Healthcare cost analysis: Source code

```
#Checking working directory
getwd()

#loading the dataset
library(readxl)
hospital_df = read_excel("1555054100_hospitalcosts.xlsx")

#checking nuber of observations and variables
dim(hospital_df)

#Viewing the dataset to understand the data variables,obsersavtions and stractures and summary
View(hospital_df)
str(hospital_df)
summary(hospital_df)

#Initial Observations:
#by looking at the data, we see the variables are:
#Age, Gender, Race of the patients and APRDRG (All Patient Refined Diagnosis Related Groups) are
# categorical/factors varaiables
# LOS -length of stay in days and TOTCHG- Hospital discharge costs are continous varibles

#looking at the summary, we can see that, there are:
#1.Age ranging from 0 to 17 years old
#2.Race types from 1 to 6
#3.we can also see that maximun LOS is 41 days but the mean is 2.828, which tells us that
# records with  41 days LOS cloud be an outlier
#4.we can also see that there is 1 missing value under RACE vaiable.

df_null = subset(hospital_df, is.na(hospital_df$RACE))
df_null
#since it is only one observations, we can drop the record from the dataset

# Dropping the row entry where RACE data is missing
hospital_df = hospital_df[!is.na(hospital_df$RACE),]
dim(hospital_df)
summary(hospital_df)

# we saw that,Age ranging from 0 to 17 years old, we can create a grouping of the age
# creating one new col for Age category

Age_cat =c()
i=1
for (age in hospital_df$AGE){
  if (age <= 1){
    Age_cat = c(Age_cat,"Infants")            #0-1=Infants
  }else if (age <=3){
    Age_cat = c(Age_cat,"Toddlers")           #2-3= Toddlers
  } else if (age <=5){
    Age_cat = c(Age_cat,"PreSchoolers")        #4-5= PreSchoolers
  }else if (age <=11){
    Age_cat = c(Age_cat,"Middle Childhood")     #6-11=Middle Childhood
  }else if (age<=14){
    Age_cat = c(Age_cat,"Young Teens")         #12-14=Young Teens
  }else if (age <=17){
    Age_cat = c(Age_cat,"Teenagers")          #15-17=Teenagers
  }else{
    Age_cat = c(Age_cat,"Adults")
  }
  i = i+1
}

length(Age_cat)

#merging the newly created column with the original datset
hospital_df = cbind(hospital_df,as.factor(Age_cat))
names(hospital_df)[7]='AGE_CAT'
str(hospital_df)
View(hospital_df)

# Data Analysis:
#To analyse the data to research on healthcare costs and their utilization,
#we approach the data based on each question provided


#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.

# we can approcah to check the number of patients and the TOTCHG(Hospital
# discharge costs) by Age category

# Creating a summary table by Age category, Gender, count of gender and TOTCHG
library(dplyr)

#count of patients by Age and gender
ds1=table(hospital_df$AGE_CAT,hospital_df$FEMALE)
ds1=data.frame(ds1)
names(ds1)[1:3]=c('Age_category',"Gender","Count_Gender")
ds1
ds1=mutate(ds1,new_col=paste(ds1$Age_category,ds1$Gender)) #new_col for lookup key
ds1

#sum of TOTCHG by Age and gender
ds2=aggregate(x=hospital_df$TOTCHG, by = list(hospital_df$AGE_CAT,hospital_df$FEMALE), FUN =
sum)
```

```
#ds2
names(ds2)[1:3]=c('Age_category',"Gender","Exp_by_age_cat")
ds2
ds2=mutate(ds2,new_col=paste(ds2$Age_category,ds2$Gender)) #new_col for lookup key
ds2

# merging two data sets using the new_col lookup key
ds3=merge(ds1,ds2,by="new_col",all.x=TRUE)
ds3
View(ds3)
names(ds3)
# drooping duplicate gender col
ds3=ds3[-c(1,5,6)]
ds3
names(ds3)[1:3]=c("Age_category","Gender","Gender_count")
ds3

ds3=ds3[order(-ds3$Exp_by_age_cat),]
ds3

##Visualizing the data
library(ggplot2)
library(grid)
library(gridExtra)
library(scales)
library(tidyverse)

p=ggplot(data = ds3, aes(x=Age_category, y = Gender_count,fill=Gender))+
  geom_bar(stat='identity')+
  geom_text(aes(label=Gender_count), position = position_stack(vjust = 0.5), color="white", size=3)+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF"))



p



q=ggplot(data = ds3, aes(x=Age_category, y = Exp_by_age_cat,fill=Gender))+
  geom_bar(stat='identity')+
  geom_text(aes(label=Exp_by_age_cat/1000), position = position_stack(vjust = 0.5), color="white",
size=3)+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3))

q
```

```
plots= list(p,q)
layout = rbind(c(1,2))
grid.arrange(grobs=plots,layout_matrix=layout,top="Expenditure by Age category and Gender",
          vp=viewport(width=0.8, height = 0.9))



# ans 1:From the above data out/put of ds3 and visualizing the data, we can clearly see that Infants
age = 0 to 1 are
# age category visiting hospital most frequently and therefore has the maximum expenditure =
408356 (M) and 306350(F)


#####################################################################################
##################################
#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.

