Project	Healthcare Cost Analysis
Author	Rakesh Gorai
Date	02-Feb-2020

Table of Contents

1. Introduction

1.1 Business Scenario

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.

1.2 Dataset Description

Variable	Description	Remarks
AGE	Age of the patient discharged	AGE is categorical variable.
FEMALE	A binary variable that indicates if the patient is female.	FEMALE is categorical variable and 0 represents male and 1 represents female.
LOS	Length of stay in days	LOS is continuous variable.
RACE	Race of the patient (specified numerically)	RACE is categorical variable.
TOTCHG	Hospital discharge costs	TOTCHG is continuous variable.
APRDRG	All Patient Refined Diagnosis Related Groups	APRDRG is categorical variable.

1.3 Assumptions

Assumptions
1. Assumptions for normal distribution, homoscedasticity and non-collinearity hold true wherever applicable.
2. In most of cases , the number of observation is > 30, so we assume normality due to central limit theorem and hence will use parametric formulas wherever applicable.
3. Case where the data doesn’t meet the criteria for parametric test, non-parametric tests were performed e.g kw test for comparison of means of different RACE types.
4. When we are building model for predicting cost, we will be using linear regression model (lm) instead of glm as cost (output) is continuous variable.

1.4 Summary Of Tests Used

Test Name	Test Purpose	When To Use	References

Kruskal-Wallis test	Mean Comparison between groups	When the conditions for one-way ANOVA tests are not met. e.g Average cost for Race as the number of observations for some Race categories are very less (1 0r 2) and hence will impact the median between different race groups.	https://www.statisticshowto.datasciencecentral.com/kruskal-wallis/
One-way ANOVA	Mean Comparison between groups	a) Normality b) Sample independence c) Variance Equality	https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553
Eta Square test	Correlation	This correlation test is used when a) One variable is continuos and one is categorical b) The categorical variable has level >2	https://www.researchgate.net/post/Can_I_use_Pearsons_correlation_coefficient_to_know_the_relation_between_perception_and_gender_age_income To interpret the results: https://www.rdocumentation.org/packages/sjmisc/versions/1.8/topics/eta_sq
Point-Biserial Test	Correlation	To be used when one variable is continuos and one is dichotomous categorical variable. e.g Cost vs Gender	References : https://www.statisticssolutions.com/point-biserial-correlation/

2. Solution

2.1 Business Goal 1 : Find Maximum Hospital Visit And Maximum Spend By AGE Group

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

2.1.1 Analysis, Coding and Results

2.1.1.1 Data Summary

Considering, AGE as a categorical variable and has to be converted to factor data type. We will organize and summarize the data as follows:

- Load the data into a data frame and modify the data frame to have 2 more columns. one column having the corresponding frequency (count) for each age group and other column having the sum of total charges aggregated by age group

- Arrange the above data frame by descending order of frequency of visit and Sum total charge. This will give insight into the age group having frequent hospital visit and maximum spend.

R-CODE:

```
# Set the working directory to the path having the data file.
# setwd(<Actual Path>)

#Load the relevant libraries
library(readxl) # to read xlsx file
library(dplyr) #to use the group by, add_tally and add_count, arrange functions
library(magrittr) #for piping a output to another input
library(ggplot2)

#Load The xls file from physical location to a date frame
hospitalcosts <- read_excel("1555054100_hospitalcosts.xlsx")

# 1. Copy the dataframe hospitalcosts to a new dataframe hospitalcostsAge
# 2. Group by Age
# 3. Find the sum of total charges paid grouped by age and append the new column to the
newly formed dataframe
# 4. Find the number of visits by age and append the new column to the newly formed
dataframe
hospitalcostsAge <- hospitalcosts %>% group_by(AGE) %>%
  add_tally(wt = TOTCHG,name = "TotChgByAge") %>% add_count(name = "CountByAge")

hospitalcostsAge$AGE <-  as.factor(hospitalcostsAge$AGE)

# Order the dataset in descending order of Number of visits per age, Sum Total Charge per
age
hospitalcostsAge <-  hospitalcostsAge %>%
arrange(desc(hospitalcostsAge$CountByAge,hospitalcostsAge$TotChgByAge))

# below will give which age group frequented the visit and also the maximun charge and total
expenditure in each age group
group_by(hospitalcostsAge, AGE) %>%
  summarise(
    "Frequncy Of Visits" = n(),
    "Max  Charge" = max(TOTCHG),
    "Sum Total Charge By Age " = max(TotChgByAge)
  )
```

Results

	AGE	`Frequncy Of Visits`	`Max Charge`	`Sum Total Charge By Age `
	<fct>	<int>	<dbl>	<dbl>
1	0	307	29188	678118
2	1	10	9606	37744
3	2	1	7298	7298
4	3	3	14243	30550
5	4	2	9230	15992
6	5	2	10584	18507
7	6	2	9530	17928
8	7	3	6425	10087
9	8	2	3588	4741
10	9	2	10585	21147
11	10	4	17524	24469
12	11	8	3908	14250
13	12	15	17434	54912
14	13	18	5615	31135
15	14	25	10756	64643
16	15	29	20195	111747
17	16	29	10002	69149
18	17	38	48388	174777

From above we could see that the age group 0 has the most hospital visit and also has the most spend.

