

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:

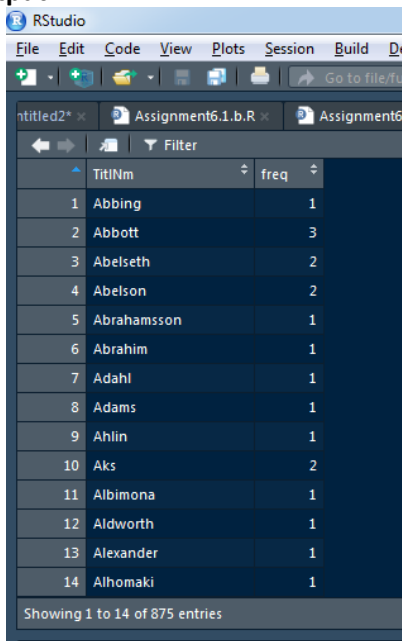
```
library(readxl)
library(stringr)
library(dplyr)
library(plyr)

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)
  }
}

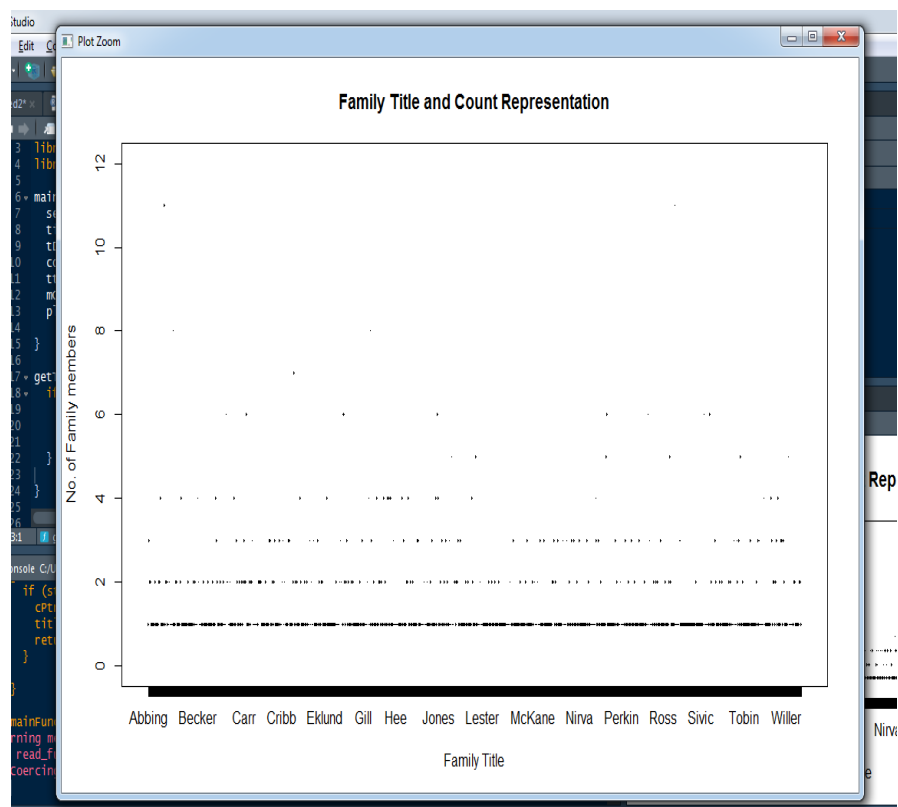
mainFunc()
```

Output:



The RStudio interface shows the script 'Assignment6.1.b.R' and the resulting data table. The table has two columns: 'TitlNm' and 'freq'. It displays the first 14 entries of 875 total entries.