#checking maximum hospitalization under APRDRG
ds4= table(hospital_df$APRDRG,hospital_df$FEMALE)
ds4 = data.frame(ds4)
names(ds4)[1:3]=c("DRG","Gender","Count_Patients")

# Top 10 count of patients under given APRDRG
head(ds4[order(-ds4$Count_Patients),],10)
#Visualization
a= ggplot(data=ds4, aes(x=(DRG=reorder(DRG,-Count_Patients)),y=Count_Patients, fill=Gender))+
  geom_bar(stat='identity')+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))
a

# we can see that maximum number of hospitalization falls under APRDRG = 640

# Checking the Expenditure for the APRDRG

df_sum_exp_by_DRG = aggregate(x=hospital_df$TOTCHG, by =
list(hospital_df$APRDRG,hospital_df$FEMALE), FUN = sum)
df_sum_exp_by_DRG
names(df_sum_exp_by_DRG)[1:3]=c('DRG','Gender','Exp_by_DRG')

head(df_sum_exp_by_DRG)


#sorting by Exp by Age cat
```

```
df_sum_exp_by_DRG=df_sum_exp_by_DRG[order(-df_sum_exp_by_DRG$Exp_by_DRG),]
df_sum_exp_by_DRG

#Top 10 expensive DRG

head(df_sum_exp_by_DRG,10)

# # we can see that maximum Expenditure falls under APRDRG = 640 which has 254659 for male and
182163 for female

#Expensive Treatment -summary of DRG with maximum Expenditure

df_sum_exp_by_DRG %>% slice_max(df_sum_exp_by_DRG$Exp_by_DRG, n = 2)


#visualization
b= ggplot(data=df_sum_exp_by_DRG, aes(x=(DRG=reorder(DRG,-Exp_by_DRG)),y=Exp_by_DRG,
fill=Gender))+
  geom_bar(stat='identity')+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3))

b

# ans: # # we can see that maximum Expenditure falls under APRDRG = 640 which has 254659 for
male and 182163 for female,
# number of patients under this category is the highest thereby contributing to maximun Total
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.



ggplot(data=hospital_df, aes(x=RACE,y=TOTCHG))+
  geom_bar(stat='identity')+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3))

boxplot(hospital_df$TOTCHG ~ hospital_df$RACE, main = 'RACE', col=c("blue","red"))

#from the barplot, we can see that RACE = 1, has maximun Total Expenditure
# but looking at the box plot, we see most of the outliers are also under RACE 1
# finding correlation
hospital_df
str(hospital_df)

library(corrplot) # for corrplot

corr2=cor(hospital_df[,1:6])

corrplot(corr2, method = "color", outline = T, cl.pos = 'n', rect.col = "black",  tl.col = "indianred4",
addCoef.col = "black", number.digits = 2,
        number.cex = 0.60, tl.cex = 0.7, cl.cex = 1, col =
colorRampPalette(c("green4","white","red"))(100))
# ans:
# the correlation between patient race and hospitalization cost  = -0.02 which shows a very weak
# correlation between RACE and hospitalization costs.

# #checking linear relationship

dataset1=data.frame(lapply(hospital_df, as.numeric))
str(dataset1)

race = lm(formula = TOTCHG ~RACE , data = dataset1)
summary(race)

# p value = 68% is very high,
# ans: so even though the barplot, has maximun total expendiure under RACE1,
# the, boxplot, the correlation and the linear relation check, shows that here 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 the proper allocation of resources.