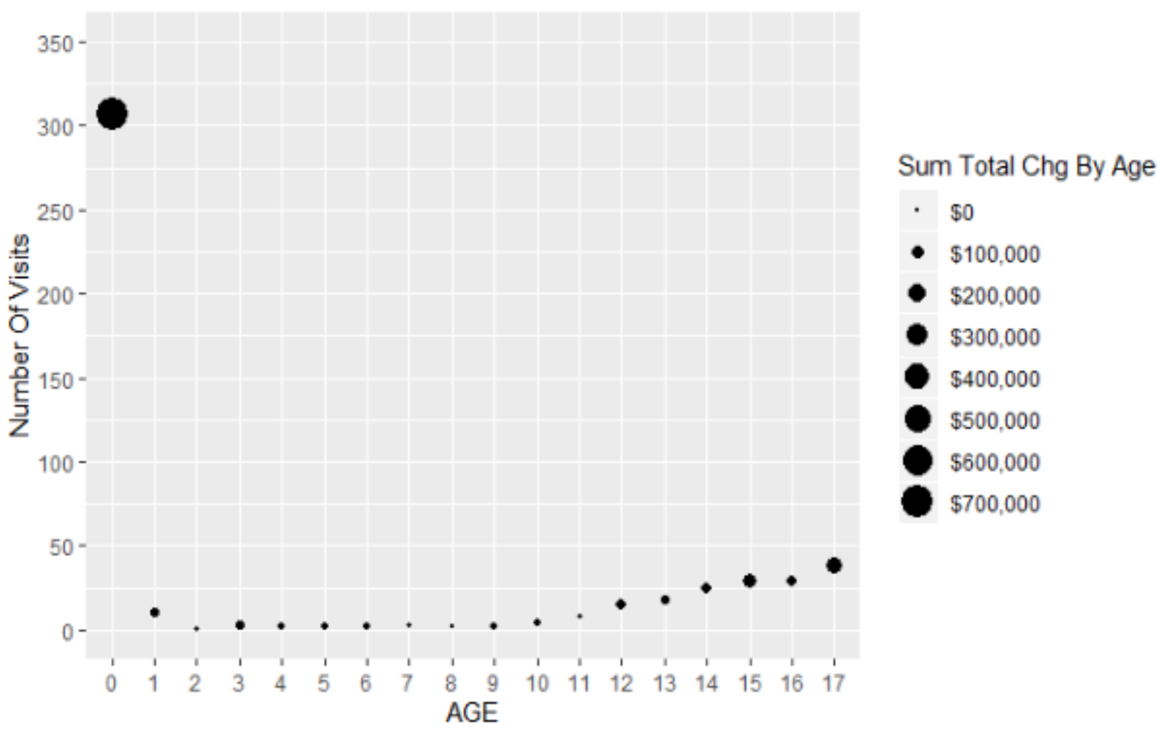
2.1.1.2 Data Visualization

We can draw a scatter plot between age and frequency of visit to get a visual insight of which age group frequently visit the hospital.

And then we add the “Sum Total Charge by Age” representing the size of the points to get a glimpse of how the “frequency of visit” and “Sum Total Charge by Age” are distributed for different age groups.

```
# Draw a scatter plot with "number of visits" on Y-axis and AGE on X-axis and "Total
Charge by Age" as a size factor
# This will help us understand which age group have maximum visits and maximum total
spend as hospital charges
ggplot(hospitalcostsAge,aes(y = hospitalcostsAge$CountByAge,x =
hospitalcostsAge$AGE))+
  geom_point(aes(size=hospitalcostsAge$TotChgByAge) )+
  scale_y_continuous (name = "Number Of Visits", breaks
                    = seq (0, 350, 50), limits=c(0,350))+
  scale_size_area(name = "Sum Total Chg By Age", breaks
                 = seq (0, 700000, 100000), limits=c(0,700000),labels = scales::dollar_format())+
  scale_x_discrete(name="AGE" ,limits = c("0",1:17),
                  labels =  c("0",1:17))
```

Screenshots



Again, we could see above that the age group 0 has the most hospital visit and also has the most spend.

2.2 Business Goal 2 : Maximum Hospital Visit And Maximum Spend By APRDRG

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.

2.2.1 Analysis, Coding and Results

2.2.1.1 Data Summary

Again APRDRG is also a categorical variable so we will factor it. We can use following code to group the data by APRDRG and also check the factor level for APRDRG. We will follow below steps to summarize the data :

- Load the data into a data frame and modify the data frame to have 2 more columns. one column having the corresponding frequency of visit (count) for each APRDRG group and other column having the sum of total charges aggregated by APRDRG

- Also, we will find the top 10 record by APRDRG that has frequent visit to hospital and top 10 records by APRDRG that has most spend.

