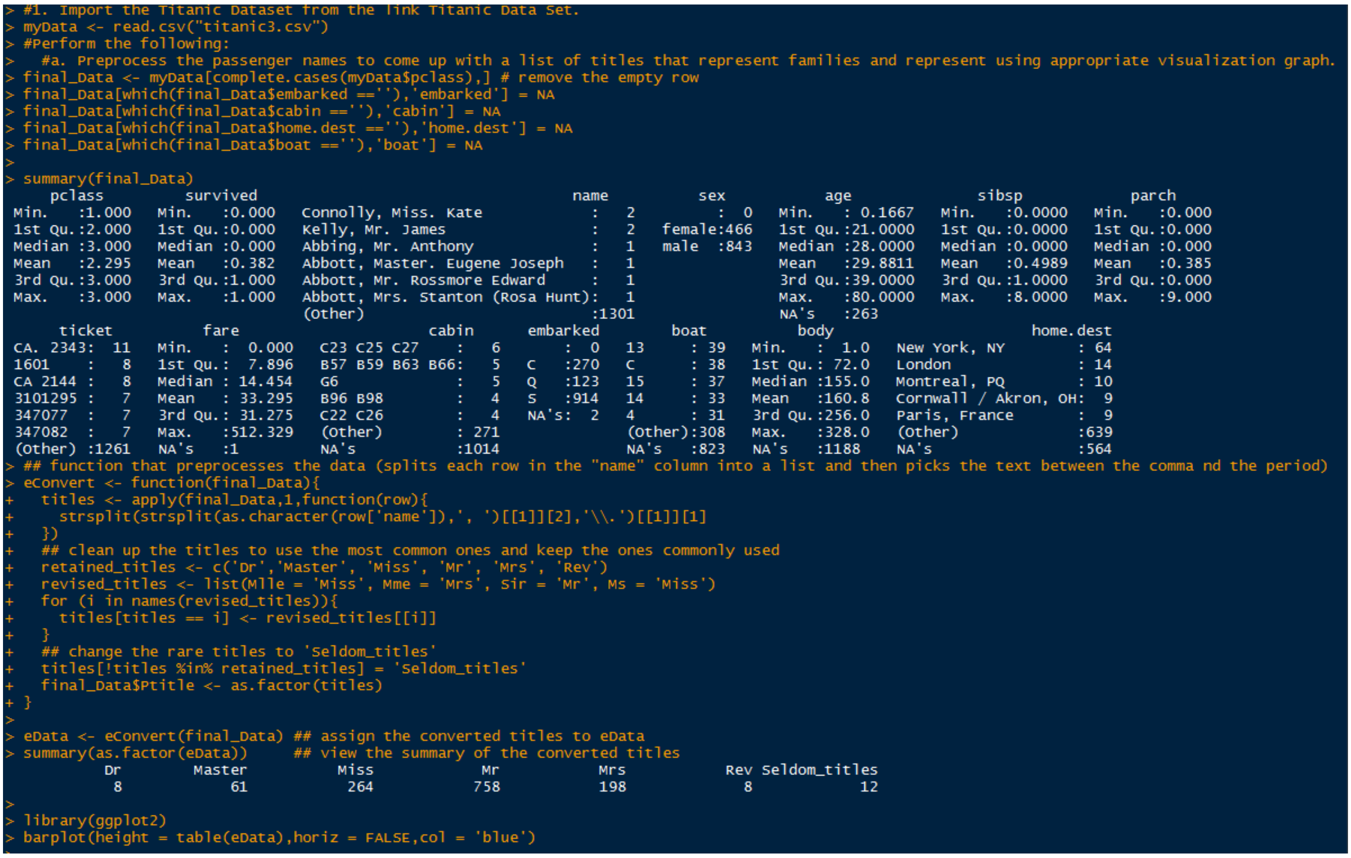
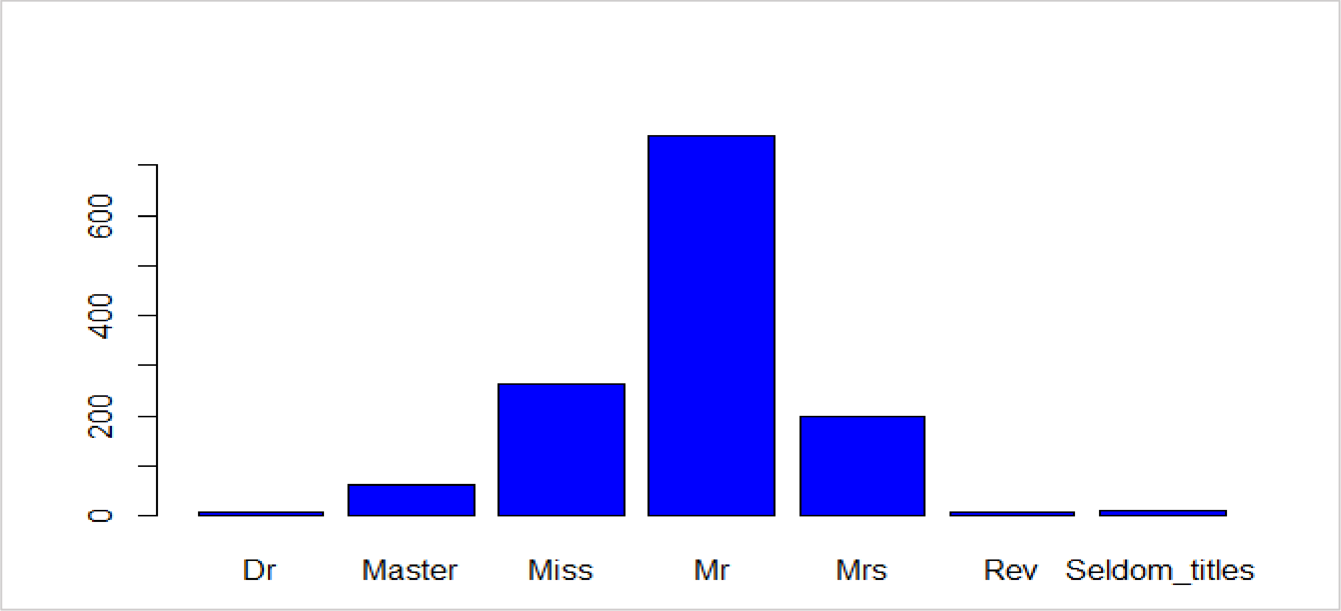
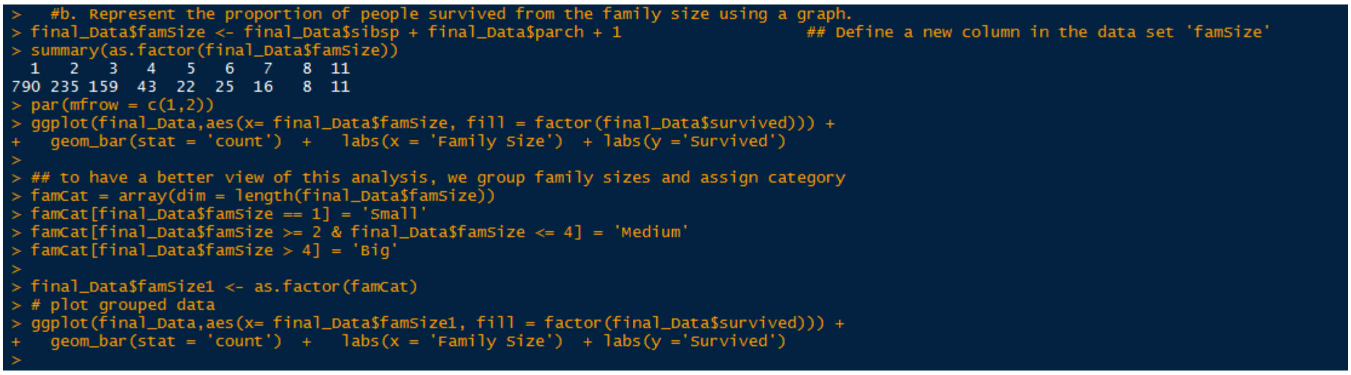
Code

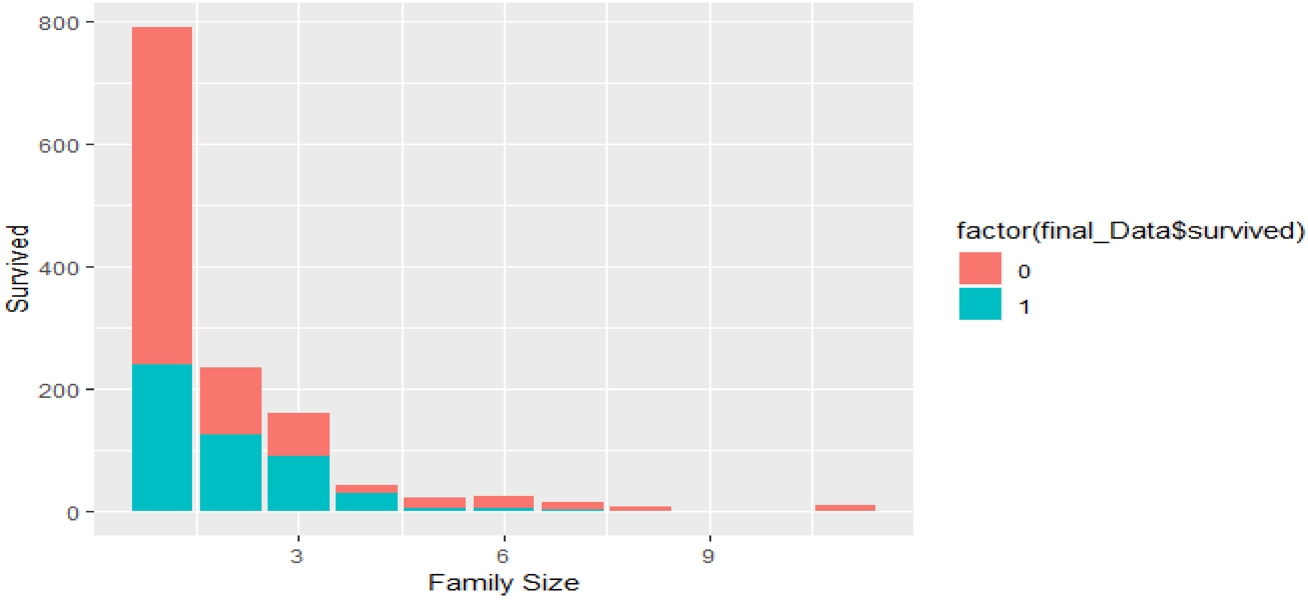


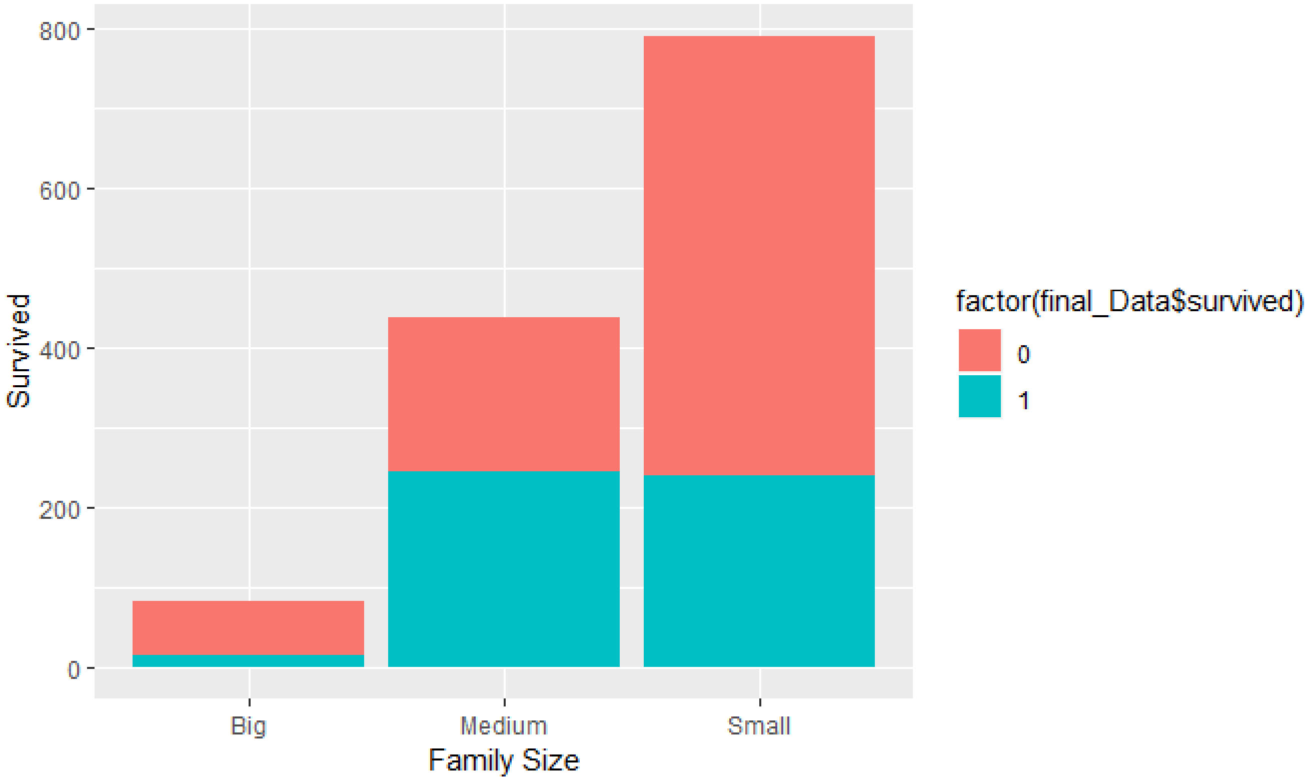
```
> #1. Import the Titanic Dataset from the link Titanic Data Set.
> myData <- read.csv("titanic3.csv")
> #Perform the following:
>   #a. Preprocess the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.
> final_Data <- myData[complete.cases(myData$pclass),] # remove the empty row
> final_Data[which(final_Data$embarked ==''),'embarked'] = NA
> final_Data[which(final_Data$cabin ==''),'cabin'] = NA
> final_Data[which(final_Data$home.dest ==''),'home.dest'] = NA
> final_Data[which(final_Data$boat ==''),'boat'] = NA
> 
> summary(final_Data)
     pclass         survived                                   name          sex           age              sibsp            parch      
