R CODE FOR HEALTH CARE COST ANALYSIS

```
#Author: U. SAI RAVI TEJA

getwd()

setwd(choose.dir())

library(readxl)

library(dplyr)

dataset=read_xlsx(path = "D:/Simplilearn/DS with R/Healthcare
Project2/HospitalCosts.xlsx")


#check for missing values

check=(is.na(dataset))

#Used to find the column of the missing value

colnames(dataset)[colSums(is.na(dataset)) > 0]

which(check[,4]==TRUE)

dataset$RACE[277]

#replace missing data

dataset$RACE=ifelse(is.na(dataset$RACE),
                    ave(dataset$RACE,FUN = function(x) round(mean(x,na.rm = TRUE))),
                    dataset$RACE)

#Categorise the Age

Age_Categorised<- transform(dataset, Age_Category = ifelse(AGE<=1, 'Infant',
                                    ifelse(AGE<=4, 'Todler',
                                    ifelse(AGE<13, 'Children', 'Teenagers'))))

df=as.data.frame(Age_Categorised %>% group_by(Age_Category)
                 %>% summarise(Hospitalcosts= sum(TOTCHG),
                FrequentVisit=round(mean(LOS)))%>%
                arrange(desc(Hospitalcosts)))


#Age category of the people who frequent visit the hospital and has max expenditure
```

```
sol1=df[1,]

#Diagnosis related group that has max hospitalization and max expenditure
df1=as.data.frame(Age_Categorised %>% group_by(APRDRG)
            %>% summarise(Hospitalcosts= sum(TOTCHG),
                    Stay=sum(LOS))%>%
             arrange(desc(Hospitalcosts,FrequentVist)))
sol2=df1[1,]

#check if race is normally distributed
qqnorm(dataset$RACE)
qqline(dataset$RACE,col="red")

#Answer: No

#check through hist plot
hist(dataset$RACE,breaks=3,col="green")

#Answer: No

#check if Totchg is normally distributed
qqnorm(dataset$TOTCHG)
qqline(dataset$TOTCHG,col="blue")

#Answer: No

#Is race of the patient related to Hospitalization costs
#Dependent variable: hospitalization costs
#Independent variable: Race
```

```
#Perform Anova
race=as.factor(dataset$RACE)
anova_sol=aov(TOTCHG~RACE,data = dataset)
summary(anova_sol)

#Sol: Race is related to hospitaliztion costs as P>o.o5 significance level

#check if age is normally distributed
qqnorm(dataset$AGE)
qqline(dataset$AGE,col="red")

#Answer: NO
#Not preferring 2-way ANOVA Because there are no two factor variables. only one FACTOR which is RACE
library(caTools)
# df=dataset[,c(1,2,5)]
# df[,1:3]=scale(df[,1:3])
check_relation1=lm(TOTCHG~FEMALE+AGE,data=dataset)
summary(check_relation1)
#Both GENDER AND AGE are statistically significant to TOTCHG
#The lower the pvalue the high the impact of independent on var on dependent



#LOS Can be predicted by AGE,GENDER,RACE
library(caTools)
split=sample.split(dataset$LOS,SplitRatio = 0.8)
train_set=subset(dataset,split==T)
test_set=subset(dataset,split==F)
model=lm(LOS~AGE+FEMALE+RACE,data=train_set)
```

```
summary(model)
#FEMALE AND RACE DONT HAVE MUCH IMPACT ON LOS while AGE has little impact
pred=predict(model,newdata = test_set)
round(pred)



#Finding the var the affects the hospital costs
check_relation2=lm(TOTCHG~.,data=dataset)
summary(check_relation2)

#FROM the summary it seems that AGE,LOS AND APRDRG mainly affects the hospital costs
```