SESSION 6: Visualization & Plotting

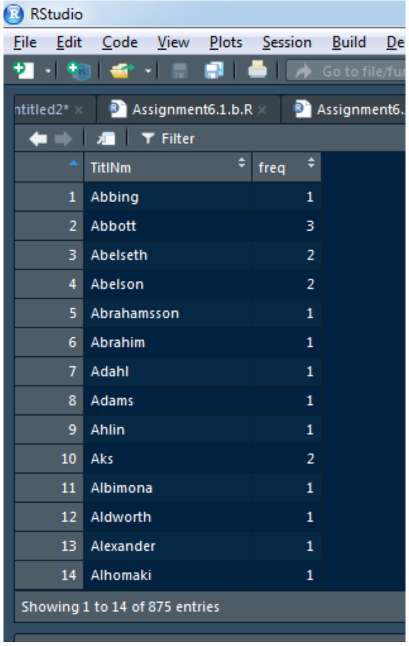
Assignment 1

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:

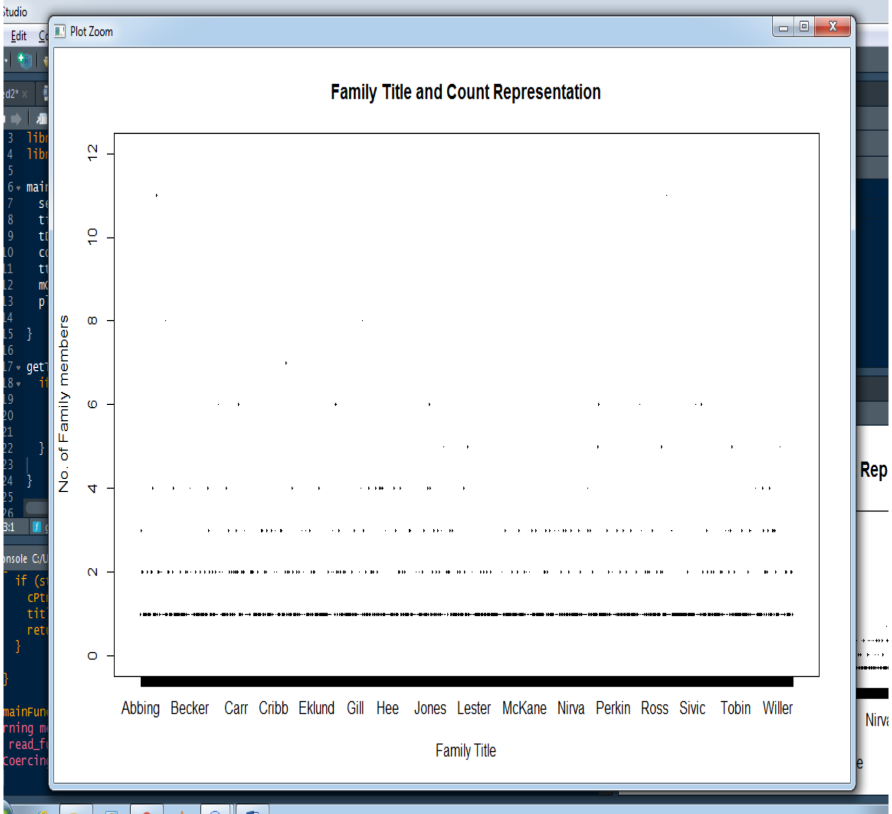
```
library(readxl)
library(stringr)
library(dplyr)
library(plyr)

mainFunc<-function(){
        setwd("C://Users//DELL//Desktop//Assignments//Session6")
        titanicDF <- read_excel("titanic3.xls")
        tDf<-data.frame(cbind(sapply(titanicDF$name,function(x)  getTitle(x),simplify = T)))
        colnames(tDf)<-"TitlNm"
        ttlCnt<-count(tDf, "TitlNm")
        mCnt <- max(ttlCnt[,2])+1
        plot(ttlCnt,type="p",main="Family Title and Count Representation", ylab="No. of Family members",
                xlab="Family Title", ylim=c(0,mCnt))
        View(ttlCnt)
}

getTitle <- function (x){
        if (str_detect(x,",") == T ) {
                cPtr <- str_locate(x,",")
                titleNm<-substr(x,1, cPtr-1)
                return (titleNm)
        }
}
mainFunc()
```