 Min.   :1.000   Min.   :0.000   Connolly, Miss. Kate            :   2         :  0   Min.   : 0.1667   Min.   :0.0000   Min.   :0.000  
 1st Qu.:2.000   1st Qu.:0.000   Kelly, Mr. James                :   2   female:466   1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000  
 Median :3.000   Median :0.000   Abbing, Mr. Anthony             :   1   male  :843   Median :28.0000   Median :0.0000   Median :0.000  
 Mean   :2.295   Mean   :0.382   Abbott, Master. Eugene Joseph   :   1                Mean   :29.8811   Mean   :0.4989   Mean   :0.385  
 3rd Qu.:3.000   3rd Qu.:1.000   Abbott, Mr. Rossmore Edward     :   1                3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000  
 Max.   :3.000   Max.   :1.000   Abbott, Mrs. Stanton (Rosa Hunt):   1                Max.   :80.0000   Max.   :8.0000   Max.   :9.000  
                                 (Other)                         :1301                NA's   :263                                       
      ticket          fare                         cabin      embarked        boat            body                    home.dest  
 CA. 2343:  11   Min.   :  0.000   C23 C25 C27    :   6       :  0   13     : 39   Min.   :  1.0   New York, NY        : 64  
 1601    :   8   1st Qu.:  7.896   B57 B59 B63 B66:   5   C   :270   C      : 38   1st Qu.: 72.0   London              : 14  
 CA 2144 :   8   Median : 14.454   G6             :   5   Q   :123   15     : 37   Median :155.0   Montreal, PQ        : 10  
 3101295 :   7   Mean   : 33.295   B96 B98        :   4   S   :914   14     : 33   Mean   :160.8   Cornwall / Akron, OH:  9  
 347077  :   7   3rd Qu.: 31.275   C22 C26        :   4   NA's:  2   4      : 31   3rd Qu.:256.0   Paris, France       :  9  
 347082  :   7   Max.   :512.329   (Other)        : 271              (Other):308   Max.   :328.0   (Other)             :639  
