Project: Healthcare cost analysis

DESCRIPTION OF PROBLEM

**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.

**Analysis to be done:**

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 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

**SOLUTION R CODE:**

rm(list=ls())

setwd(choose.dir())

hops <- read.csv('HospitalCosts.csv')

head(hops)

colSums(is.na(hops))

hops <- na.omit(hops)

colSums(is.na(hops))

summary(hops)

# 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

str(hops)

# Race and Female are be categorical variable

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

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

# 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.

# a. To find the category that has the highest frequency of hospital visit

# We can use graphical analysis. A histogram would display the number of

# occurrences of each age category. The as.factor() is called to make sure

# that the categories are not treated as numbers.

# Code:

hist(hops$AGE)

summary(as.factor(hops$AGE))

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

# 306 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38

# Result: From the graph that is displayed, we can see that infants AGE = 0) have the maximum

# frequency of hospital visit, going above 300. The summary of AGE attribute gives

# the numerical output (after converting the age from numeric to factor) - and we

# can see that there are 306 entries for those in the range of 0-1 year.

# b. To find the age category with the maximum expenditure

# we need to add the expenditure for each age, and find the maximum value

# from the sum. We will use the aggregate function to add the values of

# total expenditure according to the values of age.

# Age TotChgAge

library(dplyr)

df1 <- hops %>% group\_by(AGE) %>% summarise(TotchgAge = sum(TOTCHG)) %>% arrange(desc(TotchgAge))

df1

df1[1,]

# AGE TotChgAge

# <int> <int>

# 1 0 676962

# Result: From the result we can see that the infant category (AGE = 0) has maximum hospital costs

# as well (in accordance with the number or frequency of visit). Following the infants,

# 15 and 17 year old individuals have high hospitalization costs.

#####################################################################################

#####################################################################################

# 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.

# Similar to the previous analysis, we can find the diagnosis

# related group with maximum hospitalization and expenditure. For this, we will use the summarise

# and the arrange functions.

#Code:

df2 <- hops %>% group\_by(APRDRG) %>% summarise(AprdrgLos = sum(LOS), AprdrgTotchg = sum(TOTCHG)) %>% arrange(desc(AprdrgLos, AprdrgTotchg))

df2

df2[1,]

# APRDRG AprdrgLos AprdrgTotchg

# <int> <int> <int>

# 1 640 650 436822

# Result: From the result we can see that the Diagnosis Related category (APRDRG = 640) has maximum hospital costs

# As well as hospitalization. Following the 640 Group,

# 753 and 754 Groups have high 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.

# if there is any effect of RACE on TOTCHG

# Then, to verify if the races made an impact on the costs, perform an ANOVA with the

# following variables:

# ANOVA dependent variable: TOTCHG

# Categorical/grouping variable: RACE Missing values: 1 NA value, use na.omit to remove the NA value

# Code:

str(hops$RACE)

str(hops$TOTCHG)

model <- aov(TOTCHG ~ RACE, data = hops) # numerical/int ~ categorical varibale

# dependent variable ~ independent variable

summary(model)

alpha = 0.05

pvalue = 0.943

pvalue < alpha # if this is true = whenever p\_value is less than alpha; we reject the null hypothesis

# Result: We do not reject the null hypothesis. The p-value is very high specifying that there is

# no relation between the race of patient and the hospital cost.

#####################################################################################

#####################################################################################

# 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.

#Code:

model2 <- aov(TOTCHG ~ AGE + FEMALE, data = hops)

summary(model2)

alpha = 0.05

pvalueAge = 0.00323

pvalueGender = 0.03638

pvalueAge < alpha # if this is true = whenever p\_value is less than alpha; we reject the null hypothesis

pvalueGender < alpha # if this is true = whenever p\_value is less than alpha; we reject the null hypothesis

# Result: We reject the null hypothesis. The p-values is very low in case of Age and relatively low in case of Gender,

# specifying that there is a significant relation between the age and gender of patient and the hospital cost.

#####################################################################################

