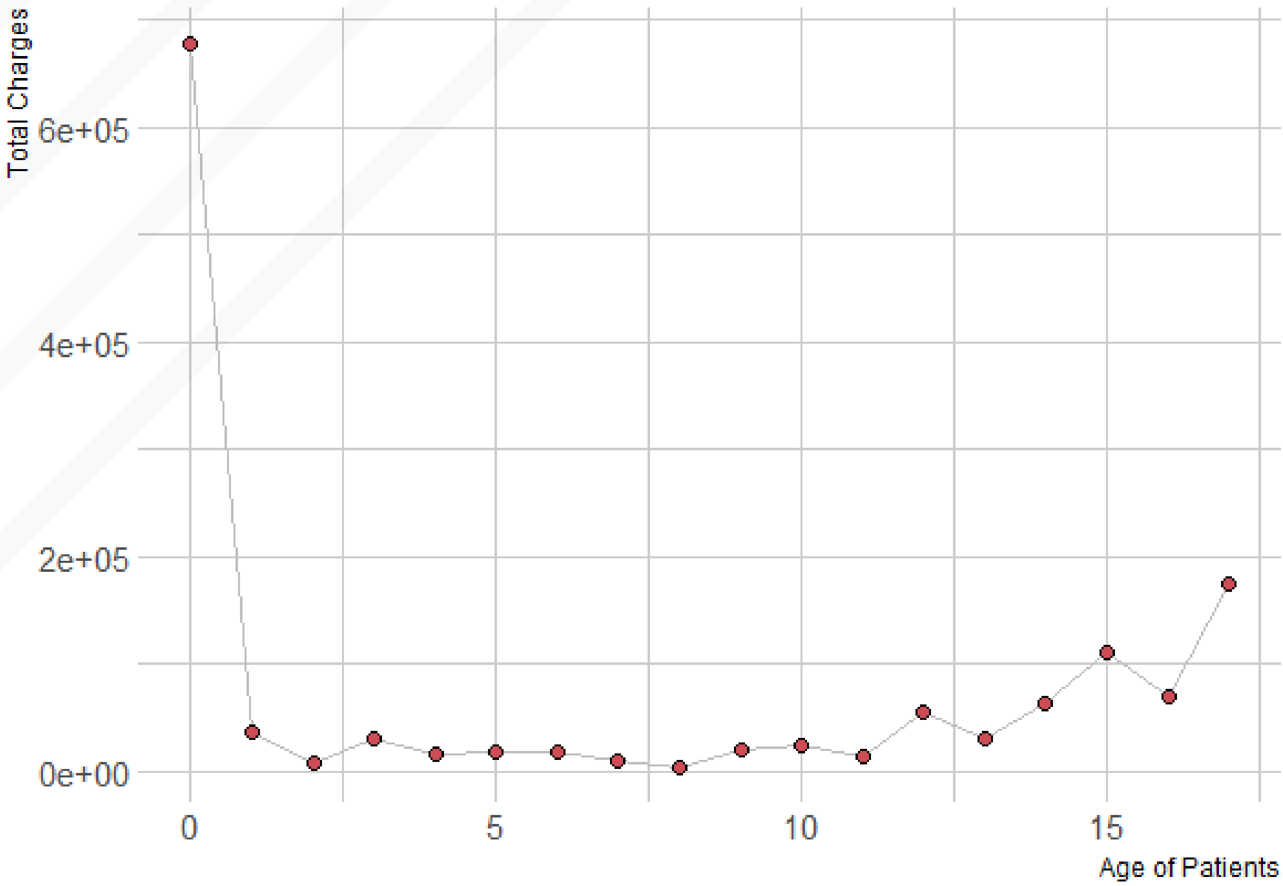
I've named this newly created data frame as **AGE_Aggregated**.

```
> AGE_Aggregated <- aggregate(TOTCHG ~ AGE, FUN = sum,
 data = Healthcare)
> AGE_Aggregated[which.max(AGE_Aggregated$TOTCHG),]
  AGE TOTCHG
1   0 678118
```

It is clear from the above observation that Age group - "0" has the maximum expenditure. This can be again represented graphically using the same libraries as before. The plot name is **Plot_2**.

```
> Plot_2 <- AGE_Aggregated %>%
+
+   filter(AGE < 18) %>%
+
+   ggplot(aes(x=AGE, y=TOTCHG)) +
+
+   geom_line( color="grey") +
+
+   geom_point(shape=21,
+              color="black",
+              fill="#cc4d56", size=2) +
+
+   labs(y= "Total Charges",
+        x = "Age of Patients") +
+
+   theme_ipsum() +
+
+   ggtitle("Maximum Expenditure by Age")
> Plot_2
```

Maximum Expenditure by Age



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.

- First of all let's understand what is APRDRG. All Patients Refined Diagnosis Related Groups (APR-DRG) is a classification system that classifies patients according to their:
    - Reason of Admission.
    - Severity of Illness.
    - Risk of Mortality.
    - Average length of Stay.
    - Cost Outlier Threshold.

DRGs were first implemented nationwide by the Health Care Financing Administration, **HCFA** to help control costs for inpatient services billed to Medicare. Hawaii Medical Service Association, **HMSA** first began using DRGs when the Preferred Provider Plan was developed in 1989 and gradually implemented DRGs for other HMSA plans.

Here's the list of few APR-DRG Weights provided by Department of Health, NY.

July 1, 2018 APR-DRG Service Intensity Weights, Average Length of Stay and High Cost Outlier Thresholds

APR-DRG	Severity	APR-DRG Description	Service Intensity Weight	Average Length of Stay	Cost Outlier Threshold
021	1	Craniotomy except for trauma	1.8941	8	$2,11,794
023	1	Spinal procedures	1.6410	6	$1,57,615
049	1	Bacterial & tuberculous infections of nervous system	1.1648	9	$1,63,210
050	1	Non-bacterial infections of nervous system exc viral meningitis	0.7505	7	$1,14,573
051	1	Viral meningitis	0.6172	4	$35,920
053	1	Seizure	0.4880	3	$36,007
640	1	Neonate birthwt >2499g, normal newborn or neonate w other problem	0.1749	2	$10,159
751	1	Major depressive disorders & other/unspecified psychoses	0.8581	7	$58,876
753	1	Bipolar disorders	0.8284	6	$64,804
754	1	Depression except major depressive disorder	0.6929	4	$21,794
755	1	Adjustment disorders & neuroses except depressive diagnoses	0.5103	3	$31,892
758	1	Behavioral disorders	0.5380	5	$45,049
952	1	Nonextensive procedure unrelated to principal diagnosis	0.7849	6	$88,793

- Now, let's start with finding the maximum hospitalization by Diagnostic-related Group. We will use a simpler method than what we used when answering Frequency of Patients by Age.

Using **plyr** library and **count()** function, we will create a new data frame which will have Frequency of Patients by APRDRG.

The name of this data frame is **APRDRG_Dataframe**.

```
> APRDRG_Dataframe <- count(Healthcare, 'APRDRG')
> head(APRDRG_Dataframe)
  APRDRG freq
1     21    1
2     23    1
3     49    1
4     50    1
5     51    1
6     53   10
```

To print the maximum value of frequency from above data frame we can use **which.max()** function as:

```
> APRDRG_Dataframe[which.max(APRDRG_Dataframe$freq),]
   APRDRG freq
44    640  267
```

This tells us that the maximum number of hospitalization is from Diagonstic-related Group - **640** with a frequency of **267**.

- Secondly, to figure out which Diagnostic-related group has the maximum expenditure we will use **aggregate()** function along side with **which.max()** function.