ds3

178+138

178/316
138/316

# looking at the data we, see most of the patients are under Infants category
# there are 316 (178 Male Infants and 138 Female infants) infants
```

```
#combo chart (col and line chart)

ggplot(ds3) +
  geom_col(aes(x = Age_category, y = Exp_by_age_cat,fill=Gender), size = 1, color = "darkblue") +
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF"))+
  geom_line(aes(x = Age_category, y = 4000*Gender_count,
color=Gender,group=Gender),stat='identity',size=1)+
  scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3),sec.axis = sec_axis(~./4000, name
= "Gender_count"))+
  theme_minimal()

#looking at the number of patients by gender and age, we see that most of the patients are under
Infants category.
# within Infants category, we see 56 % (178/316) male infants and 44% (138/316) female infants.

#checking linear relationship
str(dataset1)

age_gender = lm(formula = TOTCHG ~AGE+FEMALE , data = dataset1)
summary(age_gender)

# Age has a very less P value which means that it has hig sifnificance in the hospitalization costs
# gender also has less p value and we can say that it is also has some signifiance in the hospitlization
costs

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

#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.

# dependent variable = LOS
# independent variable = AGE,FEMALE,RACE

head(dataset1)
str(dataset1)

attach(dataset1)
par(mfrow=c(2,3)) # setting the graph area to add multiple plots
plot(AGE,LOS, main="AGE vs LOS")
plot(FEMALE,LOS, main="Gender vs LOS")
plot(RACE,LOS, main="RACE vs LOS")
boxplot(dataset1$LOS ~ dataset1$AGE, main = 'AGE', col=c("blue","red"))
boxplot(dataset1$LOS ~ dataset1$FEMALE, main = 'Gender', col=c("blue","red"))
boxplot(dataset1$LOS ~ dataset1$RACE, main = 'RACE', col=c("blue","red"))

par(mfrow=c(1,1)) # resetting the grap area to plot only single plot
```

```
#LOS is spread across diffrent across AGE and Gender but mostly concentrated to RACE 1 for RACE

# Dropping variables that are not needed for prediction.

names(dataset1)

dataset2 = subset(dataset1, select = -c(TOTCHG,APRDRG,AGE_CAT))

names(dataset2)

str(dataset2)

#checking linear relationship
model = lm(formula = LOS ~. , data = dataset2)
summary(model)

# p value is very high, so it occurs there is no linear relationship between the given variables

# but we can try to improve the model

#1:
#Cecking LOS Outliers

boxplot(dataset2$LOS)

# #Outlier Treatment: using the IQR for Upper cap and Floor LOS

bench_upper = 3.00 +1.5*IQR(dataset2$LOS)
bench_upper

bench_lower = 2.00 - 1.5*IQR(dataset2$LOS)
bench_lower

# Creating a new col for LOS_treated
LOS_treated =c()
i=1
for (los in dataset2$LOS){
  if(los > bench_upper){
    LOS_treated =c(LOS_treated,bench_upper)
  }else if (los < bench_lower){
    LOS_treated =c(LOS_treated,bench_lower)

  }else{
    LOS_treated =c(LOS_treated,los)
  }
}
```

```
length(LOS_treated)
dataSet3 = cbind(dataset2,LOS_treated)
summary(dataSet3)
names(dataSet3)
boxplot(dataSet3$LOS_treated)

#checking linear relationship
model = lm(formula = LOS_treated ~AGE+FEMALE+RACE , data = dataSet3)
summary(model)

# P value of RACE is very high, so we can say that there is no linear relationship between LOS and RACE

# we can try the model removing the RACE variable
model = lm(formula = LOS_treated ~AGE+FEMALE , data = dataSet3)
summary(model)

# by looking at he P value has improved after removing RACE variable but we see a very small R squared and Adjusted R squared,
# almost close to zero, it again signifies that there is no linear relationship between LOS, AGE,FEMALE variables
# we cannot build a model to predict the LOS using the AGE and Gender of the patients


###################################################################################################################
#6. To perform a complete analysis, the agency wants to find the variable that mainly affects hospital costs.
library(caret)

str(dataset1)

##checking linear relationship
check_variable = lm(formula = TOTCHG ~AGE_CAT+FEMALE+LOS+APRDRG , data = dataset1)
summary(check_variable)

# p value of FEMALE is 14%, so we can say that Gender of a patients does not affect the hospital costs.
# but we can see that AGE, LOS and APRDRG affects the total hospital costs

# validating using varImp
varImp(check_variable)
#We can see that, LOS, APRDRG and AGE affect the total hospital cost

#########################################################################################################
```