R-Code:

```
# 1. Copy the dataframe hospitalcosts to a new dataframe hospitalcostsDRG
# 2. Group by APRDRG
# 3. Find the sum of toal charges paid grouped by the APRDRG and append the new column
to the newly formed dataframe
# 4. Find the number of visits by APRDRG and append the new column to the newly formed
dataframe

hospitalcostsDRG <- hospitalcosts %>% group_by(APRDRG) %>%
  add_tally(wt = TOTCHG,name = "TotChgByDRG") %>% add_count(name = "CountByDRG")


# Order the dataset in descending order of Number of visits by APRDRG, Sum Total Charge
by APDRG
hospitalcostsDRG <- hospitalcostsDRG %>%
arrange(desc(hospitalcostsDRG$CountByDRG,hospitalcostsDRG$TotChgByDRG))

# Convert the APRDRG to factors
hospitalcostsDRG$APRDRG <- as.factor(hospitalcostsDRG$APRDRG)
str(hospitalcostsDRG) # verify that APRDRG is converted to factors
length(levels(hospitalcostsDRG$APRDRG))

# Find the top 10 APRDRG by "number of visits". From this the first row will be the APRDRG
having maximum visits
top_DRG_ByCount <- group_by(hospitalcosts, APRDRG) %>%
  summarise(Count=n()) %>%
  top_n(n=10) %>% arrange(desc(Count))


# Find the top 10 APRDRG by Total Charges. From this the first row will be the APRDRG
having maximum Total Charges
top_DRG_ByTotChg <- group_by(hospitalcosts,APRDRG) %>%  summarise(sum =
sum(TOTCHG)) %>%
  top_n(n=10) %>% arrange(desc(sum))
```

Screenshots:

1. Levels of APRDRG

```
> length(levels(hospitalcostsDRG$APRDRG))
[1] 63
```

2. Top 10 APRDRG group having frequent visits:

	APRDRG	Count
	<dbl>	<int>
1	640	267
2	754	37
3	753	36
4	758	20
5	751	14
6	755	13
7	53	10
8	249	6
9	626	6
10	139	5

3. Top 10 APRDRG group having most spend:

	APRDRG	sum
	<dbl>	<dbl>
1	640	437978
2	53	82271
3	753	79542
4	754	59150
5	911	48388
6	758	34953
7	602	29188
8	614	27531
9	930	26654
10	421	26356

From above we could see that the APRDRG 640 has the most hospital visit and also has the most spend.

2.2.1.2 Data Visualization

To visualize the data in a scatter plot, we follow below steps:

- We will draw a scatter plot between APRDRG and frequency of visit to get a visual insight of which APRDRG group frequently visit the hospital.
- And then we add the "Sum Total Charge by APRDRG" representing the size of the points to get a glimpse of how the "frequency of visit" and "Sum Total Charge by APRDRG" are distributed for different APRDRG groups.

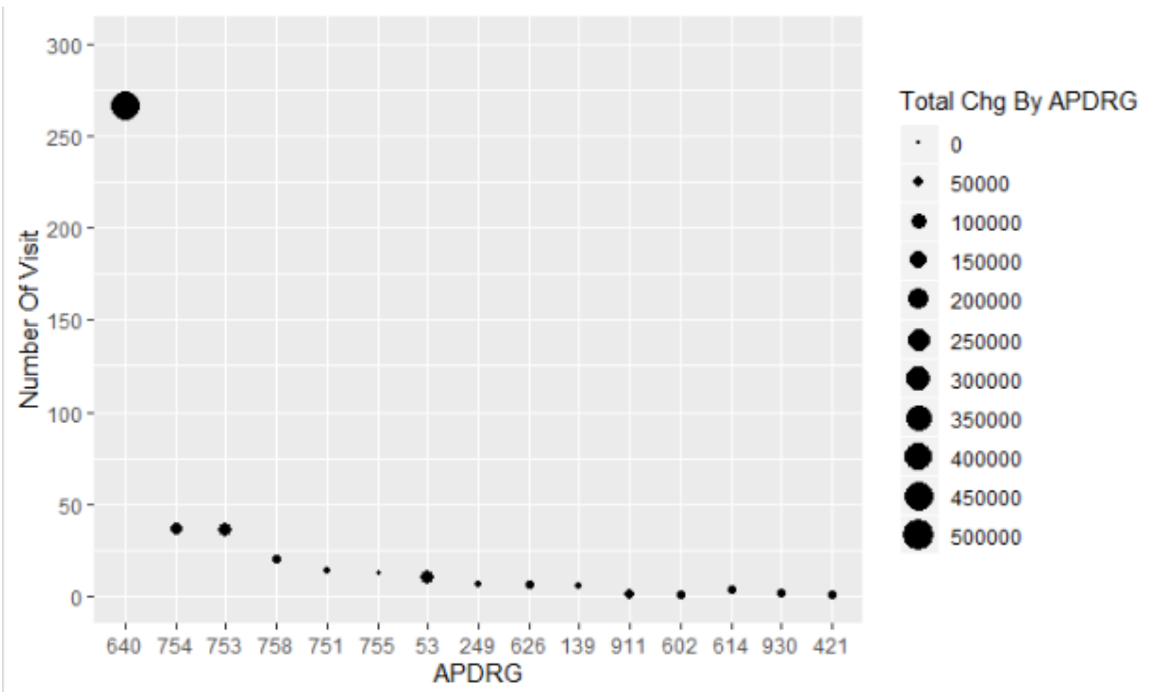
- But before we could use "Sum Total Charge by APRDRG" as a size attribute we have to scale it as there are 63 levels for APRDRG. So we will just take the top 10 records as follows:
  i. Find the top 10 APRDRG that visit the hospital frequently.
  ii. Find the top 10 APRDRG that has most spend.
  iii. Combine both the top 10 group to have all the APRDRG from both the above group.