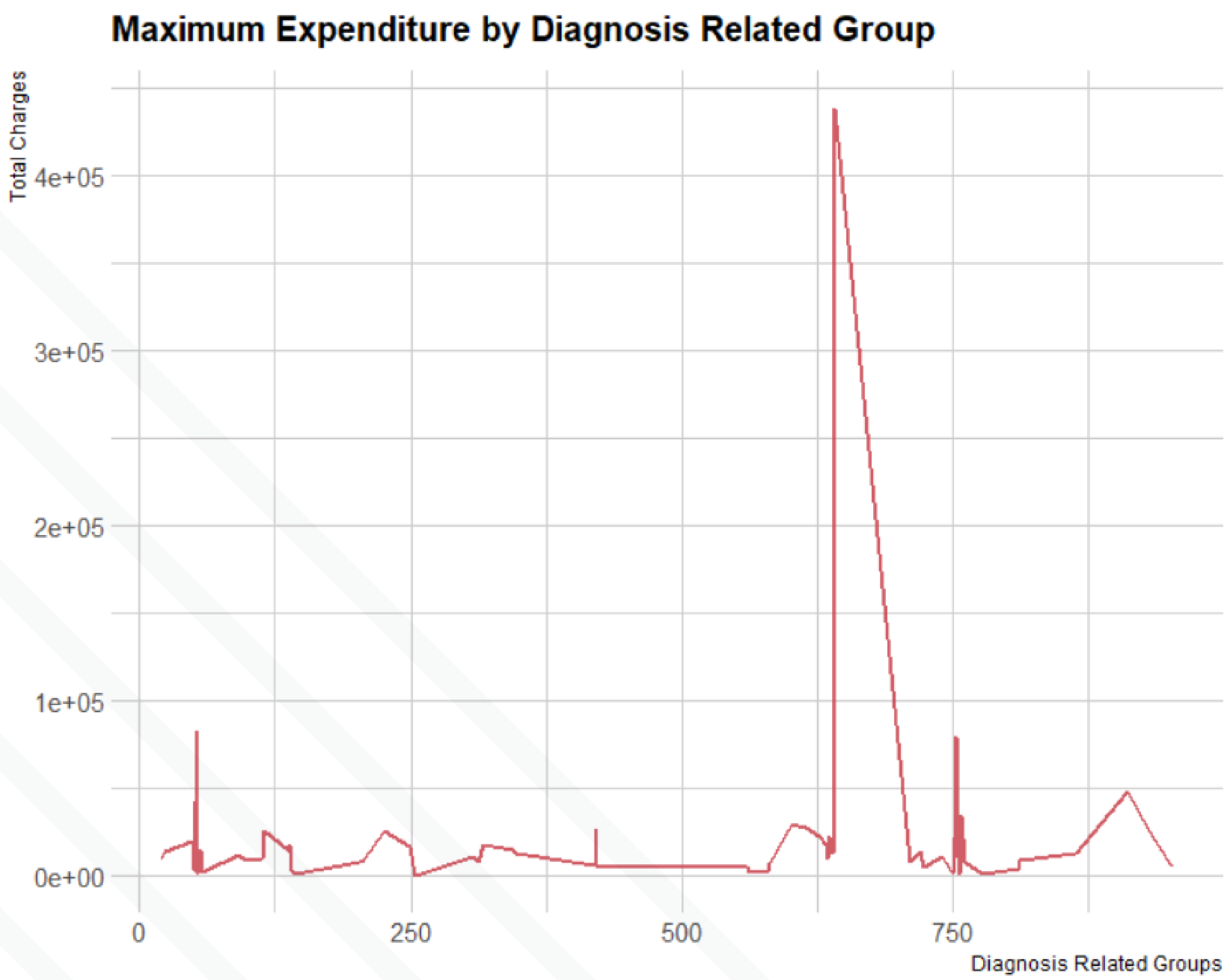
I've named this newly created data frame as **APRDRG_Aggregated**.

```
> APRDRG_Aggregated <- aggregate(TOTCHG ~ APRDRG,
+                               FUN = sum,
+                               data = Healthcare)
> APRDRG_Aggregated[which.max(APRDRG_Aggregated$TOTCHG),]
   APRDRG TOTCHG
44    640 437978
```

It is clear from the above observation that APRDRG group - "640" has the maximum expenditure of **437978 units**.

The above observation can be again represented graphically using the same libraries as before. The plot name is **Plot_3**.

```
> Plot_3 <- ggplot(APRDRG_Aggregated, aes(x=APRDRG, y=TOTCHG)) +
+
+   geom_line( color="#cc4d56", size=1, alpha=0.9) +
+
+   labs(y= "Total Charges",
+        x = "Diagnosis Related Groups") +
+
+   ggtitle("Maximum Expenditure by Diagnosis Related Group") +
+
+   theme_ipsum() +
+
+   theme(plot.title = element_text(size=15))
> Plot_3
```



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.

- In this dataset, RACE column is defined by Numeric codes, thus we cannot identify the Race of a patient. Though, a little research would tell us that **US Department of Labor** issues a Coding Instruction for Race & Ethnicity, which defines the codes in our dataset as follows:

VALUE	DESCRIPTION
1	White, not Hispanic
2	Black, not Hispanic
3	Hispanic
4	American Indian or Alaskan Native
5	Asian or Pacific Islander
6	Other
-1	Missing or Unknown

There is a missing value which we saw in our initial analysis and this is the time to handle it.

We will create a new dataset, **Healthcare_New** by omitting the NA value using **na.omit()** function and then we will again check if we have any NA's in our newly created dataset using **colSums(is.na())** function, as follows:

```
> Healthcare_New <- na.omit(Healthcare)
> colSums(is.na(Healthcare_New))
    AGE  FEMALE     LOS    RACE  TOTCHG  APRDRG AGE_New
      0       0       0       0       0       0       0
```

As we can see, all columns are showing "0" values, which means there are no blanks or NA's left in our new dataset.

To check if Race impacts Hospitalization costs, we will run Analysis of Variance, ANOVA test using **aov()** function where TOTCHG will be dependent and RACE an independent variable.

