```
## ----setup, include=FALSE----------------------------------------------------
knitr::opts_chunk$set(echo = TRUE)


## ----Hospital----------------------------------------------------------------
hospital_data<-read.csv(file.choose(),header = T)
names(hospital_data) #Get the columns
head(hospital_data) #First 6 rows
str(hospital_data) #Structure, datatype of each column


## ----Task 1------------------------------------------------------------------
#hospital_data$AGE<-as.factor(hospital_data$AGE)
levels(as.factor(hospital_data$AGE))
table(as.factor(hospital_data$AGE)) #here AGE 0 has max number, getting an idea
summary(hospital_data) #Summary of each variable in the dataset

#Now the max cost analysis, we need the aggregated values of TOTCHRG based on AGE, we can use aggregate fun
ction
cost_aggregate<-aggregate(TOTCHG~AGE,data = hospital_data,FUN = sum)
cost_aggregate
max(cost_aggregate) #678118
cost_aggregate[which.max(cost_aggregate$TOTCHG),] #Tells AGE=0 as maximum entry in aggregate


## ----Task 1 plot-------------------------------------------------------------
hist(hospital_data$AGE,breaks = nlevels(as.factor(hospital_data$AGE)),xlab = "Age",ylab = "Total Records",col =
"green",freq = T,density = 100,border = 4)
barplot(table(hospital_data$LOS,hospital_data$AGE),xlab = "Age",ylab = "Total rows, Days Stayed stacked",col =
"green",density = 100,border = 4)



## ----Task 2------------------------------------------------------------------
summary(as.factor(hospital_data$APRDRG)) #we can see code 640 has 267 records out of 500
cost_diag_aggregate<-aggregate(TOTCHG~APRDRG,data = hospital_data,FUN = sum)
cost_diag_aggregate
max(cost_diag_aggregate) #678118
cost_diag_aggregate[which.max(cost_diag_aggregate$TOTCHG),]


## ----Task 2 Plot-------------------------------------------------------------
hist(hospital_data$APRDRG,breaks = nlevels(as.factor(hospital_data$APRDRG)),xlab = "Diagnosis Code",ylab = "
Sum of expenses",col = "green",freq = T,density = 100,border = 4)
#Histo chart shows the same results,
#Barplot won't be practical here, as it would take a lot of space to draw x bars. As can be seen with unique number o
f values of diagnosis codes
nlevels(as.factor(hospital_data$APRDRG))
#or
unique(hospital_data$APRDRG)
#We have 63 unique codes
```

```
## ----Task 3----------------------------------------------------------------
anyNA(hospital_data) #this tells TRUE
anyNA(hospital_data$RACE) #RACE has some NA entries
summary(hospital_data$RACE) # 1 NA value, we can omit this, as we have no way to predict this as we do for num
erical variables
hospital_data<-na.omit(hospital_data)
anyNA(hospital_data) #Now we have no NA values
#ready to test ANOVA now
model_race_vost_aov<-aov(hospital_data$TOTCHG~hospital_data$RACE)
summary(model_race_vost_aov)


## ----Task 4----------------------------------------------------------------
hospital_data$FEMALE<-as.factor(hospital_data$FEMALE)
summary(hospital_data$FEMALE) #almost similar number of MALES and FEMALES
fit_1<-lm(formula =  TOTCHG~AGE+FEMALE,data = hospital_data)
fit_1
summary(fit_1)


## ----Task 4 Plot-----------------------------------------------------------
library(ggplot2)
ggplot(hospital_data,aes(y=TOTCHG,x=AGE,color=factor(FEMALE)))+geom_point()+stat_smooth(method="lm",
se=FALSE)
#The plot clearly defines our regression model, female have lesser costs than males,
#and More the age, more the costs


## ----Task 5----------------------------------------------------------------
hospital_data$RACE<-as.factor(hospital_data$RACE)
hospital_data<-na.omit(hospital_data)
fit_los<-lm(data = hospital_data,formula = LOS~AGE+FEMALE+RACE)
fit_los
summary(fit_los)


## ----Task 6----------------------------------------------------------------
fit_costs_all<-lm(data = hospital_data,formula = TOTCHG~AGE+FEMALE+RACE+APRDRG+LOS)
fit_costs_all
summary(fit_costs_all)
```