

**Presented by: Ankita Agarwal**

Abstract

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 the healthcare costs and their utilization.

**Analyze the Healthcare cost and Utilization in Wisconsin hospitals**

***Business Analytic Foundation with R Tools- Question***

**Problem 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 the healthcare costs and their utilization.

**Detailed description of the given dataset:**

AGE: Age of the patient discharged

FEMALE: 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

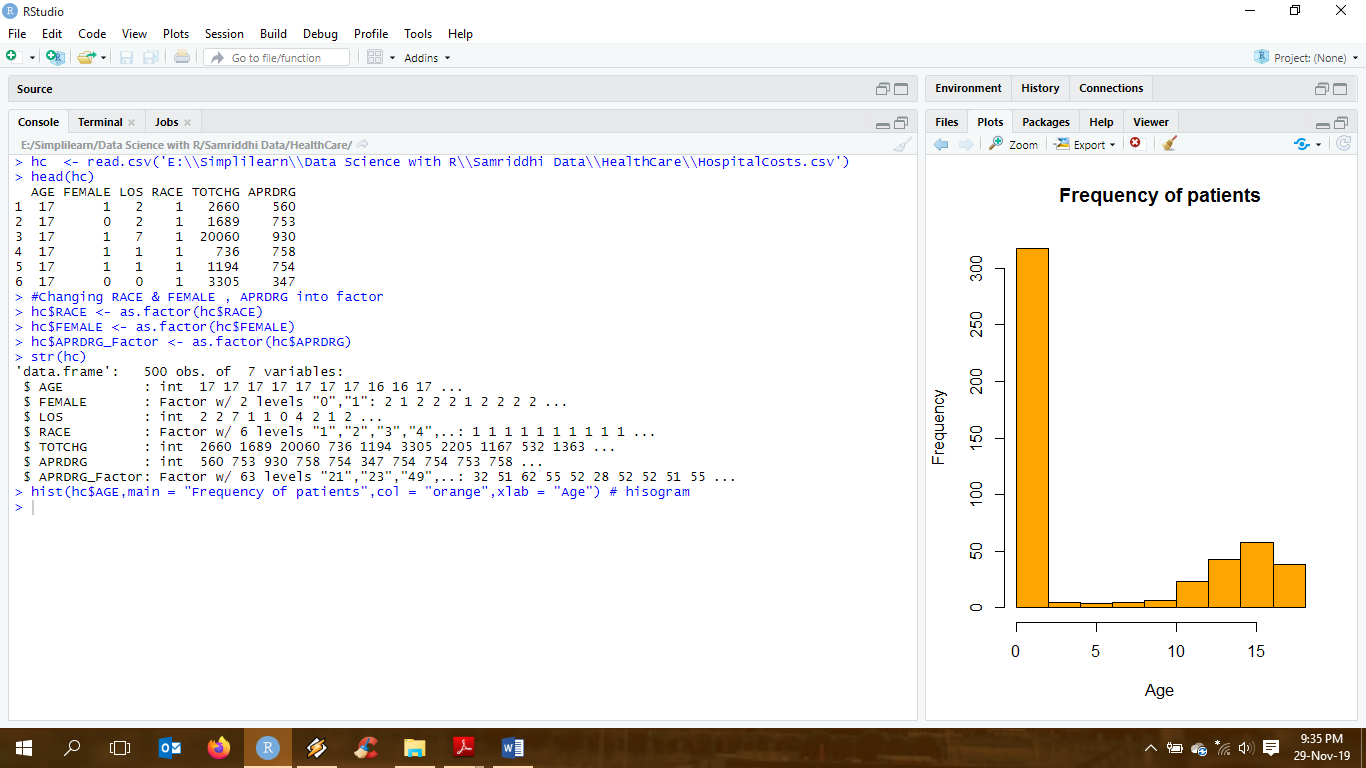
**To Analyze:**

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.
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 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 the hospital costs.

