Session 6 – Visualization and Plotting
Assignment – 1

Problem Statement
1. Import the Titanic Dataset from the link Titanic Data Set.
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.
Ans.

```
extractAndConvertTitles <- function(dataset){ titles <- apply(dataset,1,function(row){
strsplit(strsplit(as.character(row['Name']),', ')[[1]][2],'\.')[[1]][1] }) keep_titles <-
c('Dr','Master', 'Miss', 'Mr', 'Mrs') replacementTitles <- list(Mlle = 'Miss', Mme = 'Mrs',
Sir = 'Mr', Ms = 'Miss') for(r_title in names(replacementTitles)){ titles[titles == r_title]
<- replacementTitles[[r_title]] }

titles[!titles %in% keep_titles] = 'Rare Title' dataset$Title <- as.factor(titles)
invisible(dataset) } dataset <- extractAndConvertTitles(dataset) dataset$Name <-
NULL summary(dataset$Title[dataset$mode == 'Training'])
```

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

```
familySize <- dataset$SibSp + dataset$Parch + 1 familySizeClass = array(dim =
length(familySize)) familySizeClass[familySize == 1] = 'Small'
familySizeClass[familySize >= 2 & familySize <= 4] = 'Medium'
familySizeClass[familySize > 4] = 'Big'

dataset$FamilySize <- as.factor(familySizeClass)

ggplot(training, aes(FamilySize, fill = Survived)) + geom_bar(position = 'fill') +
ggtitle('Family Size Impact on Survival') + labs(y = '%')
```

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

```
ageImputBySex_Pclass <- function(dataset, averageAgeStats){
calculateImputedAge <- function(dfRow, ageEvaluationSex_Pclass){ filterIndex <-
ageEvaluationSex_Pclass$Sex == dfRow['Sex'] & ageEvaluationSex_Pclass$Pclass
== dfRow['Pclass'] impAge <- ageEvaluationSex_Pclass[filterIndex,]$meanAge }

dataset$Age[is.na(dataset$Age)] <- apply(dataset[is.na(dataset$Age),], 1,
calculateImputedAge, ageEvaluationSex_Pclass) invisible(dataset) }

dataset <- ageImputBySex_Pclass(dataset, averageAgeStats)
```