	TitlNm	freq
1	Abbing	1
2	Abbott	3
3	Abelseth	2
4	Abelson	2
5	Abrahamsson	1
6	Abraham	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



- 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()

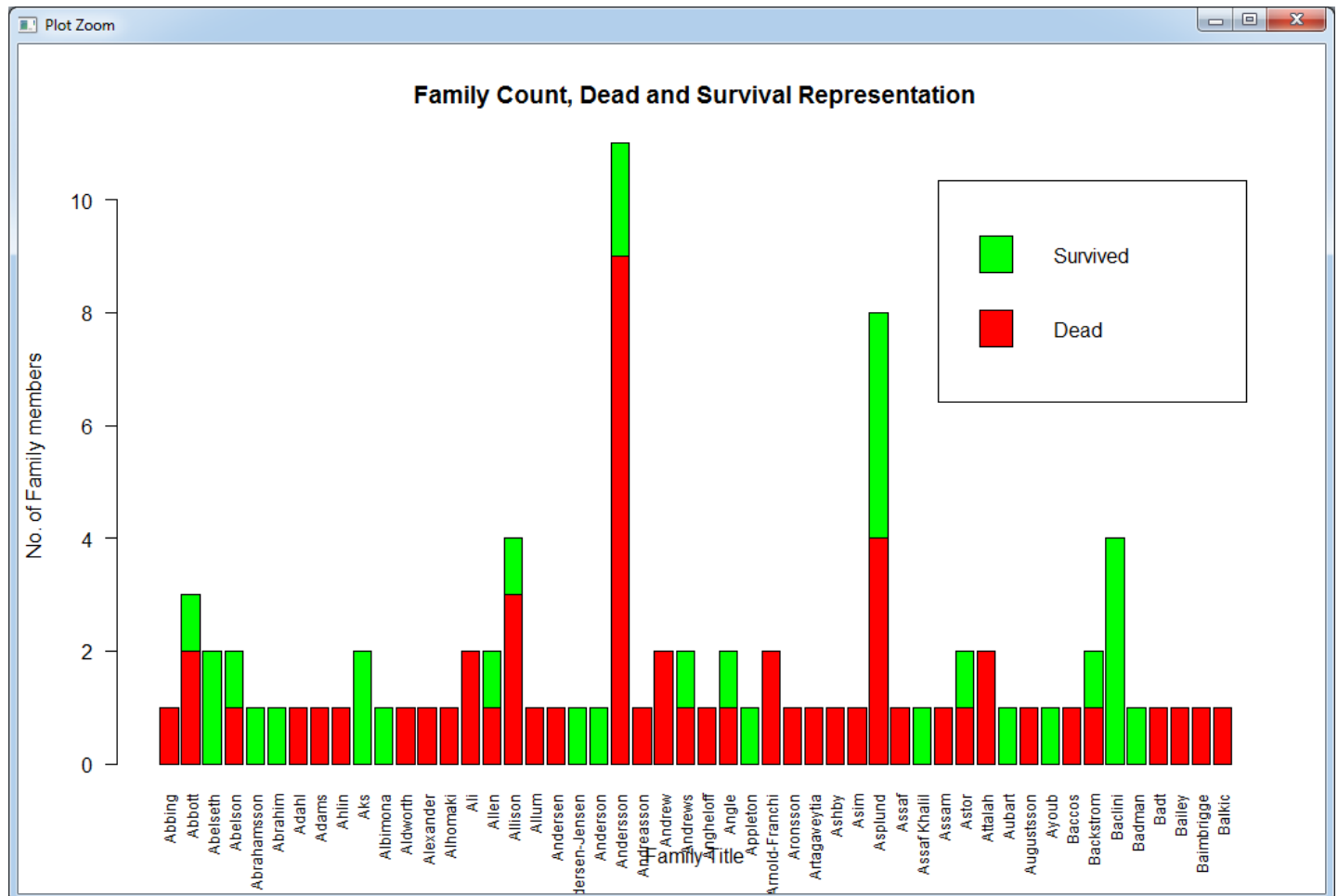
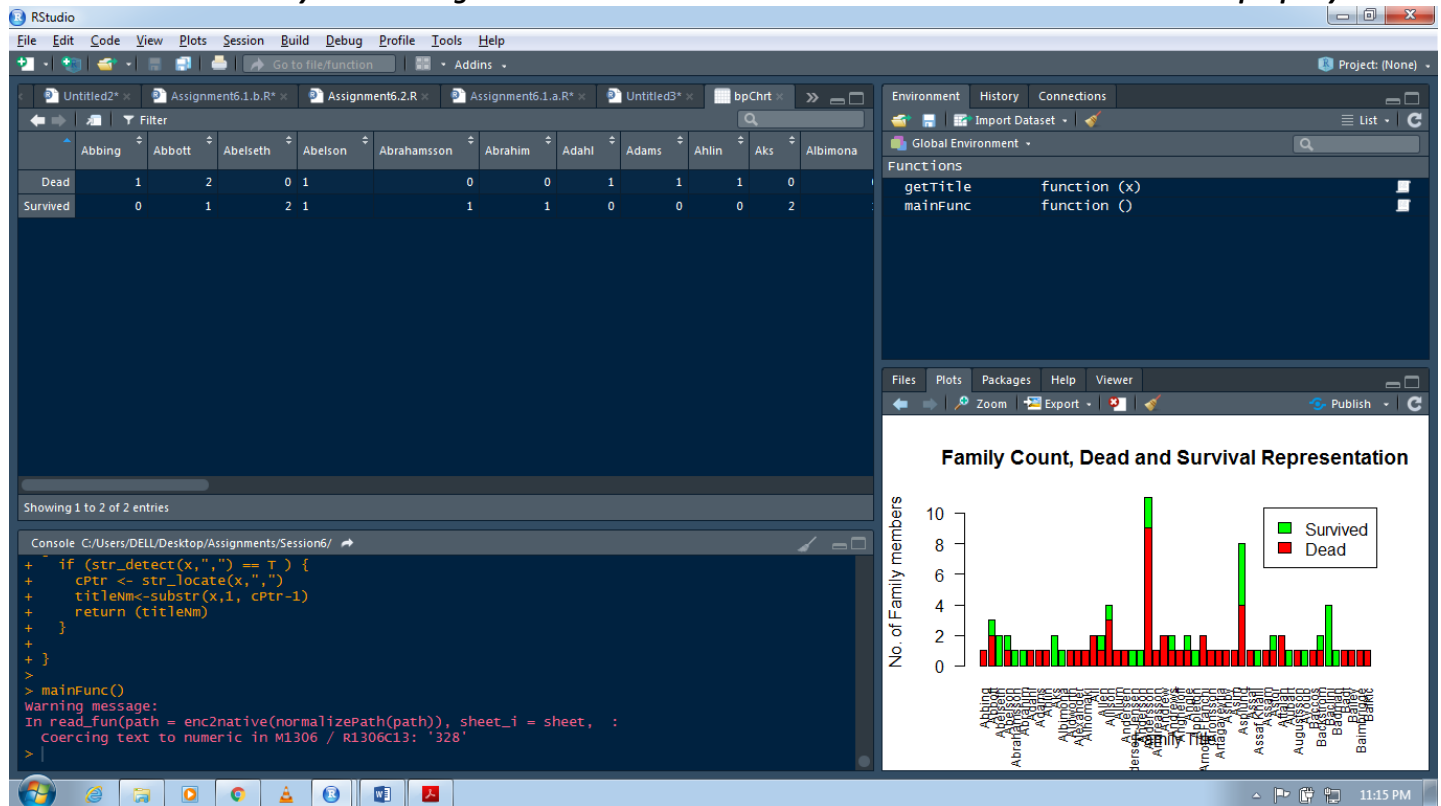
  for (i in 1:length(tNsChrt[,1])){
    if ( is.na(tNsChrt[i,3]) == T ) {
      tNsChrt[i,3] = 0
      nChrt[1,i]=tNsChrt[i,2]
      nChrt[2,i]=tNsChrt[i,3]
    }else{
      tNsChrt[i,2] = tNsChrt[i,2] - tNsChrt[i,3]
      nChrt[1,i]=tNsChrt[i,2]
      nChrt[2,i]=tNsChrt[i,3]
    }
  }
  bpChrt<-data.matrix(nChrt[1:50]) # 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.



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

Answer:

Output: