Problem Statement

1. Import the Titanic Dataset from the following link:

https://drive.google.com/file/d/1JTJCjdGuUxzKXYlwOavwovB01k6FWg3r/view?ts=5b42ea10

Perform the below operations:

a. Pre-process the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.

Answer :

```
head(Titanic3)
tail(Titanic3)
str(Titanic3$name) # check structure, as only charecter vectors can be split using strsplit function
Titanic3$name<-as.character(Titanic3$name)
str(Titanic3$name)
#telling R to call rbind, on two charecters split by strsplit.
#in strsplit, as the data has many " ", and all breaks in many pieces
# hence, using sub() {and not gsub()}, which replaces only first pattern
# so, sub changes first space in ; and the strsplit splits along ; and then rbind binds along colums, which is called by do.call
namessplit<-do.call(rbind,strsplit(sub(" ",";",Titanic3$name),";"))

head(namessplit)
#converting the charecters to data frame and naming the columns
namessplit<-data.frame(namessplit)
names(namessplit)<-c("family_name", "name")
head(namessplit)
str(namessplit)

#getting title separated from first name
Title<-do.call(rbind,strsplit(sub(" ",";",namessplit$name),";"))
head(Title)
Title<-data.frame(Title)
names(Title)<-c("title", "first_name")
head(Title)
str(Title)
head(Title)
```

```
#merging the rownames in titanic survival data to form new data set
#similar to text to columns in excel
#tried merge function which didnt work as expected, but cbind is simpler and gives right data.
str(Titanic3)
TitanicData<-cbind(namessplit,Titanic3)
head(TitanicData)
View(TitanicData)
str(TitanicData)
TitanicData<-cbind(Title,TitanicData)
head(TitanicData)
View(TitanicData)

# There is one more effective way of doing this, and more efficiently
#in the names, we want only titles, i.e Mr or Ms etc.
# names are like this - Braund Mr. Owen Harris
# from these, we need to remove everything after the "."
subtitles<-gsub("\\..*", "", TitanicData$name) # "\\." is read as ".", one more . after that indicates one more charecter after that, and * after . (.*) means all charecters post "."
head(subtitles)
# from subtitles, we need to remove everything before title, including space.
Title<-gsub(".*\\ ", "", subtitles) # putting "." before any charecter, here space represented as "\\ ", selects one charecter before it, and putting * makes it ALL charecters before it.
head(Title)
#graphical representation of the data in various forms
#barplot -No. of passangers by Family name

familyname<-table(TitanicData$family_name)
View(familyname)
barplot(familyname,main = "survival as per family name", xlab = "family_name", ylab = "count",col ="red")

#barplot -No. of passangers by Title

Title<-table(Title)
Title
View(Title)
barplot(Title,xlab = "Title", ylab = "No. of Passangers",
main = "survival as per Title" , col = c("blue", "red"), las=3)
text(Title, 0,table(Title), pos = 3, srt = 90)
```

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

Answer :

```
View(TitanicData)
SurvivedTitle<-table(TitanicData$Survived, TitanicData$title)
#survived is 0, first row. we will take only that
p<-SurvivedTitle[1,]

#barplot of survived numbers per title
barplot(p,xlab = "Title", ylab = "survived",
        main= "Survival as per title", col=rainbow(length(p)))
#pie chaart showing proportion of survival title wise

pie_chart<-pie(p, main = "Pie-Chart of Titles survived", col = rainbow(length(p)) )
legend("topright", names(p), cex= 0.5, fill = rainbow(length(p)))
```

c. Impute the missing values in Age variable using Mice library, create two different graphs showing Age distribution before and after imputation

Answer :

```
library(mice)
sum(is.na(TitanicData$age))
str(TitanicData)


#Removing columns 1,2,3,4,5,7,12,13,14,16,17,18

mini_data <- TitanicData[-c(1,2,3,4,5,7,12,13,14,16,17,18)]
View(mini_data)

md.pattern(mini_data)

library(dplyr)
mini_data <- mini_data %>%
mutate(
survived = as.factor(survived),
sex = as.factor(sex),
age = as.numeric(age),
sibsp = as.factor(sibsp),
parch = as.factor(parch),
embarked = as.factor(embarked)
)
str(mini_data)
mice_data <- mice(mini_data, m=5, maxit=10,seed=500)
summary(mini_data)
```

```
Imputed=complete(mice_data,5)
hist(TitanicData$age, main='Actual Data',col="green")
hist(Imputed$age, main='Imputed Data',col="black")
```