Also, we will convert values in column RACE into factors before using ANOVA function. The model is named as **AOV_Model**.

```
> Healthcare_New$RACE <- as.factor(Healthcare_New$RACE)
> AOV_Model <- aov(TOTCHG ~ RACE, data = Healthcare_New)
> AOV_Model
Call:
   aov(formula = TOTCHG ~ RACE, data = Healthcare_New)

Terms:
                     RACE  Residuals
Sum of Squares   18593279 7523518505
Deg. of Freedom         5        493

Residual standard error: 3906.493
Estimated effects may be unbalanced
```

Running **summary()** function on above Model will give us the p-value and F- value, which can be analyzed for our further study.

```
> summary(AOV_Model)
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          5 1.859e+07  3718656   0.244  0.943
Residuals   493 7.524e+09 15260687
```

Conclusion

As we can see, the p-value is **0.943** which is quite higher than the significance level, Alpha which is generally considered to be **0.05**.

Therefore, we can reject the assumption that Race affects the expenditure of Patients. Also, by looking at the F-value of **0.244** which is quite low, it is safe to say that variation between hospital costs among different races is much smaller than the variation of hospital costs within each race.

To support the above statement we will also check the frequency of patients for each Race by using summary() function.

```
> summary(Healthcare_New$RACE)
  1   2   3   4   5   6
484   6   1   3   3   2
```

The above observation tells us that Race - "**1**" has **484/500** patients which makes the analysis skewed and thus we can say that we don't have enough data to verify whether Race of a patient affects hospital costs.

4 To properly utilize the costs, the agency must analyze the severity of the hospital costs by Age and Gender for the proper allocation of resources.

- To check if Age and Gender impacts Hospitalization costs, we will do a Linear Regression test using **lm()** function where **TOTCHG** will be dependent and **AGE + FEMALE** will be an independent variables.

Also, we will convert values in column FEMALE into factors before using LM function. The model is named as **LM_Model**.

```
> Healthcare$FEMALE <- as.factor(Healthcare$FEMALE)
> LM_Model <- lm(TOTCHG~AGE + FEMALE, data = Healthcare_New)
> summary(LM_Model)

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

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 ***
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
```

Conclusion

The above output tells us that both Age and Gender are significant enough to impact Hospitalization Costs.

However, The p-value for AGE is much lower than the p-value for FEMALE which tells that AGE has more significance on Cost of Hospitalization than Gender.

Also, the Estimate value for Gender is **-744.21** which means that Males incur more expenses than a Female patient by a value of 744.21**.**

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.

- To check if Age, Gender & Race impacts Hospitalization costs, we will again use Linear Regression Model where **LOS** will be dependent and **AGE + FEMALE + RACE** will be an independent variables.

The model is named as **LM_Model_2**. Also, we have used **Healthcare_New** dataset this time so as to exclude NULL value from RACE column.

```
> LM_Model_2 <- lm(LOS ~ RACE + FEMALE + AGE, data = Healthcare_New)
> summary(LM_Model_2)

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

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 ***
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    
AGE         -0.03938    0.02258  -1.744   0.0818 .  
---
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

The above output tells us that none of the Independent variables are significant enough as the p-values are more than Alpha value, 0.05.

So, we can conclude that Length of Stay is not significantly affected by Age, Gender or Race.

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

- Again, I'll be using Linear Regression Model to check which variable affects Hospitalization costs.

The model is named as **LM_Model_3**.

```
> LM_Model_3 <- lm(TOTCHG ~ AGE + FEMALE +
+                    RACE + LOS + APRDRG,
+                  data = Healthcare_New)
> summary(LM_Model_3)

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

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

Conclusion

The above output tells us that:

- Age,
- Length of Stay &
- Diagnosis related Groups

highly affects the Hospitalization costs as the p-values for them are very small compared to significance value, Alpha.

Also, by looking at the estimate value of LOS, we can say that for every increment in LOS the Total Charges are increased by a value of **742.9637**

Summary

This concludes our analysis of the dataset on Healthcare cost and Utilization in Wisconsin hospitals.

To sum up everything, here are the snaps which concludes our analysis:

1 Age group - "0" or **Infants**, have the most number of admissions in the Hospital and therefore incurs the highest amount of Total Charges all together.

2 Diagnosis related Group - "**640**", has the highest number of admissions in the Hospital with maximum expenditures incurred.

- Moreover, when I checked for APR-DRG value 640, it is described as **Neonate birthwt >2499g, normal newborn or neonate w other problem** which is related to Infants. This confirms above observation that the maximum number of admissions in the Hospital are indeed Infants.

3 We don't have enough evidence to claim any malpractices in hospital based on Race.

4 AGE has more significance on Cost of Hospitalization than Gender although they are both significant.

5 Length of Stay is not significantly affected by Age, Gender or Race.

6 Hospitalization Cost is highly affected by the following factors:

- Age,
- Length of Stay &
- Diagnosis related Groups.