Output:



	TitlNm	freq
1	Abbing	1
2	Abbott	3
3	Abelseth	2
4	Abelson	2
5	Abrahamsson	1
6	Abrahim	1
7	Adahl	1
8	Adams	1
9	Ahlin	1
10	Aks	2
11	Albimona	1
12	Aldworth	1
13	Alexander	1
14	Alhomaki	1

Showing 1 to 14 of 875 entries



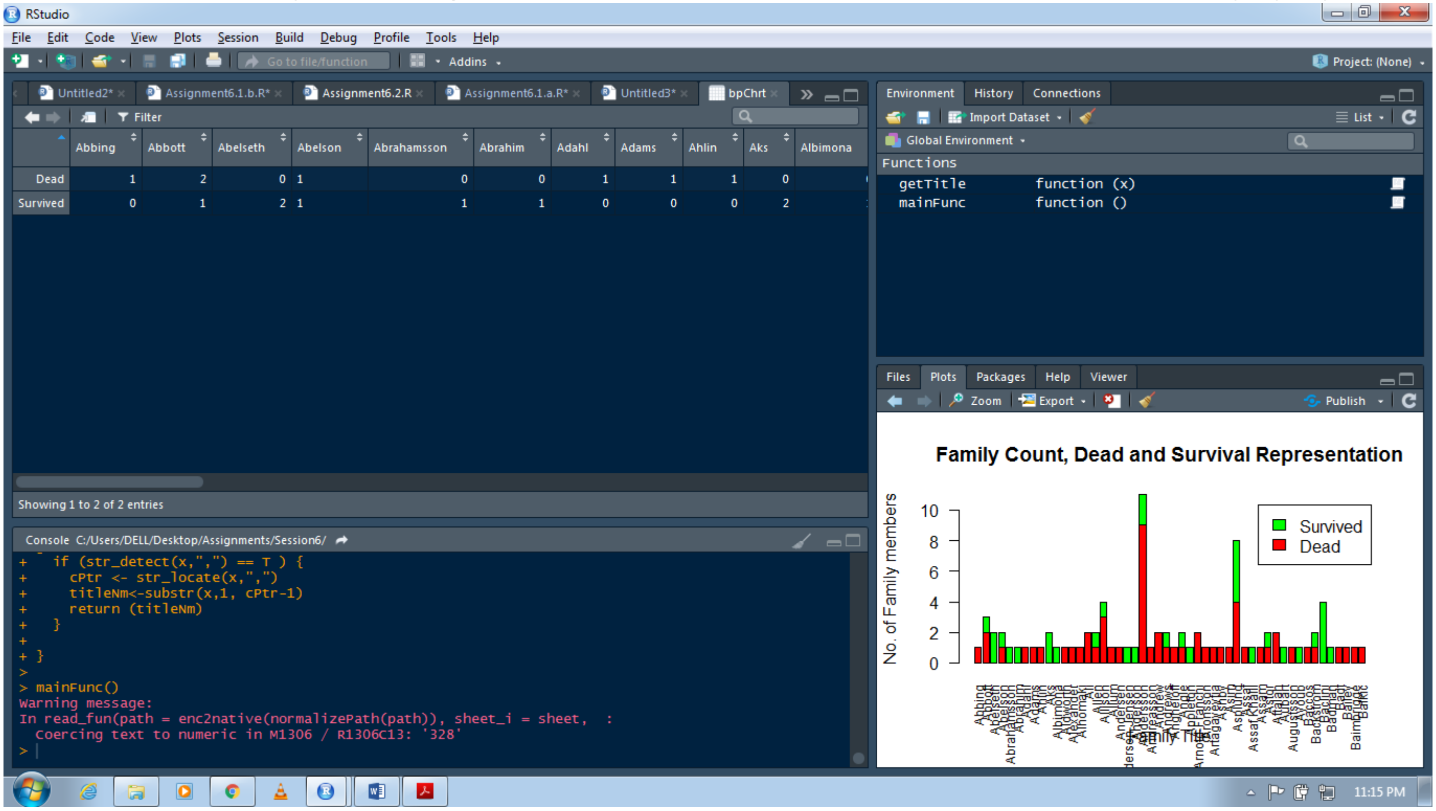
b. Represent the proportion of people survived by family size using a graph.

