Assignment 6.1

Ans

```
library('ggplot2') # visualization
library('ggthemes') # visualization
library('scales') # visualization
library('dplyr') # data manipulation
library('mice') # imputation
library('randomForest') # classification algorithm

train <- read.csv('../input/train.csv', stringsAsFactors = F)
test  <- read.csv('../input/test.csv', stringsAsFactors = F)

full  <- bind_rows(train, test) # bind training & test data

# check data
str(full)


# Create a family size variable including the passenger themselves
full$Fsize <- full$SibSp + full$Parch + 1

# Create a family variable
full$Family <- paste(full$Surname, full$Fsize, sep='_')


# Use ggplot2 to visualize the relationship between family size & survival
ggplot(full[1:891,], aes(x = Fsize, fill = factor(Survived))) +
  geom_bar(stat='count', position='dodge') +
  scale_x_continuous(breaks=c(1:11)) +
  labs(x = 'Family Size') +
  theme_few()




library(mice)
init = mice(tr, maxit=0)
predM = init$predictorMatrix
# Do not use following columns to impute values in 'Age'. Use the rest.
predM[, c("PassengerId", "Name","Ticket","Cabin")]=0
imp<-mice(tr, m=5, predictorMatrix = predM)
# Get the final data-frame with imputed values filled in 'Age'
tr <- complete(imp)
View(tr)
```