R-CODE:

```
# Following will be used in selecting only top APRDRG in the x-axis for scatter plot below.
# As the top 10 APRDRG for "number of visits" and "Total Charges" might be different,
union will make sure data from both list are considered
DRG_DATA_LIMITS <- union(top_DRG_ByCount[1],top_DRG_ByTotChg[1])

# Draw a scatter plot with "number of visits" on Y-axis and APRDRG on X-axis and "Total
Charge by APRDRG" as a size factor
# This will help us understand which APRDRG group have maximum visits and maximum
spend as hospital charges
ggplot(hospitalcostsDRG,aes(y = hospitalcostsDRG$CountByDRG,x =
hospitalcostsDRG$APRDRG))+
  geom_point(aes(size=hospitalcostsDRG$TotChgByDRG) )+
  scale_y_continuous (name = "Number Of Visit", breaks
                  = seq (0, 300, 50), limits=c(0,300))+
  scale_size_area(name = "Total Chg By APDRG", breaks
                = seq (0, 500000, 50000), limits=c(0,500000))+
  scale_x_discrete(name="APDRG" ,limits = as.character(unlist(DRG_DATA_LIMITS)),
                 labels = as.character(unlist(DRG_DATA_LIMITS)))
```

Screenshots:



Again, we could see above that the APRDRG 640 has the most hospital visit and also has the most spend.

2.3 Business Goal 3 : Analyze If Malpractice In Hospital Cost Based On Race

To make sure that there is no malpractice, the agency needs to analyse if the race of the patient is related to the hospitalization costs and also if the average cost varies for different RACE type.

2.3.1 Analysis, Coding and Results

2.3.1.1 Kruskal-Wallis test and ETA Square Test to establish relation between cost and RACE/RACE type

We will do a Kruskal-Wallis test to establish if there is malpractice based on RACE for costing.

Note: Kruskal-Wallis test is to establish if the average cost varies based on RACE type. Where as ETA Square test will establish if the cost and RACE are co-related.

We will see below that cost is not co-related to RACE and also average cost doesn't vary based on RACE type. However, in next section we will see that cost is related to gender but the average cost doesn't vary with gender type.

RACE is a categorical variable here and we will factor it as below:

```
# Copy the dataframe hospitalcosts to a new dataframe hospitalcostsRace
hospitalcostsRace <- hospitalcosts

# Clean data to remove missing record for RACE
hospitalcostsRace <- hospitalcostsRace[!is.na(hospitalcostsRace$RACE),]

# Convert the RACE variable into a category variable
hospitalcostsRace$RACE <- as.factor(hospitalcostsRace$RACE)
#check the number of categories for RACE
str(hospitalcostsRace)
length(levels(hospitalcostsRace$RACE))

group_by(hospitalcostsRace, RACE) %>%
  summarise(NoOfObservations=n()) %>%
  top_n(n=10) %>% arrange(desc(NoOfObservations))
```

Screenshot:

1. No Of levels for RACE is 6 from below:

```
> str(hospitalcostsRace)
Classes 'tbl_df', 'tbl' and 'data.frame':	499 obs. of  6 variables:
 $ AGE   : num  17 17 17 17 17 17 17 16 16 17 ...
 $ FEMALE: num  1 0 1 1 1 0 1 1 1 1 ...
 $ LOS   : num  2 2 7 1 1 0 4 2 1 2 ...
 $ RACE  : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ TOTCHG: num  2660 1689 20060 736 1194 ...
 $ APRDRG: num  560 753 930 758 754 347 754 754 753 758 ...
> length(levels(hospitalcostsRace$RACE))
[1] 6
```

2. No of observations available for different RACES:

```
  RACE  NoOfObservations
  <fct>            <int>
1 1                  484
2 2                    6
3 4                    3
4 5                    3
5 6                    2
6 3                    1
```

We have to find the correlation between the TOTCHG (cost) and APRDRG variables whose nature are follows:

- TOTCHG is the dependent variable which is continuous.
- RACE is the independent variable which is categorical and has more than 2 level/categories (63 levels from above results).

