ACADGILD

SESSION 6: Visualization & Plotting

Assignment 1

Table of Contents

1. Introduction

This assignment will help you understand the concepts learnt in the session.

2. Objective

This assignment will test your skills on Visualization and Plotting operations in R.

3. Prerequisites

Not applicable.

4. Associated Data Files

Not applicable.

5. Problem Statement

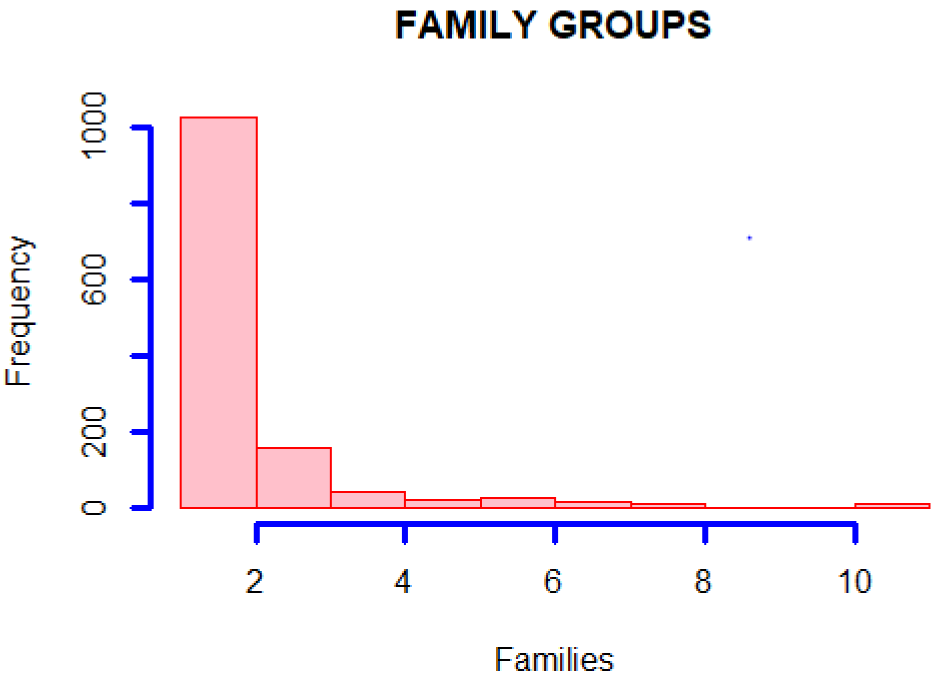
1. Import the Titanic Dataset from the following link: https://drive.google.com/file/d/1JTJCjdGuUxzKXYlwOavwovB01k6FWg3r/view?ts=5b42ea10

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.

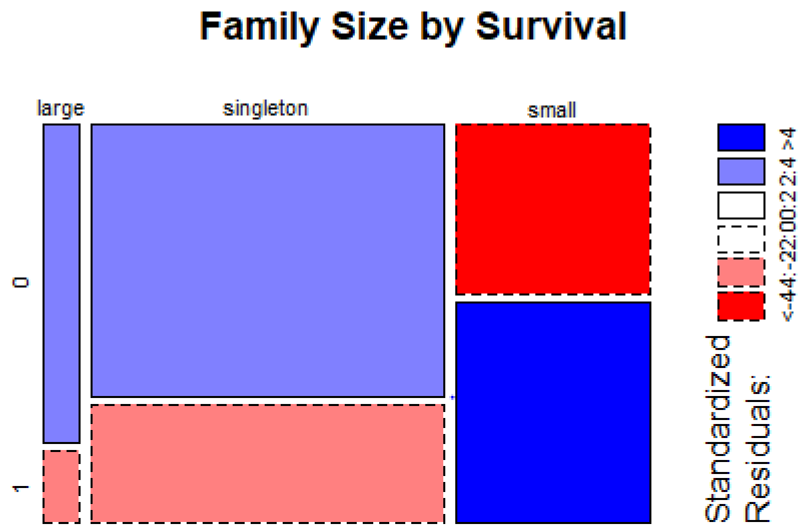
ANS:
```
full=titanic3
  View(full)
  #grab the passenger name
  full$titles=gsub('(.*, )|(\\..*)', '', full$name)
  full$titles
  table(full$sex, full$titles)
  # Titles with very low cell counts to be combined to "rare"
level
  rare_title <- c('Dona', 'Lady', 'the Countess','Capt', 'Col',
'Don','Dr', 'Major', 'Rev', 'Sir', 'Jonkheer')
 # Also reassign mlle, ms, and mme accordingly
 full$titles[full$titles=='Mlle']<- 'Miss'
```

```
full$titles[full$titles=='Ms']<- 'Miss'
full$titles[full$titles=='Mme']<-' Mrs'
full$titles[full$titles %in% rare_title]<- 'Rare Title'
full$titles[full$titles=='Mrs']<-' Mrs'
#Show title counts by sex again
table(full$sex, full$titles)
# Finally , grab surname from passenger name
full$surname= sapply(full$name, function(x) strsplit(x, split =
    '[,.]')[[1]][1])
#cat(paste('We have <b>', nlevels(factor(full$surname)), '</b>
    unique surnames.
#       I would be interested to infer ethnicity based on surname ---
    another time.'))
full$surname
# Create a family size variable including the passenger themselves
full$Fsize <- full$sibsp + full$parch + 1
# Create a family variable
    full$Family <- paste(full$surname, full$Fsize, sep='_')
hist(full$Fsize,col = "pink", bg='red', border = 'red',fg= 'blue',main
    = "FAMILY GROUPS",
       xlab = 'Families', lwd=3)
```