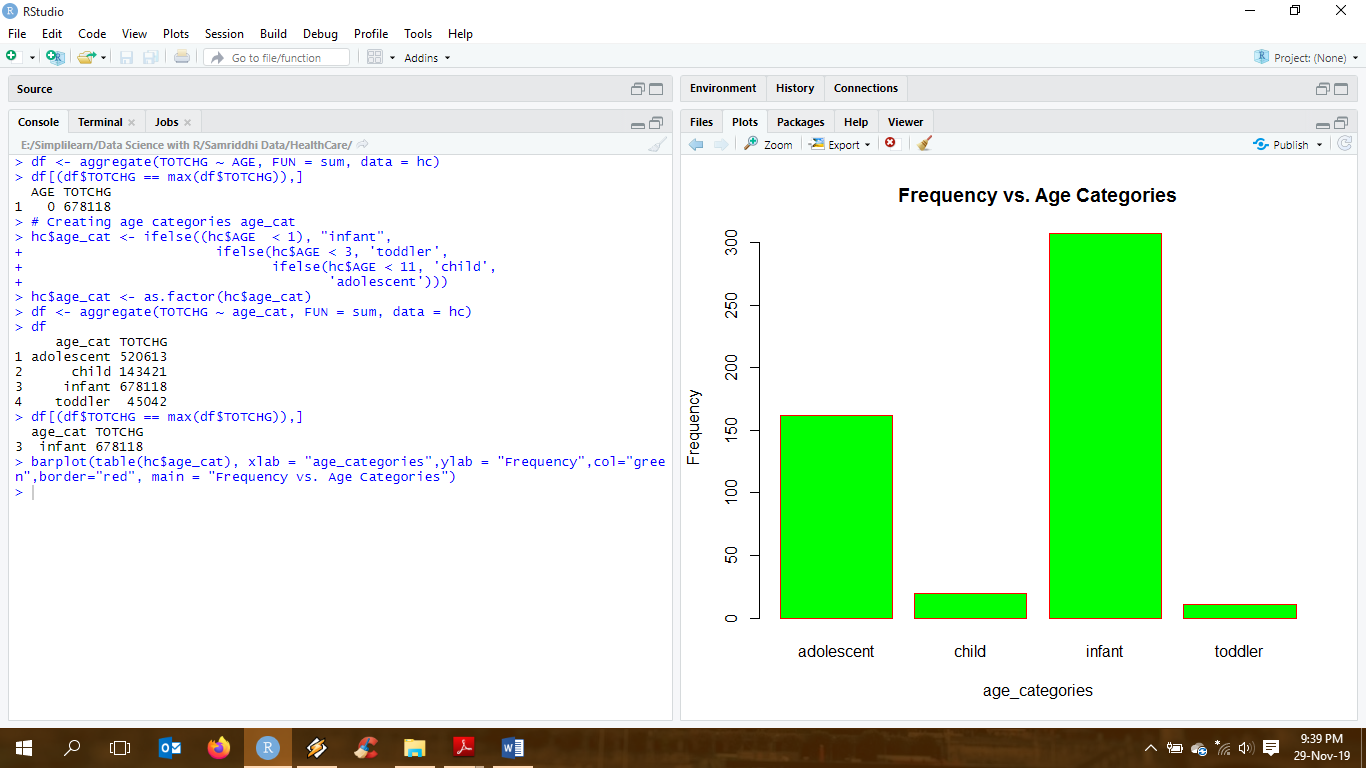
**Analysis and Interpretations:**

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.**

a. To find the age category that has the highest frequency of hospital visit, a histogram can be plotted that would display the number of occurrences of each age category.



b. To find the age category with the maximum expenditure, we use the aggregate function to add the values of total expenditure according to the values of age.