So, eta-square test is more suitable than the pearson correlation test to find correlation(Reference: https://www.researchgate.net/post/Can_I_use_Pearsons_correlation_coefficient_to_know_the_relation_between_perception_and_gender_age_income)

For interpreting the eta square test result, we can use below (Reference : https://www.rdocumentation.org/packages/sjmisc/versions/1.8/topics/eta_sq)

.02 ~ small or negligible relation between dependant and independent variable.

.13 ~ medium relation between dependant and independent variable

.26 ~ large relation between dependant and independent variable

Also, we will be doing a Kruskal-Wallis test rather than one-way ANOVA test, to establish if RACE type has an impact on cost, as the number of observations available are significantly different for different RACE. From above results we see that for RACE1 there are sufficient observations available whereas for some RACE only 1 or 2 observations are available.

Null Hypothesis: H_0 = Average hospital cost is same for all RACE categories i.e average cost doesn't vary according to RACE type.

Alternate Hypothesis: H_1 = Average cost varies with RACE type.

R-CODE:

```
# Kruskal-Wallis Test
kwResult <- kruskal.test(hospitalcostsRace$TOTCHG ~ hospitalcostsRace$RACE, data =
hospitalcostsRace)

# ETA square test
library(sjstats) # for eta square function
anovaResult <- aov(hospitalcostsRace$TOTCHG ~ hospitalcostsRace$RACE, data =
hospitalcostsRace)
eta_sq(anovaResult)
```

Screenshots:

1. Kruskal-Wallis Summary:

```
> kwResult

        Kruskal-Wallis rank sum test

data:  hospitalcostsRace$TOTCHG by hospitalcostsRace$RACE
Kruskal-Wallis chi-squared = 3.2701, df = 5, p-value = 0.6584
```

2. ETA Square Result:

```
> eta_sq(anovaResult)
                    term etasq
1 hospitalcostsRace$RACE 0.002
```

From above Kruskal-Wallis summary results, we notice that the p-value 0.6584 > .05 , thus we fail to reject the null hypothesis at 5% significance level and hence we conclude that RACE type doesn't impact the cost.

Also, from the ETA square results, as the value .002 < .02, thus we can conclude that cost is not related to RACE at all.

2.3.1.2 F-Statistics test via Linear Regression to check if cost and Race are related

As the output is cost and it's a continuous variable, we can use a linear regression model of the form:

$TOTCHG = B_0 + B_1(RACE_1) + B_1(RACE_2) \ldots + B_6(RACE_6) + A_0(AGE_0) + \ldots + A_{17}(AGE_{17}) + F_0(FEMALE_0) + F_1(FEMALE_1) + RESIDUALS$

Following is the null hypothesis for the F-test :

Null Hypothesis: H_0 = The coefficients are 0.

Alternate Hypothesis: H_1 = The co-efficients are not 0.

We will not be using a glm (generalised linear model as the output is not categorical/discrete)

R-CODE:

```r
library(car) #for Anova function
hospitalcostsRace$FEMALE = as.factor(hospitalcostsRace$FEMALE)
hospitalcostsRace$AGE = as.factor(hospitalcostsRace$AGE)
hospitalcostsRace$APRDRG = as.factor(hospitalcostsRace$APRDRG)

RACE_model <- lm(TOTCHG ~ . , data = hospitalcostsRace )

summary(RACE_model)

Anova(RACE_model)
```

Screenshots:

1. F-test results : For space constraint, some of the AGE and APRDRG coefficients has been clipped out in the screenshot.

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

Residuals:
    Min      1Q  Median      3Q     Max 
-5431.1  -199.9   -54.4    91.1  5431.1 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7136.47     967.14   7.379 8.84e-13 ***
AGE17        1444.21     467.46   3.089 0.002141 ** 
FEMALE1      -191.55      74.02  -2.588 0.009998 ** 
LOS           650.96      19.87  32.767  < 2e-16 ***
RACE2         253.31     409.40   0.619 0.536437    
RACE3         630.32     791.17   0.797 0.426086    
RACE4          84.36     426.49   0.198 0.843307    
RACE5        1531.61     833.44   1.838 0.066826 .  
RACE6         -52.82     526.05  -0.100 0.920073    
APRDRG930    1683.07    1000.60   1.682 0.093313 .  
APRDRG952   -4398.64    1117.56  -3.936 9.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 720.4 on 413 degrees of freedom
Multiple R-squared:  0.9716,    Adjusted R-squared:  0.9657 
F-statistic: 166.1 on 85 and 413 DF,  p-value: < 2.2e-16
```

2. ANOVA summary:

```
> Anova(RACE_model)
Note: model has aliased coefficients
      sums of squares computed by model comparison
Anova Table (Type II tests)