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

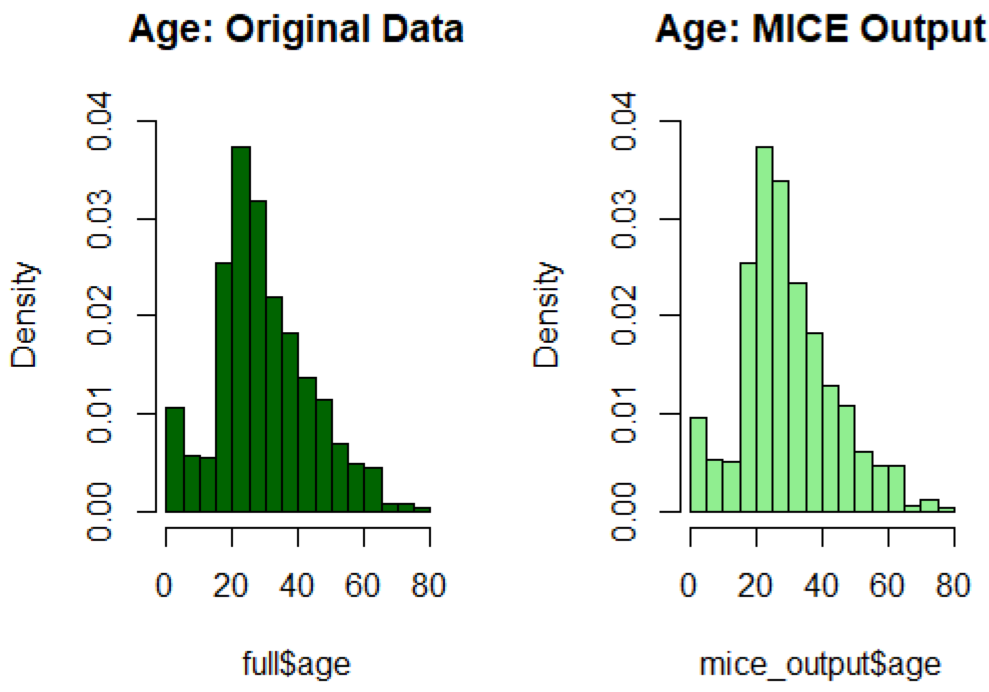
ANS:
```
#graph of familes survived
 attach(full)
# Discretize family size
full$FsizeD[full$Fsize == 1] <- 'singleton'
full$FsizeD[full$Fsize < 5 & full$Fsize > 1] <- 'small'
full$FsizeD[full$Fsize > 4] <- 'large'

# Show family size by survival using a mosaic plot
mosaicplot(table(full$FsizeD, full$survived), main='Family Size by Survival', shade=TRUE)
```



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

ANS:

```
# Make variables factors into factors
      factor_vars <- c('PassengerId','Pclass','Sex','Embarked',
       'Title','Surname','Family','FsizeD')
      full[factor_vars] <- lapply(full[factor_vars], function(x) as.factor(x))
       # Set a random seed
       set.seed(129)
        # Perform mice imputation, excluding certain less-than-useful variables:
      mice_mod=
mice(full[,!names(full)%in%c('PassengerId','Name','Ticket','Cabin','Family','Surname','Survived')],
method='rf')
          # Save the complete output
         mice_output <- complete(mice_mod)
        # Plot age distributions
        par(mfrow=c(1,2))
       hist(full$age, freq=F, main='Age: Original Data',  col='darkgreen', ylim=c(0,0.04))
       hist(mice_output$age, freq=F, main='Age: MICE Output', col='lightgreen', ylim=c(0,0.04))
```


Age: Original Data
Age: MICE Output
Density
full$age
mice_output$age