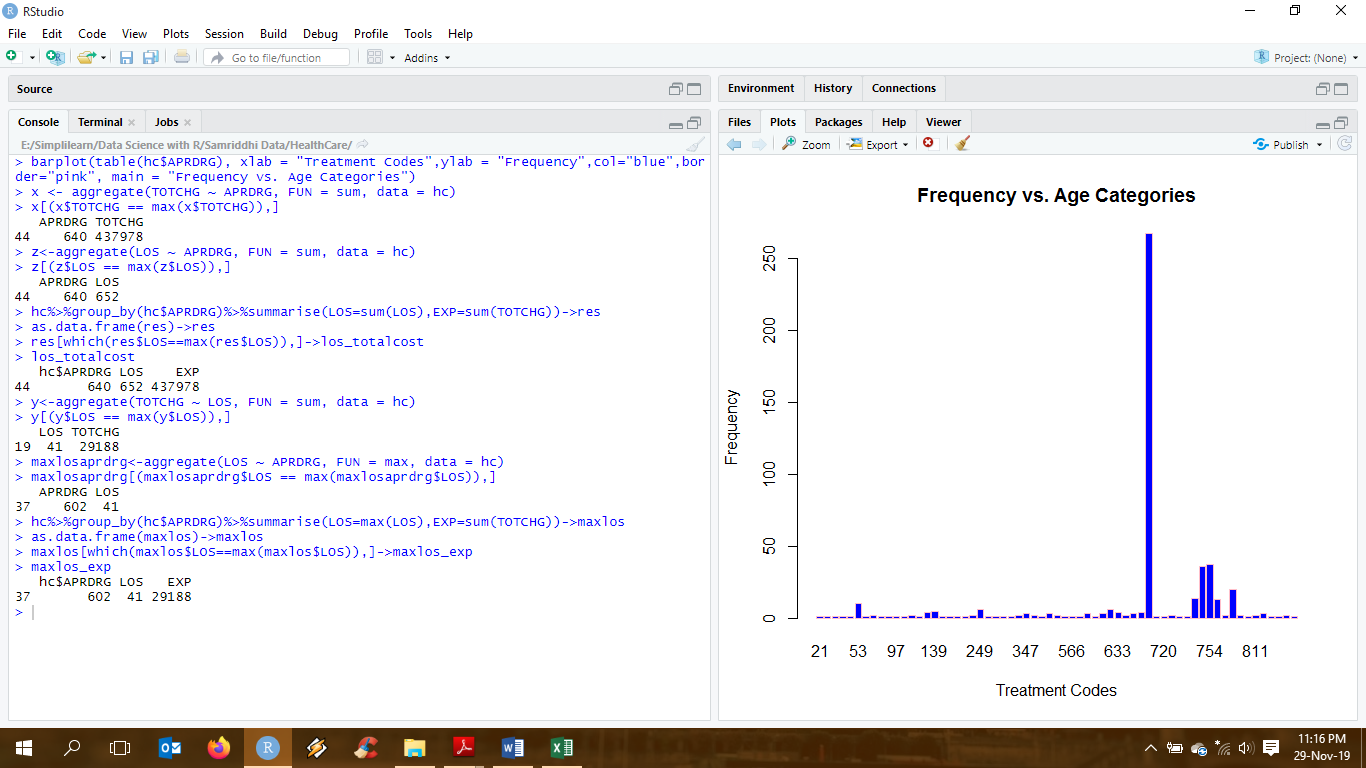


**Interpretation:**

By observing the above table, it is very clear that the people of age '0' frequent the hospital most often and has the most expenditure (678118). By converting age into age categories, we can observe that infants (0-1 years) are more prone to visit the hospital, fetching in most expenditure. Following the infants, adolescents between the ages of 11 to 17 years have high hospitalization costs.

1. **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.**

To find the diagnosis related group that has maximum hospitalization and expenditure, we use the aggregate function or we can use group by to summarize the characteristics of the variables LOS and TOTCRG with respect to APRDRG.



**Interpretation:**

As seen in the graph, maximum people are diagnosed with group 640, inferring that the most expensive treatment by far is for diagnosis group 640 with a total charge of 437978 and length of stay is summed up to 652 days. However, the longest length of stay is 41 days costing 29188 for diagnosis group 602.

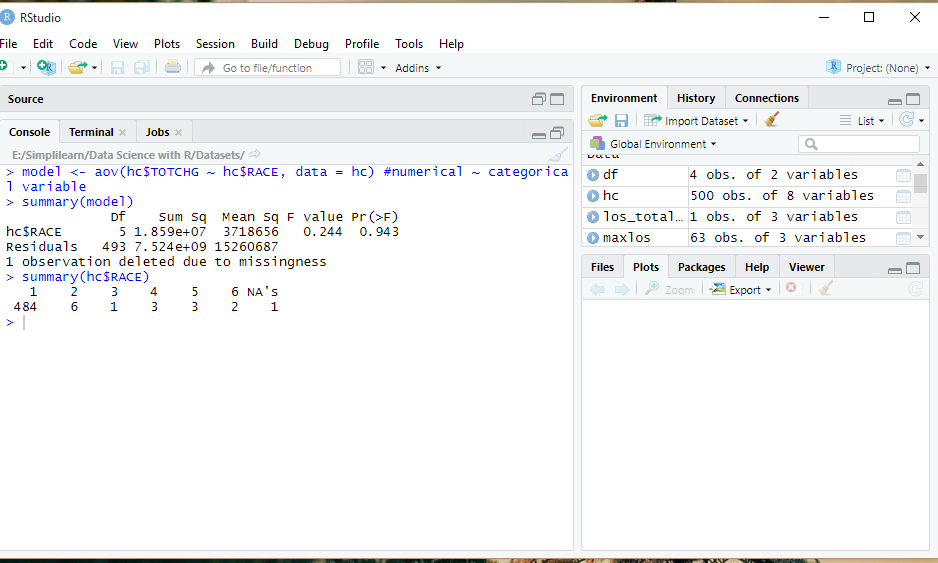
1. **To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.**

To analyze if there is a relation between Race and Total Cost, we need to convert the Race variable to factor (done in the beginning of the project). We now do an ANOVA test on RACE and TOTCHG to verify the impact of Race on Total Costs.