Response: TOTCHG
              Sum Sq  Df   F value    Pr(>F)    
AGE         60217402  16    7.2515 6.283e-15 ***
FEMALE       3475803   1    6.6970  0.009998 ** 
LOS        557252369   1 1073.6856 < 2.2e-16 ***
RACE         2310047   5    0.8902  0.487531    
APRDRG    3309860592  61  104.5454 < 2.2e-16 ***
Residuals  214350663 413                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that the RACE co-efficients are 0. For categorical variable, as it will be divided into categories , one of the category will be taken as base category and the F-test result for other RACE co-efficients will be interpreted as below : (Reference- https://www.listendata.com/2016/07/insignificant-levels-of-categorical-variable.html)

The category 1 (or level 1) of 'RACE' variable has been set as reference category and the coefficient of RACE2 means the difference between the coefficient of RACE1 and RACE2. The p-value tells us whether the difference between the coefficient of RACE1 and RACE2 differs from zero. In this case, as the p-value for RACE2 is 0.536 > .05 , hence we can conclude that RACE2 and RACE1 doesn't vary significantly for determining cost. Same hold true for other RACE categories.

Thus we can conclude that there is no malpractice based on RACE.

The ANOVA summary also suggest that RACE is not significant in deciding cost.

Also, we notice from the F-test output, that the adjusted R-squared value is 0.9657. Now, let's remove the RACE variable from the linear model and see if its deteriorates the adjusted R square value.

R-CODE:

```
# Removed the RACE predictor from the regression input
RACE_model1 <- lm(TOTCHG ~ AGE + FEMALE +  LOS + APRDRG, data = hospitalcostsRace
)
summary(RACE_model1)
```

Screenshots:

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

Residuals:
    Min      1Q  Median      3Q     Max 
-5433.7  -204.3   -64.6    91.8  5433.7 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7419.33     857.99   8.647  < 2e-16 ***
AGE1         -561.00     463.37  -1.211 0.226701    
AGE2          201.36     849.86   0.237 0.812821    
AGE17        1415.77     465.26   3.043 0.002491 ** 
FEMALE1      -196.63      73.73  -2.667 0.007953 ** 
LOS           649.65      19.81  32.800  < 2e-16 ***
APRDRG23     4039.60    1040.61   3.882 0.000120 ***

APRDRG952   -4651.75    1040.85  -4.469 1.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 719.9 on 418 degrees of freedom
Multiple R-squared:  0.9713,	Adjusted R-squared:  0.9658 
F-statistic: 176.7 on 80 and 418 DF,  p-value: < 2.2e-16
```

We can see the adjusted R-square didn't deteriorate on removing the RACE variable, rather it improves though by .0001 unit. Thus RACE can be ignored while predicting cost.

2.4 Business Goal 4 : Analyze Severity Of Hospital Cost Based On Age and Gender

To properly utilize the costs, the agency has to analyse the severity of the hospital costs by age and gender for the proper allocation of resources.

2.4.1 Analysis, Coding and Results

2.4.1.1 ANOVA and Pint-Biserial test for Cost vs Gender and F-test

We will perform an ANOVA test to establish the relation between cost and gender type.

Null Hypothesis for ANOVA Test: H_0 = The average cost for Female is same as average cost of Male

Alternate Hypothesis : H_1 = The average cost for Female is different than that of Male.

Also, we will do a Point-Biserial test to establish the correlation between cost and gender. We use Poin-Biserial test as FEMALE is a dichotomous categorical variable (i.e no of categories =2) and cost is a continuos variable.

R-CODE:
```
hospitalcostsGender <- hospitalcosts
hospitalcostsGender$FEMALE <- as.factor(hospitalcostsGender$FEMALE)
str(hospitalcostsGender)
```

```
# Since, FEMALE is a categorical variable, we have to check the dependency of cost on gender as below:
anovaResult <- aov(hospitalcostsGender$TOTCHG ~ hospitalcostsGender$FEMALE, data = hospitalcostsGender)
summary(anovaResult) # print the anova table

hospitalcostsGender$FEMALE <- as.numeric(hospitalcostsGender$FEMALE)

#Pearson correlation test evaluates same to Point-Biserial test for a strict binary.
#https://rpubs.com/juanhklopper/biserial_correlation

cor.test(x = hospitalcostsGender$FEMALE, y = hospitalcostsGender$TOTCHG, method=c("pearson"))
```

Screenshots:

1. ANOVA Summary:

```
> summary(anovaResult) # print the anova table
                            Df    Sum Sq  Mean Sq F value Pr(>F)
hospitalcostsGender$FEMALE   1 2.734e+07 27337922   1.811  0.179
Residuals                  498 7.517e+09 15095177
```

2. Pearson correlation result:

```
        Pearson's product-moment correlation

data:  hospitalcosts$FEMALE and hospitalcosts$TOTCHG
t = -1.3457, df = 498, p-value = 0.179
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14710911  0.02764145
sample estimates:
        cor
-0.06019504
```

From the ANOVA summary results, as p-value .179 > .05, we fail to reject the null hypothesis that the average cost is same for gender type i.e we accept that the cost doesn't vary with gender type.

Also, from the Point-Biserial test we find that the cost and gender have a very low co-relation and is not statistically significant.

We can also do the F-statistic test with a linear regression to verify the adjusted R-Square value.

R-CODE:

```
hospitalcostsGender <- hospitalcosts
hospitalcostsGender$FEMALE <- as.factor(hospitalcostsGender$FEMALE)
FEMALE_model <- lm(TOTCHG ~  FEMALE , data = hospitalcostsGender)
```

```
summary(FEMALE_model)
```

Screenshots:

```
Call:
lm(formula = TOTCHG ~ FEMALE, data = hospitalcosts)

Residuals:
   Min     1Q Median     3Q    Max 
 -2464  -1602  -1221   -207  45842 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3013.9      248.7  12.117   <2e-16 ***
FEMALE        -467.8      347.6  -1.346    0.179    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3885 on 498 degrees of freedom
Multiple R-squared:  0.003623,	Adjusted R-squared:  0.001623 
F-statistic: 1.811 on 1 and 498 DF,  p-value: 0.179
```

From above F-test also we see that the R-Square and adjusted R-Square are too low indicating that gender is insignificant for cost calculations.

2.4.1.2 ETA Square for correlation between Cost and Age groups and F-test

We will do a ETA square test to establish the correlation between cost and age groups. Also we will verify the adjusted R-square for a severity of age group on cost.

R-CODE:

```
hospitalcostsAge <- hospitalcosts
hospitalcostsAge$AGE <- as.factor(hospitalcostsAge$AGE)

anovaResult1 <- aov(hospitalcostsAge$TOTCHG ~ hospitalcostsAge$AGE, data = hospitalcostsAge)

summary(anovaResult1)
eta_sq(anovaResult1)

AGE_model <- lm(TOTCHG ~ AGE , data = hospitalcostsAge )
summary(AGE_model)
```

Screenshots:

1. ETA square result:

```
> eta_sq(anovaResult1)
                    term etasq
1 hospitalcostsAge$AGE 0.117
```

2. Adjusted R-Square from the F-test

```
Call:
lm(formula = TOTCHG ~ AGE, data = hospitalcostsAge)

Residuals:
   Min     1Q Median     3Q    Max 
 -4957  -1113   -797   -122  43789 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2208.9      212.2  10.409  < 2e-16 ***
AGE1          1565.5     1194.8   1.310 0.190716    
AGE2          5089.1     3724.2   1.366 0.172422    
AGE3          7974.5     2157.2   3.697 0.000243 ***
AGE4          5787.1     2637.7   2.194 0.028712 *  
AGE5          7044.6     2637.7   2.671 0.007824 ** 
AGE6          6755.1     2637.7   2.561 0.010741 *  
AGE7          1153.5     2157.2   0.535 0.593090    
AGE8           161.6     2637.7   0.061 0.951159    
AGE9          8364.6     2637.7   3.171 0.001615 ** 
AGE10         3908.4     1871.2   2.089 0.037255 *  
AGE11         -427.6     1331.6  -0.321 0.748258    
AGE12         1451.9      983.2   1.477 0.140397    
AGE13         -479.1      901.7  -0.531 0.595417    
AGE14          376.9      773.3   0.487 0.626244    
AGE15         1644.5      722.3   2.277 0.023244 *  
AGE16          175.6      722.3   0.243 0.808034    
AGE17         2390.5      639.4   3.739 0.000207 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3718 on 482 degrees of freedom
Multiple R-squared:  0.1168,	Adjusted R-squared:  0.08563 
F-statistic: 3.749 on 17 and 482 DF,  p-value: 7.822e-07
```

From above eta square value and the F-test p-value and F-test adjusted R-square value, we can conclude that age impacts price although impact is low.

2.5 Business Goal 5 : Analyze If LOS Can 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.

2.5.1 Analysis, Coding and Results

2.5.1.1 F-test for LOS vs Age, Gender and Race

We will do a linear regression to see the dependency of LOS on Age, Gender and Race.

LOS = $B_0 + A_0(AGE_0) + \ldots + A_{17}(Age_{17}) + F_0(FEMALE_0) + F_1(FEMALE_1) + R_1(RACE_1) + .. + R_6(RACE_6)$

Null Hypothesis : H_0 = The co-efficients are 0. There is no significant dependency of LOS on Age, Gender and Race.

R-CODE:

```
hospitalcostsLOS <- hospitalcosts
hospitalcostsLOS$FEMALE <- as.factor(hospitalcostsLOS$FEMALE)
hospitalcostsLOS$RACE <- as.factor(hospitalcostsLOS$RACE)
str(hospitalcostsLOS)
hospitalcostsLOS$AGE <- as.factor(hospitalcostsLOS$AGE)

LOS_model <- lm(LOS ~ AGE+FEMALE+RACE , data = hospitalcostsLOS )

summary(LOS_model)