 (Other) :1261   NA's   :1         NA's           :1014              NA's   :823   NA's   :1188    NA's                :564  
> ## function that preprocesses the data (splits each row in the "name" column into a list and then picks the text between the comma nd the period)
> eConvert <- function(final_Data){
+   titles <- apply(final_Data,1,function(row){
+     strsplit(strsplit(as.character(row['name']),', ')[[1]][2],'\\.')[[1]][1]
+   })
+   ## clean up the titles to use the most common ones and keep the ones commonly used
+   retained_titles <- c('Dr','Master', 'Miss', 'Mr', 'Mrs', 'Rev')
+   revised_titles <- list(Mlle = 'Miss', Mme = 'Mrs', Sir = 'Mr', Ms = 'Miss')
+   for (i in names(revised_titles)){
+     titles[titles == i] <- revised_titles[[i]]
+   }
+   ## change the rare titles to 'Seldom_titles'
+   titles[!titles %in% retained_titles] = 'Seldom_titles'
+   final_Data$Ptitle <- as.factor(titles)
+ }
> 
> eData <- eConvert(final_Data) ## assign the converted titles to eData
> summary(as.factor(eData))     ## view the summary of the converted titles
           Dr        Master          Miss            Mr           Mrs           Rev Seldom_titles 
            8            61           264           758           198             8            12 
> 
> library(ggplot2)
> barplot(height = table(eData),horiz = FALSE,col = 'blue')
```

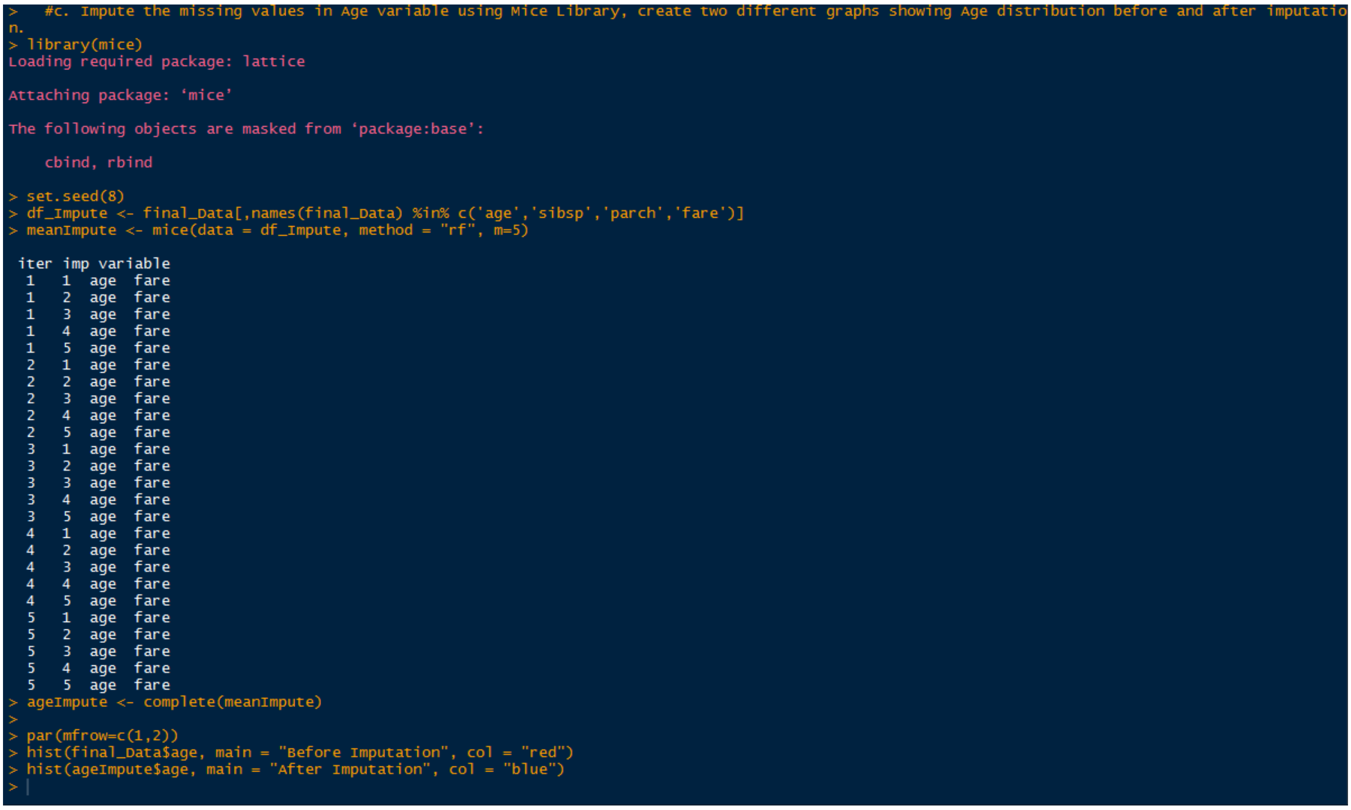




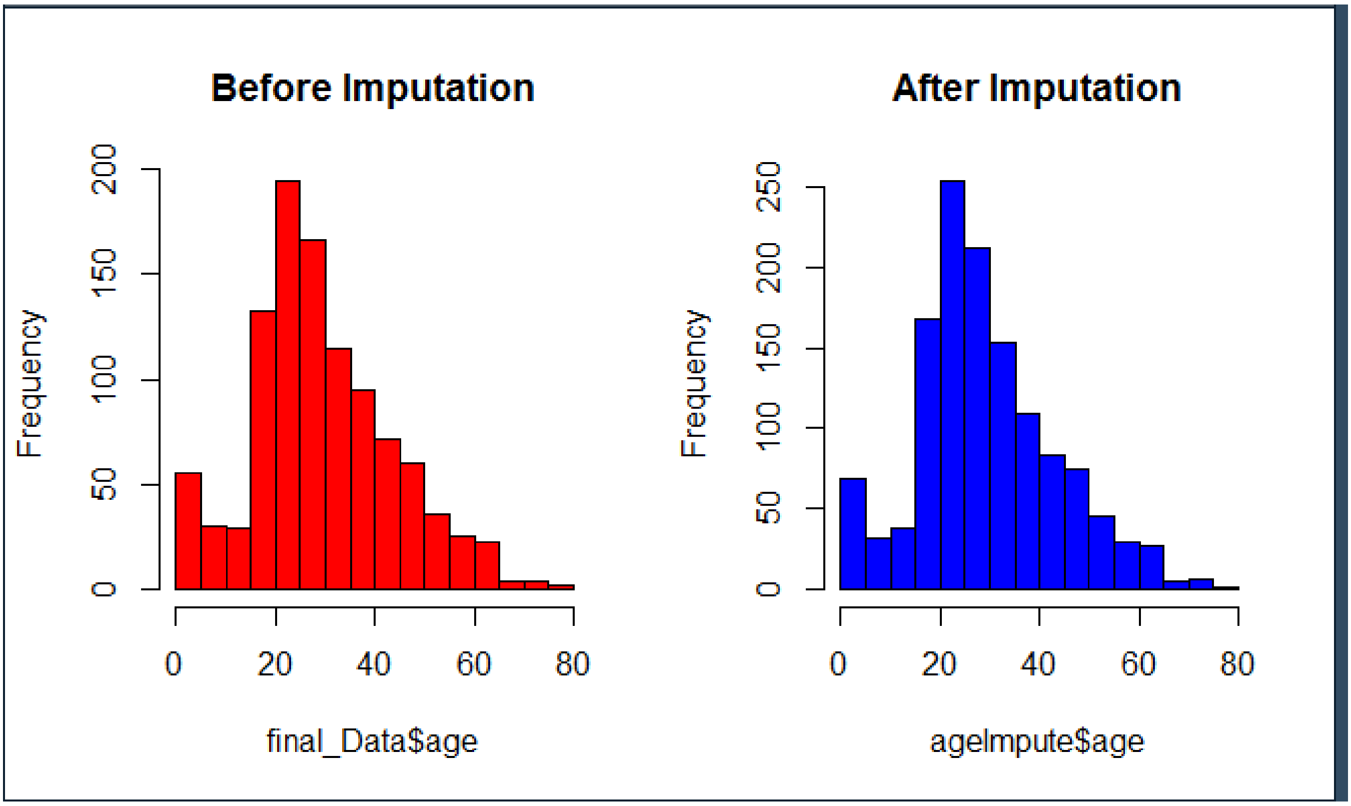
```
>   #b. Represent the proportion of people survived from the family size using a graph.
> final_Data$famSize <- final_Data$sibsp + final_Data$parch + 1                    ## Define a new column in the data set 'famSize'
> summary(as.factor(final_Data$famSize))
  1   2   3   4   5   6   7   8  11 
790 235 159  43  22  25  16   8  11 
> par(mfrow = c(1,2))