Defining Hypothesis:

Ho: The race has no an impact on the costs.

Ha: The race has an impact on the costs.



**Interpretation:**

As we know, when p-value < alpha, p-value is less than alpha; we reject the null hypothesis. We take alpha value as 0.05 at 95% confidence level.

Here, p-value = 0.943. which means that the p-value is significantly high, emphasizing that there is no relation between the race of patient and the hospital cost. Also, from the summary we notice that, the data has 484 patients of Race 1 out of the 500 entries. This will affect the results of ANOVA as well, since the number of observations is very much skewed.

In conclusion, there is not enough data to verify if the race of patient is related to the hospitalization cost.

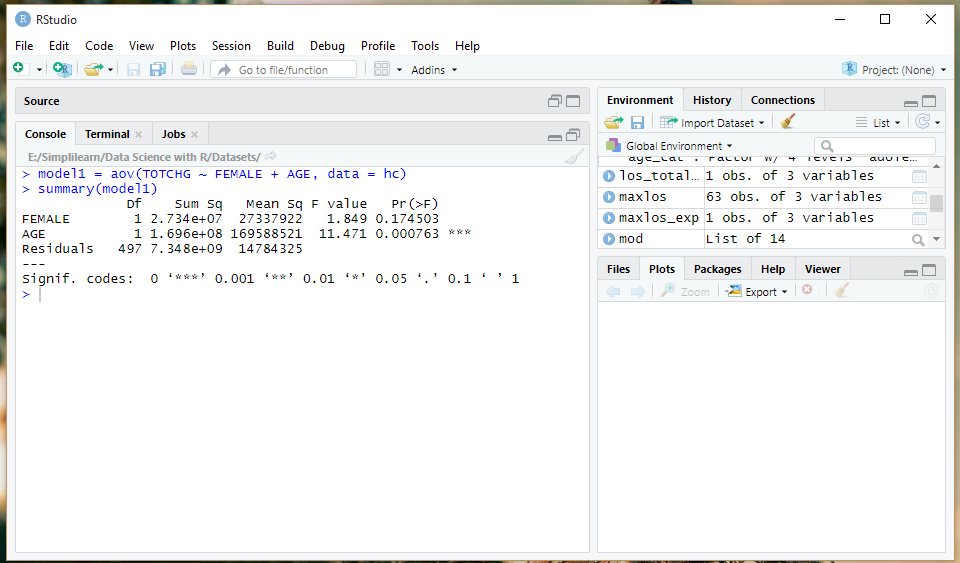
1. **To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.**

To analyze if there is a relation between Gender + Age and Total Cost, we need to convert both FEMALE and AGE variables to factors (done in the beginning of the project). Then, we do an ANOVA test with the following variables: FEMALE + AGE and TOTCHG to verify its impact on Total Costs.

Defining Hypothesis

Ho: The gender and age have no an impact on the total costs.

Ha: The gender and age have an impact on the total costs.



**Interpretation:**

From the above ANOVA model, we observe that p-value of AGE (0.000763) is much lower than 0.05, which suggests that Age is quite significant. Hence, we retain the Null Hypothesis (Ho). However, p-value of FEMALE (0.174503) is higher than 0.05, which suggests that gender has very less impact on the costs. Hence, we cannot retain the Null Hypothesis (Ho).

To conclude, Age does impact the Total Cost of hospitalization whereas Gender does not have much severity towards the Total Cost.

