```
setwd('//home//labsuser')
library("readxl")
hosp=read_excel('Hospital costs.xlsx')

# 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
head(hosp)
max(hosp$AGE)    #-> Max. age is 17

# Finding which age category has maximum expenditure
max(hosp$TOTCHG)
which(hosp$TOTCHG==48388) # -> gives the index of given element
hosp[333,5]

library('dplyr')
filter(hosp,hosp$TOTCHG==48388) # => Age category

# Finding which age category is frequently visiting hospital
table(hosp$AGE)  # => people with age 0 (i.e., infants) are frequently
visiting the hospital
hist(hosp$AGE, main="Histogram of Age Group and their hospital visits",
     xlab="Age group", border="black", col=c("grey", "violet"),
xlim=c(0,20), ylim=c(0,350))

infants<-filter(hosp,hosp$AGE==0)
max(infants$TOTCHG)

# INSIGHTS FOR AGE CATEGORY:
# Frequently visited: Age 0 (29188)
# Maximum Expenditure : Age 17 (48388)

# 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.
Age_1<-filter(hosp,hosp$AGE==1)
max(Age_1$TOTCHG)
Age_2<-filter(hosp,hosp$AGE==2)
max(Age_2$TOTCHG)
Age_3<-filter(hosp,hosp$AGE==3)
max(Age_3$TOTCHG)
Age_4<-filter(hosp,hosp$AGE==4)
max(Age_4$TOTCHG)
Age_5<-filter(hosp,hosp$AGE==5)
max(Age_5$TOTCHG)
Age_6<-filter(hosp,hosp$AGE==6)
max(Age_6$TOTCHG)
Age_7<-filter(hosp,hosp$AGE==7)
max(Age_7$TOTCHG)
Age_8<-filter(hosp,hosp$AGE==8)
max(Age_8$TOTCHG)
Age_9<-filter(hosp,hosp$AGE==9)
max(Age_9$TOTCHG)
Age_10<-filter(hosp,hosp$AGE==10)
max(Age_10$TOTCHG)
Age_11<-filter(hosp,hosp$AGE==11)
max(Age_11$TOTCHG)
Age_12<-filter(hosp,hosp$AGE==12)
max(Age_12$TOTCHG)
Age_13<-filter(hosp,hosp$AGE==13)
```

```
max(Age_13$TOTCHG)
Age_14<-filter(hosp,hosp$AGE==14)
max(Age_14$TOTCHG)
Age_15<-filter(hosp,hosp$AGE==15)
max(Age_15$TOTCHG)
Age_16<-filter(hosp,hosp$AGE==16)
max(Age_16$TOTCHG)

# Creating table with frequency visits and corresponding maximum
expenditure
Age<-c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)
max_expend<-
c(29188,9606,7298,14243,9230,10584,9530,6425,3588,10585,17524,3908,17434,
5615,10756,20195,10002,48388)
Freq_visits<-c(307,10,1,3,2,2,2,3,2,2,4,8,15,18,25,29,29,38)
log<-data.frame(Age,max_expend,Freq_visits)
log

# Finding the diagnosis related group which has high expenditure and
hospitalization
filter(hosp,hosp$TOTCHG==48388) # 911 (high expenditure)
max(hosp$LOS)
filter(hosp,hosp$LOS==41) # 602 (maximum hospitalization)

# INSIGHTS:
# APRDRG -> All Patient Refined Diagnosis Related Groups
# In the given data, diagnosis related group of 911 has high expenditure.
So they have expensive treatments.
# Diagnosis-related group of 602 has maximum hospitalization of 41 days.

# 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.
# Analyzing if the race of the patient is related to the hospitalization
table(hosp$RACE)

raceInfluence=lm(TOTCHG~ RACE, data=hosp)
summary(raceInfluence)
# RACE has no significance on hospitalization cost
# p-value=0.6856. i.e., p>0.05, it is highly confirming that race has no
relation with cost
# To have significance, p-alue should be less than 0.05 and F-value
should be greater than 0.05

ANO=aov(RACE~TOTCHG,data=hosp)
summary(ANO)
# Here, p-value= 0.686 and F-value= 0.164. Again here there is no
relationship between race and hospitalization cost
# F-value<0.05, it clearly confirming the above statement.
# So, the race of the patient is not related to hospitalization.
Therefore, there is no malpractice.

# 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.
# FEMALE
female<-filter(hosp, hosp$FEMALE==1)
plot(female$AGE,female$TOTCHG,pch=10)
# Upto 10000 costs for treatment for most of the females of ages 0,1,10-
17
```

```
# There's no costs for the females of ages 6-9. i.e., much resource
allocation is not needed for this age group
# For only very few females (9), the cost goes above 10000
# For female, the maximum expenditure is around 50000

# MALE
male<-filter(hosp, hosp$FEMALE==0)
plot(male$AGE,male$TOTCHG,pch=10)
# Here, cost falls around 5000-10000 for most of the males
# For male the maximum expenditure is around 25000

gen=aov(TOTCHG~AGE+FEMALE,data=hosp)
summary(gen)
# Here, p-value is less than 0.05, so it has statistical significance

genc=lm(TOTCHG~AGE+FEMALE,data=hosp)
summary(genc)
# Age has more influence on cost than gender

# 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.
rel_LOS = lm(LOS~ AGE + FEMALE + RACE, data = hosp)
summary(rel_LOS)
# Age, gender and race has no influence on length of stay as p-
value=0.2692 > 0.05

# To perform a complete analysis, the agency wants to find the variable
that mainly affects hospital costs.
names(hosp)
total_rel=lm(TOTCHG ~ AGE + FEMALE + LOS + RACE + APRDRG, data= hosp)
summary(total_rel)
# Variables that mainly affects hospital costs are 'AGE','LOS', and
'APRDRG'.
# LOS -> Length of stay
# APRDG -> All Patient Refined Diagnosis Related Groups
```