Answer:

```
library(readxl)
library(stringr)
library(dplyr)
library(plyr)
library(data.table)

mainFunc<-function(){
    setwd("C://Users//DELL//Desktop//Assignments//Session6")
    titanicDF <- read_excel("titanic3.xls")
    tDf<-data.frame(cbind(sapply(titanicDF$name,function(x)  getTitle(x),simplify = T)))
    colnames(tDf)<-"TitlNm"
    ttlCnt<-count(tDf, "TitlNm")

    tDf$survived<-data.frame(cbind(titanicDF$survived))
    survivedCnt<- count(filter(tDf,survived==1), "TitlNm")
    tNsChrt<-merge(ttlCnt, survivedCnt, by.x="TitlNm", by.y="TitlNm", all.x=T)
    nChrt<-data.frame()

    for (i in 1:length(tNsChrt[,1])){
        if ( is.na(tNsChrt[i,3]) == T ) {
            tNsChrt[i,3] = 0
            nChrt[1,i]=tNsChrt[i,2]
            nChrt[2,i]=tNsChrt[i,3]
        }else{
            tNsChrt[i,2] = tNsChrt[i,2] - tNsChrt[i,3]
            nChrt[1,i]=tNsChrt[i,2]
            nChrt[2,i]=tNsChrt[i,3]
        }
    }
    bpChrt<-data.matrix(nChrt[1:50])  # considering 1st 50 record to get clear graph
    rownames(bpChrt)<-c("Dead","Survived")
    colnames(bpChrt)<-tNsChrt$TitlNm[1:50]
    View(bpChrt)
    barplot(bpChrt, col=c("Red","Green"), legend=rownames(bpChrt),main="Family Count, Dead and
        Survival Representation",ylab="No. of Family members", xlab="Family Title", las=2,
        cex.names = 0.75)
}

getTitle <- function (x){
    if (str_detect(x,",") == T ) {
        cPtr <- str_locate(x,",")
        titleNm<-substr(x,1, cPtr-1)
        return (titleNm)
    }
}

mainFunc()
```

Output:

We have considered only 50 rows to get the clear chart. For around 875 rows the chart we not visible properly.



	Abbing	Abbott	Abelseth	Abelson	Abrahamsson	Abrahim	Adahl	Adams	Ahlin	Aks	Albimona
Dead	1	2	0	1	0	0	1	1	1	0	
Survived	0	1	2	1	1	1	0	0	0	2	