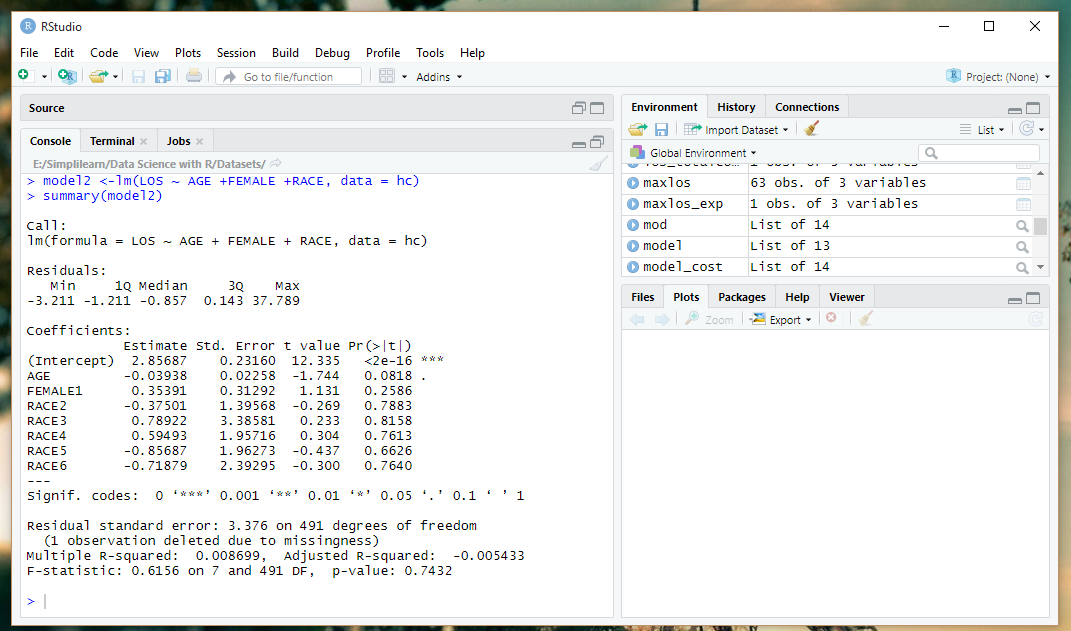
1. **Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.**

To predict if LOS is based on Gender, Age and Race, we use multiple linear regression model to analyze as we have more than one independent variable.

Defining Hypothesis:

Ho: There exists no relation between age, gender and race with LOS

Ha: There exists a relation between age, gender and race with LOS



**Interpretation:**

As we know, p-value < alpha, p-value is less than alpha; we reject the null hypothesis. We take alpha value as 0.05 at 95% confidence level.

We can clearly see that the significance codes are almost null for all the variables, except for the intercept. The p-value (0.7432) is quiet high which signifies that there is no linear relationship between the given variables. Hence, we cannot predict the length of stay of the patients with respect to the age, gender and race.

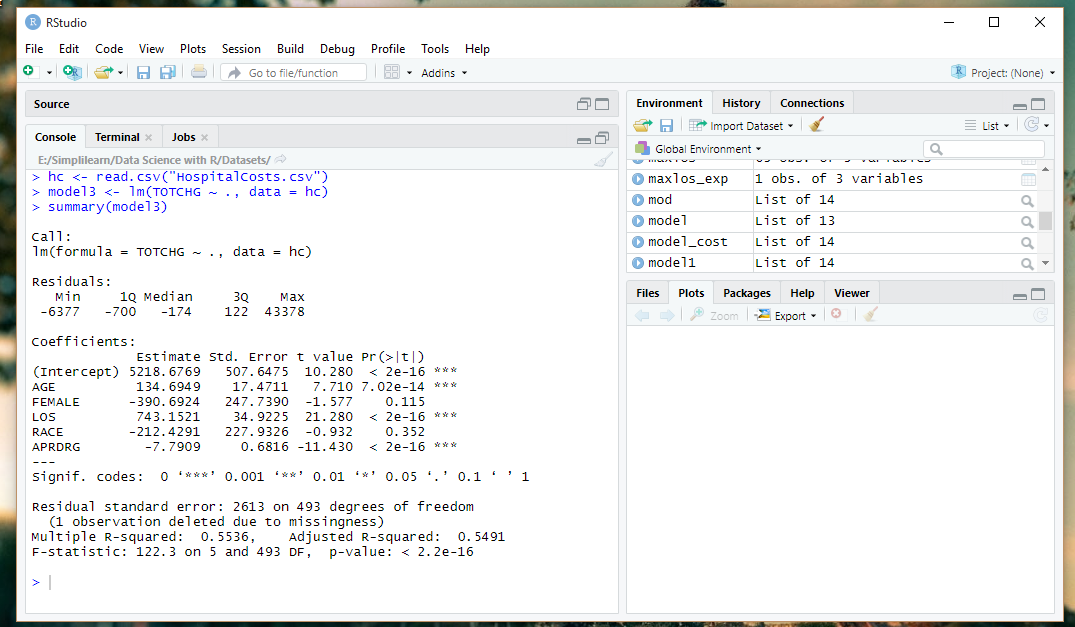
1. **To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.**

To find the variables that mainly affect the total costs, we construct a linear regression model with all the variables.