#####################################################################################

# 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.

# Since the length of stay is a continuous variable, we use linear regression to

# predict the variable. Dependent variable: LOS Independent variables: AGE, FEMALE,

# RACE. Note that RACE and FEMALE should be converted into factors, whereas AGE is a

# numerical variable.

# length of stay can be predicted from age, female, and race.

Model3 <- lm(LOS ~ AGE + FEMALE + RACE, data = hops)

summary(Model3)

predict\_los <- predict(Model3, hops)

p\_value = 0.7432

p\_value < alpha # if this is true = whenever p\_value is less than alpha; we reject the null hypothesis

#Result: We can predict the length of stay from age, female and race. But the values from the model

#are very less accurate as P-value of the model is greater than default alpha.

#Hence, there is no relation between LOS and age, gender and race.

# No need to find Rsqd and RMSE. Just check significance of the variables

#####################################################################################

#####################################################################################

# 6. To perform a complete analysis, the agency wants to find the variable that

# mainly affects the hospital costs.

#Code:

model4 <- lm(TOTCHG ~ ., data = hops)

summary(model4)

model5 <- lm(TOTCHG ~ LOS, data = hops)

summary(model5)

model6 <- lm(TOTCHG ~ APRDRG, data = hops)

summary(model6)

#Result: Upon running a linear regression model on the complete data,

# we observe that variables that mainly affacts the hospital costs are - Age, Length of Stay, and

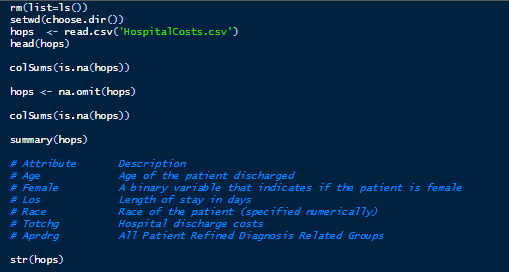
# All Patient Refined Diagnosis Related Groups, on the basis of significance level, i.e. p-value.

# Arranging the p-values in ascending order, We have the main vairiable to be LOS, then APRDRG and finally Age.

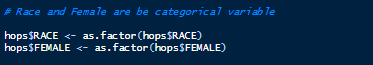
# No need to find Rsqd and RMSE. Just check significance of the variables

**SOLUTION DESCRIPTION AND SCREENSHOTS:**

With simple analysis of the dataset, using head(), str() and is.na(), we found that, there is only 1 row with NA value. Since the number of rows containing NA values is negligible and doesn’t change the statistics of the dataset, we eliminate the particular rows.

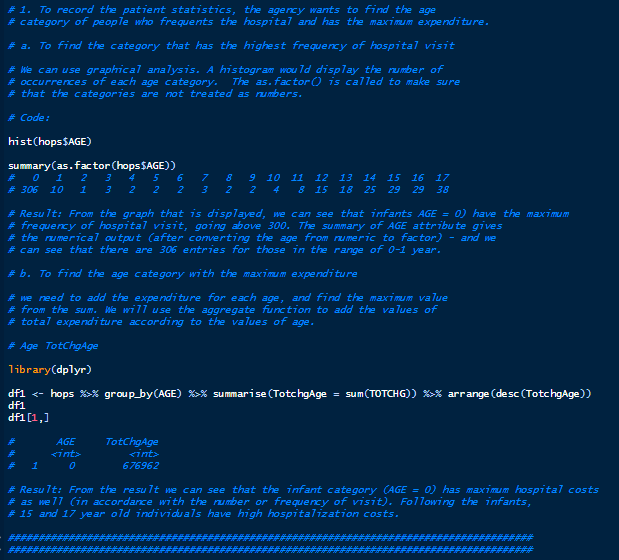


We convert all the categorical variables into factors for statistical purposes and grouping data at such levels.