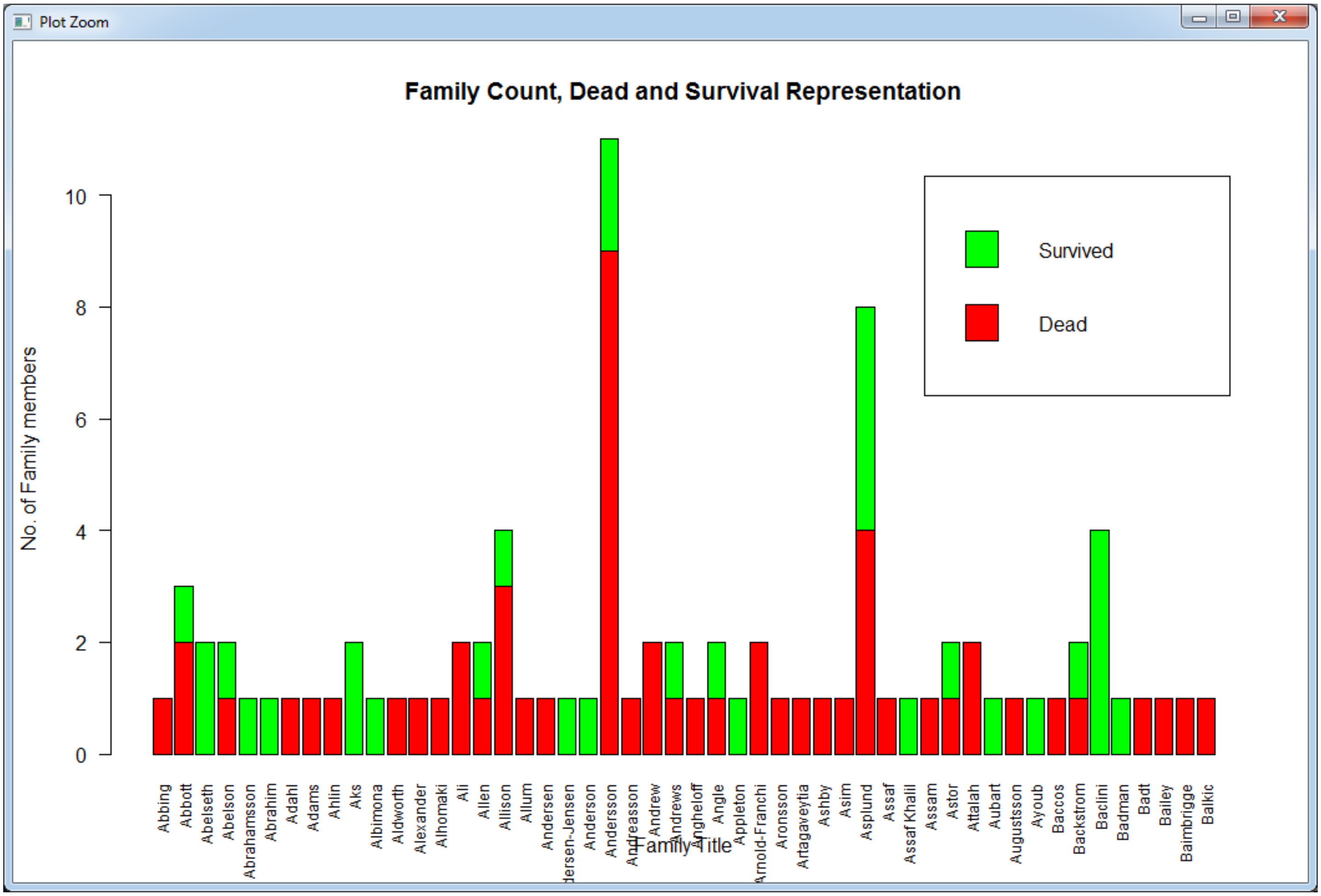
Showing 1 to 2 of 2 entries

Functions

```
getTitle          function (x)
mainFunc          function ()
```

Console C:/Users/DELL/Desktop/Assignments/Session6/

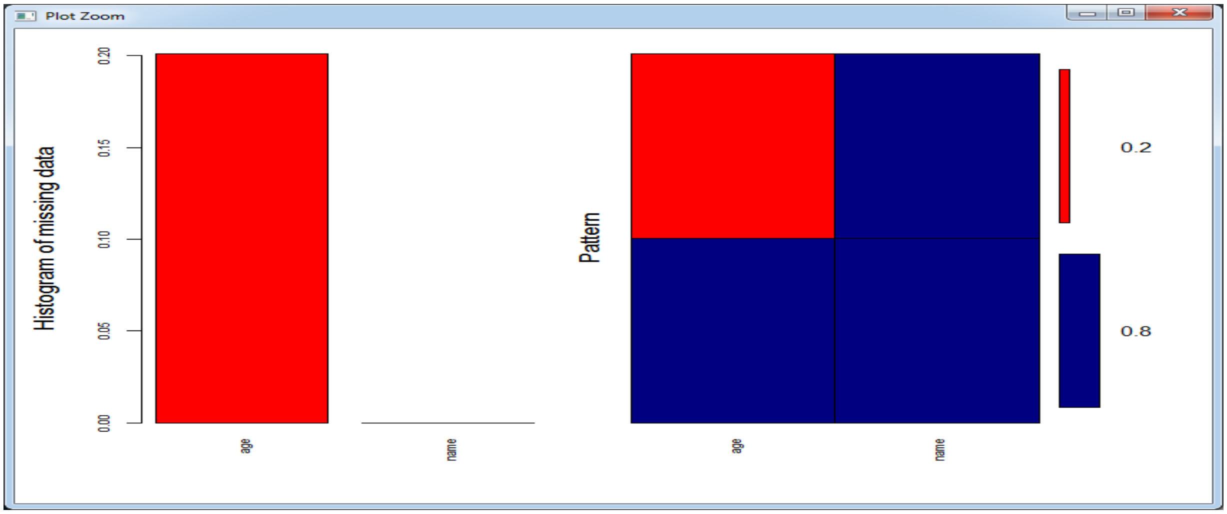
```
+   if (str_detect(x,",") == T ) {
+     cPtr <- str_locate(x,",")
+     titleNm<-substr(x,1, cPtr-1)
+     return (titleNm)
+   }
+
+ }
>
> mainFunc()
Warning message:
In read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet,  :
  Coercing text to numeric in M1306 / R1306C13: '328'
> |
```



