**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$Age[[is.na](http://is.na/)(dataset$Age)] <- apply(dataset[[is.na](http://is.na/)(dataset$Age),], 1, calculateImputedAge, ageEvaluationSex\_Pclass) invisible(dataset) }

dataset <- ageImputBySex\_Pclass(dataset, averageAgeStats)