**1. Import the Titanic Dataset from the link Titanic Data Set.**

**Perform the following:**

**a. Preprocess the passenger names to come up with a list of titles that represent families**

***Ans:***

setwd("C:/Users/hemakumar/Downloads")

data1<-read\_excel("titanic3.xls")

Rpt<-data.frame(sapply(strsplit(data1$name, " "),`[`,1))

h<-data.frame(table(Rpt))

colnames(h)<-c("family","no-of-people")

#removing the persons alone

fam<-h[!h$`no-of-people` %in% h$`no-of-people`[h$`no-of-people`==1],]

n<-fam$`no-of-people`

famnames<-fam$family

barplot(n,names.arg = famnames,xlab = 'familynames',ylab='num\_of\_people',

col="blue",main="families")

**b. Represent the proportion of people survived from the family size using a graph.**

***Ans:***

###dataframe---1####

>setwd("C:/Users/hemakumar/Downloads")

>data1<-read\_excel("titanic3.xls")

>Rpt<-data.frame(sapply(strsplit(data1$name, " "),`[`,1))

>h<-data.frame(table(Rpt))

>colnames(h)<-c("family","no-of-people")

>head(h,100)

####dataframe--2##

g<-data.frame(data1$survived,sapply(strsplit(data1$name, " "),`[`,1))

r<-data.frame(table(g))

colnames(r)<-c("survied","family","total-num-in-family-survied")

r <- r[c("family", "total-num-in-family-survied", "survied")]

a<-r[!r$survied %in% r$survied[r$survied=="0"],]

a$survied<-NULL

head(a,100)

j<- merge(h,a,by="family")

head(j,100)

##removing people alone

fam<-j[!j$`no-of-people` %in% j$`no-of-people`[j$`no-of-people`==1],]

colors = c("brown","green")

family=fam$family

status=c('dead','survied')

mymat<-data.frame(fam$`no-of-people`,fam$`total-num-in-family-survied`)

barplot(t(as.matrix(mymat)),main = "TITANIC", names.arg = family,

xlab = "names", ylab = "familynums", col = colors)

legend("topleft", status, cex = 1.3, fill = colors)

**c. Impute the missing values in Age variable using Mice Library, create two different**

**graphs showing Age distribution before and after imputation.**

***Ans:***

library(mice)

library(VIM)

md.pattern(data1)

imp<-mice(data1)

imp$imp$age

head(data1$age,30)

head(complete(imp),30)

data2<-complete(imp)

head(data2$age,30)

a=sum(!is.na(data1$age))

b=sum(is.na(data1$age))

c=sum(!is.na(data2$age))

d=sum(is.na(data2$age))

df=matrix(c(a,b,c,d),nrow = 2)

colors = c("green","orange")

imps <- c("before","after")

values<- c("values","Null values")

barplot(df, main = "Before and After imputation", names.arg =imps, xlab = "difference",

ylab = "total values",xlim=c(-3,5),ylim =c(0,2000) ,col = colors)

legend("topleft", values, cex = 1.3, fill = colors)