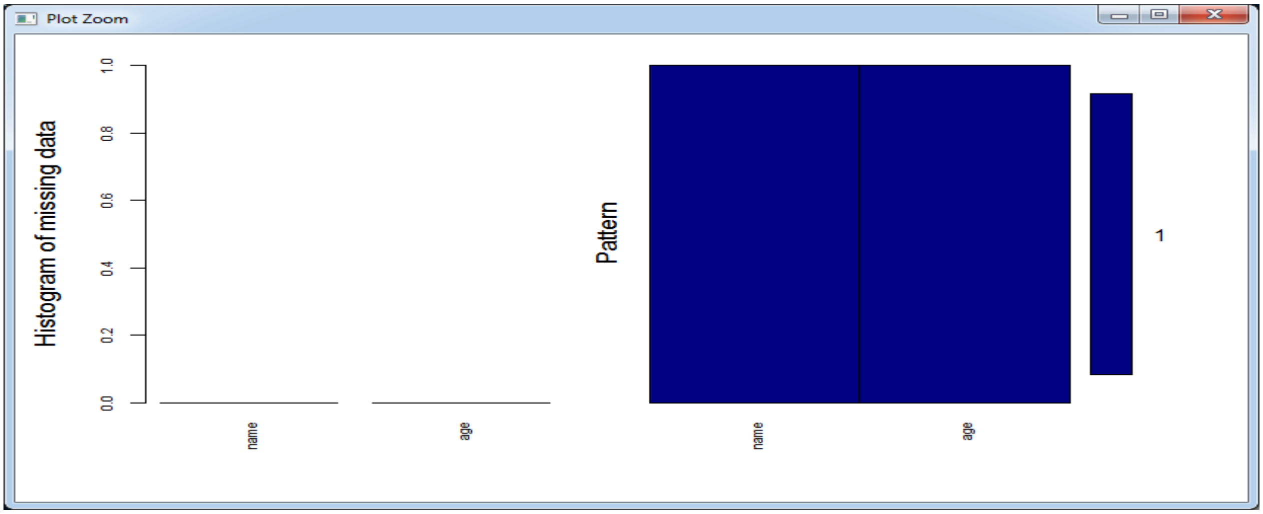
c. Impute the missing values in Age variable using Mice library, create two different graphs showing Age distribution before and after imputation

Answer:

```
library(dplyr)
library(Amelia)
library(mice)
library(ggplot2)
library(lattice)
library(readxl)
library(VIM)

setwd("C://Users//DELL//Desktop//Assignments//Session6")
titanicDF <- read_excel("titanic3.xls")
titanic<-select(titanicDF,name,age)

#before
aggr_plot <- aggr(titanic, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data),
            cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
```



```
#after
imp <- mice(titanic, method = "norm.predict", m = 1)
data_imp <- complete(imp)
aggr_plot <- aggr(data_imp, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data),
            cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
```


Plot Zoom
Histogram of missing data
1.0
0.8
0.6
0.4
0.2
0.0
name
age
Pattern
name
age
1