Defining Hypothesis:

Ho: There is no linear relationship between dependent and independent variables

Ha: There is a linear relationship between dependent and independent variables



**Interpretation:**

Since p-value (2.2e-16) < alpha, there is a linear relationship between the all variables.

Based on the output, the Age and LOS are statistically significant with a very low p-value and affects the total cost. Also, cost is directly proportional to the LOS i.e. higher the LOS, higher is the cost; increase in a day of stay, increase the cost by 743.

All Patient Refined Diagnosis Related Groups with a low p-value, is also affecting the costs. As seen in our prior analysis that APRDRG 640 has high expenditure and highest LOS. So, we can say that people with APRDRG 640 are more likely to be hospitalized and increase the hospitalization costs.

From the significance codes we can see RACE and Gender are clearly not affecting the hospital costs.

**Programming Codes:**

**#Reading Hospital cost data**

library(dplyr)

hc <- read.csv('E:\\Simplilearn\\Data Science with R\\Samriddhi Data\\HealthCare\\HospitalCosts.csv')

head(hc)

#Changing RACE & FEMALE, APRDRG into factor

hc$RACE <- as.factor(hc$RACE)

hc$FEMALE <- as.factor(hc$FEMALE)

hc$APRDRG\_Factor <- as.factor(hc$APRDRG)

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

hist(hc$AGE,main = "Frequency of patients",col = "orange",xlab = "Age") # hisogram

summary(as.factor(hc$AGE)) # summary of age data

df <- aggregate(TOTCHG ~ AGE, FUN = sum, data = hc)

df[(df$TOTCHG == max(df$TOTCHG)),]

# Creating age categories age\_cat

hc$age\_cat <- ifelse((hc$AGE < 1), "infant",

ifelse(hc$AGE < 3, 'toddler',

ifelse(hc$AGE < 11, 'child',

'adolescent')))

hc$age\_cat <- as.factor(hc$age\_cat)

df <- aggregate(TOTCHG ~ age\_cat, FUN = sum, data = hc)

df[(df$TOTCHG == max(df$TOTCHG)),]

barplot(table(hc$age\_cat), xlab = "age\_categories",ylab = "Frequency",col="green",border="red", main = "Frequency vs. Age Categories")

**#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.**

barplot(table(hc$APRDRG), xlab = "Treatment Codes",ylab = "Frequency",col="blue",border="pink", main = "Frequency vs. Age Categories")

x <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = hc)

x[(x$TOTCHG == max(x$TOTCHG)),]

z<-aggregate(LOS ~ APRDRG, FUN = sum, data = hc)

z[(z$LOS == max(z$LOS)),]

#or using this

hc%>%group\_by(hc$APRDRG)%>%summarise(LOS=sum(LOS),EXP=sum(TOTCHG))->res

as.data.frame(res)->res

res[which(res$LOS==max(res$LOS)),]->los\_totalcost

los\_totalcost

y<-aggregate(TOTCHG ~ LOS, FUN = sum, data = hc)

y[(y$LOS == max(y$LOS)),]

s<-aggregate(LOS ~ APRDRG, FUN = max, data = hc)

s[(s$LOS == max(s$LOS)),]

#or using this

hc%>%group\_by(hc$APRDRG)%>%summarise(LOS=max(LOS),EXP=sum(TOTCHG))->maxlos

as.data.frame(maxlos)->maxlos

maxlos[which(maxlos$LOS==max(maxlos$LOS)),]->maxlos\_exp

maxlos\_exp

**#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.**

model <- aov(hc$TOTCHG ~ hc$RACE, data = hc) #numerical ~ categorical variable

summary(model)

summary(hc$RACE)

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

model1 = aov(TOTCHG ~ FEMALE + AGE, data = hc)

summary(model1)

**#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.**

model2 <-lm(LOS ~ AGE +FEMALE +RACE, data = hc)

summary(model2)

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

hc <- read.csv("HospitalCosts.csv")

model3 <- lm(TOTCHG ~ ., data = hc)

summary(model3)

-------------------------------------------------------------------The End-------------------------------------------------------------------