Anova(LOS_model)
```

Screenshots:

1. F-test results: Some co-efficients are clipped out for space constraint.

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

Residuals:
   Min     1Q Median     3Q    Max
-3.262 -1.224 -0.892  0.045 37.776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95535    0.24457  12.084   <2e-16 ***
AGE1        -1.20910    1.09842  -1.101   0.2716
AGE2        -0.95535    3.41674  -0.280   0.7799
AGE3         0.28840    1.97773   0.146   0.8841
AGE16       -1.33221    0.68452  -1.946   0.0522 .
AGE17       -0.50059    0.59066  -0.848   0.3971
FEMALE1      0.26877    0.32509   0.827   0.4088
RACE2        0.08552    1.49616   0.057   0.9544
RACE3        0.77589    3.41835   0.227   0.8205
RACE4        0.54007    2.00086   0.270   0.7873
RACE5       -0.95535    1.98274  -0.482   0.6301
RACE6       -0.42362    2.43389  -0.174   0.8619
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.408 on 475 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.02263,   Adjusted R-squared:  -0.0247
F-statistic: 0.4781 on 23 and 475 DF,  p-value: 0.982
```

2. ANOVA of the F-test

```
> Anova(LOS_model)
Anova Table (Type II tests)

Response: LOS
          Sum Sq  Df F value Pr(>F)
AGE        113.3  17  0.5737 0.9115
FEMALE       7.9   1  0.6835 0.4088
RACE         4.5   5  0.0783 0.9955
Residuals 5516.8 475
```

From above results we can conclude that the LOS cannot be predicted from Age, Gender or Race.

2.6 Business Goal 6 : Find the Regressor to predict cost

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.

2.6.1 Analysis, Coding and Results

2.6.1.1 F-test for LOS vs Age, Gender and Race

From above, we have already found that Gender and Race doesn't impact cost. So, we will try to build a linear regression model for cost using LOS, APRDRG and Age variables and see if they are significant using an F-test.

R-CODE:

```
hospitalcostsAll <- hospitalcosts
hospitalcostsAll$FEMALE <- as.factor(hospitalcostsAll$FEMALE)
hospitalcostsAll$APRDRG <- as.factor(hospitalcostsAll$APRDRG)
hospitalcostsAll$AGE <- as.factor(hospitalcostsAll$AGE)
hospitalcostsAll$RACE <- as.factor(hospitalcostsAll$RACE)

All_model <- lm(TOTCHG ~ AGE +  LOS + APRDRG , data = hospitalcostsAll )
summary(All_model)
Anova(All_model)
```

Screenshots:

1. F-test summary:

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

Residuals:
    Min      1Q  Median      3Q     Max 
-5433.2  -233.6   -67.1    94.9  5433.2 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7378.92     863.14   8.549 2.33e-16 ***
AGE1         -461.38     464.71  -0.993 0.321362    
AGE2          352.36     853.20   0.413 0.679825    
AGE17        1494.53     467.19   3.199 0.001484 ** 
LOS           649.91      19.93  32.615  < 2e-16 ***
APRDRG23     4000.72    1046.91   3.821 0.000153 ***
APRDRG952   -4690.36    1047.16  -4.479 9.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 724.4 on 420 degrees of freedom
Multiple R-squared:  0.9708,	Adjusted R-squared:  0.9653 
F-statistic: 176.7 on 79 and 420 DF,  p-value: < 2.2e-16
```

2. ANOVA summary of the F-test

```
> Anova(All_model)
Note: model has aliased coefficients
      sums of squares computed by model comparison
Anova Table (Type II tests)

Response: TOTCHG
              Sum Sq  Df   F value    Pr(>F)    
AGE         58216064  16    6.9341 3.386e-14 ***
LOS        558175427   1 1063.7404 < 2.2e-16 ***
APRDRG    3370534573  61  105.3013 < 2.2e-16 ***
Residuals  220386175 420                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From above, we can conclude that LOS,APRDRG and Age can be used to predict cost and below is a good model to predict the cost:

$TOTCHG = I_0 + A_0(AGE_0) + \ldots + A_{17}(AGE_{17}) + L(LOS) + AP_1(APDRG_1)$

3. Conclusion

3.1 Summary Of Business Goals and Results

SL No	Business Goal	Research Results
1	Find Maximum Hospital Visit And Maximum Spend By AGE Group	The 0 Age group has the maximum hospital visits and they have the maximum spend.
2	Maximum Hospital Visit And Maximum Spend By APRDRG	The 640 APRDRG group has the maximum hospital visits and they have the maximum spend.
3	Analyse If Malpractice In Hospital Cost Based On Race	There is no malpractice in the hospital cost based on Race.

4	Analyse Severity Of Hospital Cost Based On Age and Gender	The impact of Gender on cost is not significant. The impact of Age on cost is significant but low.
5	Analyse If LOS Can Be Predicted From Age, Gender and Race	LOS can’t be predicted from Age, Gender and/or Race.
6	Find the Regressor to predict cost	The variables to predict cost are LOS, APRDRG and Age.