> ggplot(final_Data,aes(x= final_Data$famSize, fill = factor(final_Data$survived))) +
+   geom_bar(stat = 'count')  +   labs(x = 'Family Size')  + labs(y ='Survived')
> 
> ## to have a better view of this analysis, we group family sizes and assign category
> famCat = array(dim = length(final_Data$famSize))
> famCat[final_Data$famSize == 1] = 'Small'
> famCat[final_Data$famSize >= 2 & final_Data$famSize <= 4] = 'Medium'
> famCat[final_Data$famSize > 4] = 'Big'
> 
> final_Data$famSize1 <- as.factor(famCat)
> # plot grouped data
> ggplot(final_Data,aes(x= final_Data$famSize1, fill = factor(final_Data$survived))) +
+   geom_bar(stat = 'count')  +   labs(x = 'Family Size')  + labs(y ='Survived')
> 
```







```
>   #c. Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputatio
n.
> library(mice)
Loading required package: lattice

Attaching package: ‘mice’

The following objects are masked from ‘package:base’:

    cbind, rbind

> set.seed(8)
> df_Impute <- final_Data[,names(final_Data) %in% c('age','sibsp','parch','fare')]
> meanImpute <- mice(data = df_Impute, method = "rf", m=5)

 iter imp variable
  1   1  age  fare
  1   2  age  fare
  1   3  age  fare
  1   4  age  fare
  1   5  age  fare
  2   1  age  fare
  2   2  age  fare
  2   3  age  fare
  2   4  age  fare
  2   5  age  fare
  3   1  age  fare
  3   2  age  fare
  3   3  age  fare
  3   4  age  fare
  3   5  age  fare
  4   1  age  fare
  4   2  age  fare
  4   3  age  fare
  4   4  age  fare
  4   5  age  fare
  5   1  age  fare
  5   2  age  fare
  5   3  age  fare
  5   4  age  fare
  5   5  age  fare
> ageImpute <- complete(meanImpute)
>
> par(mfrow=c(1,2))
> hist(final_Data$age, main = "Before Imputation", col = "red")
> hist(ageImpute$age, main = "After Imputation", col = "blue")
> |
```


Before Imputation
Frequency
final_Data$age
After Imputation
Frequency
ageImpute$age