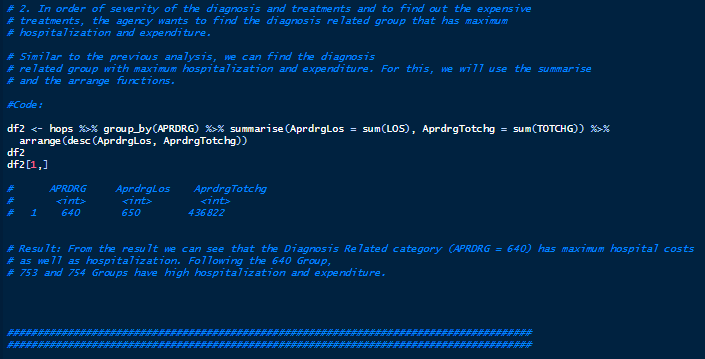


Using dplyr package, summarise() and arrange(), we have successfully answered Question 1 and Question 2. Using ANOVA and Linear Regression models, we came to various relations between the categorical, quantitative and qualitative variables. It is concluded that Length of Stay (LOS) is the most significant factor or variable in the dataset, which affects Hospital Costs and Expenditure.

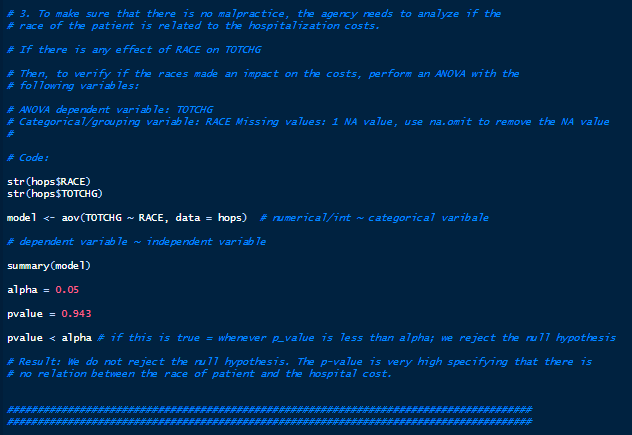
Question and Answer 1:



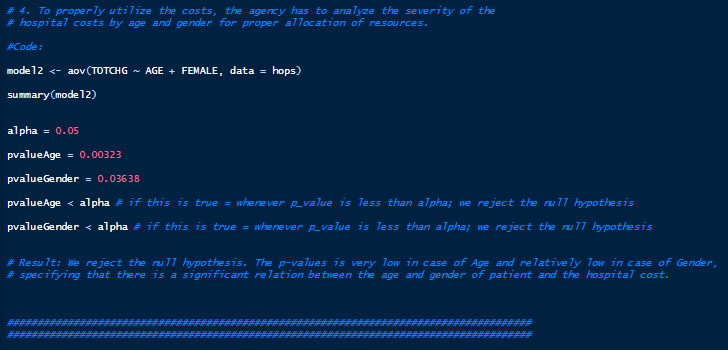
Question and Answer 2:



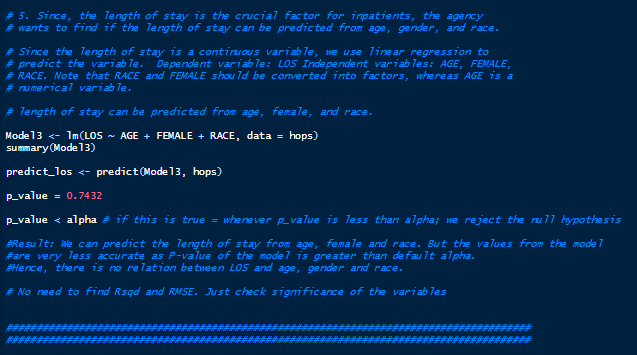
Question and Answer 3:



Question and Answer 4:



Question and Answer 5:



Question and Answer 6:

