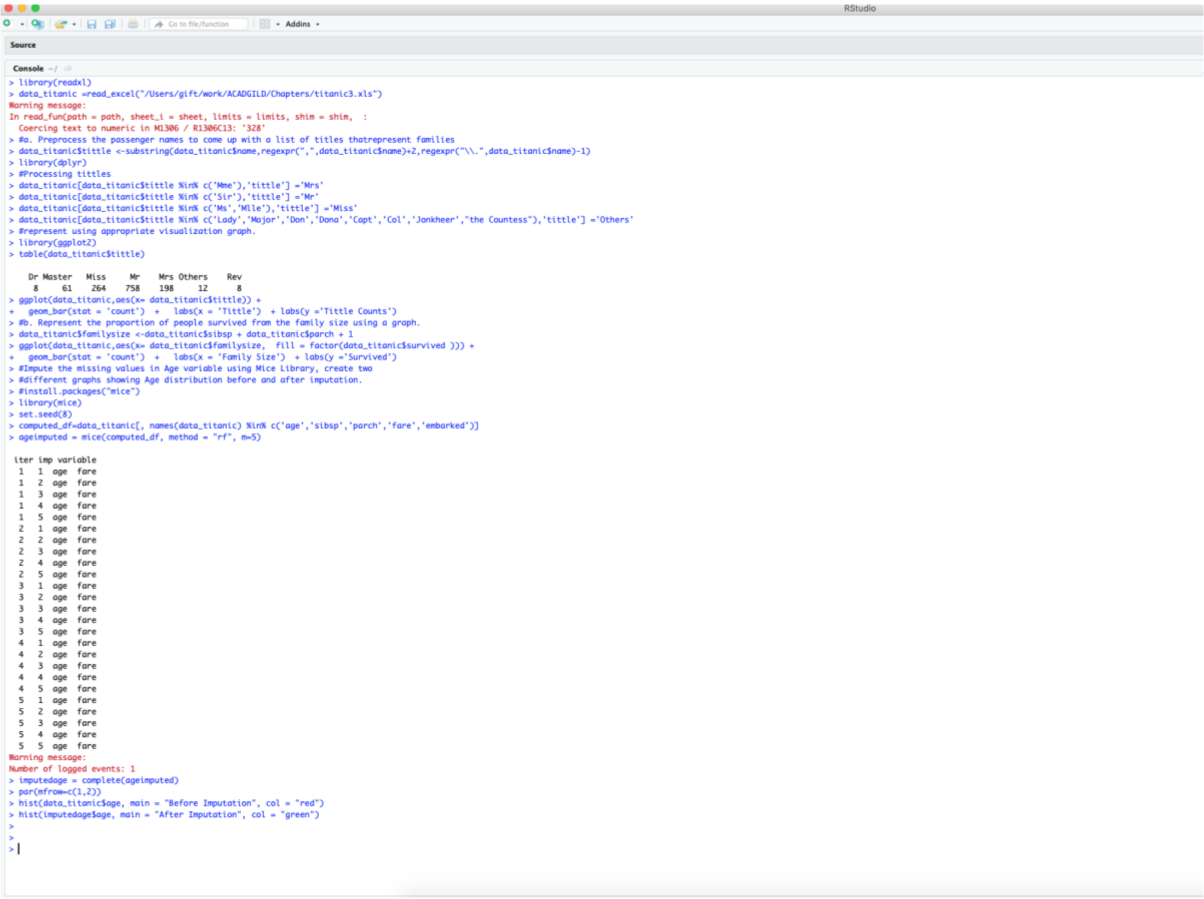
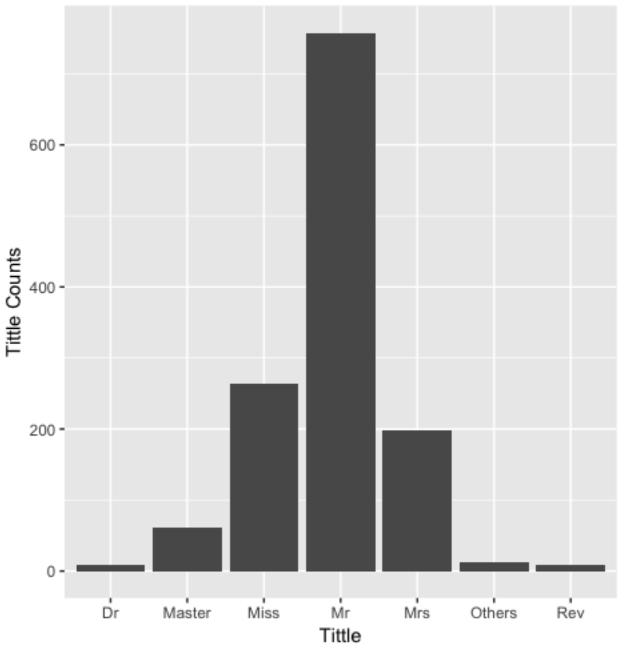
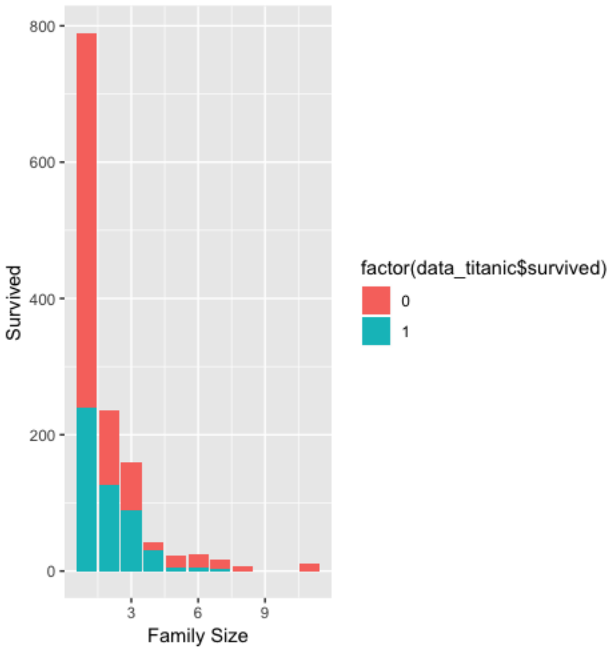
Assignment 3 Results



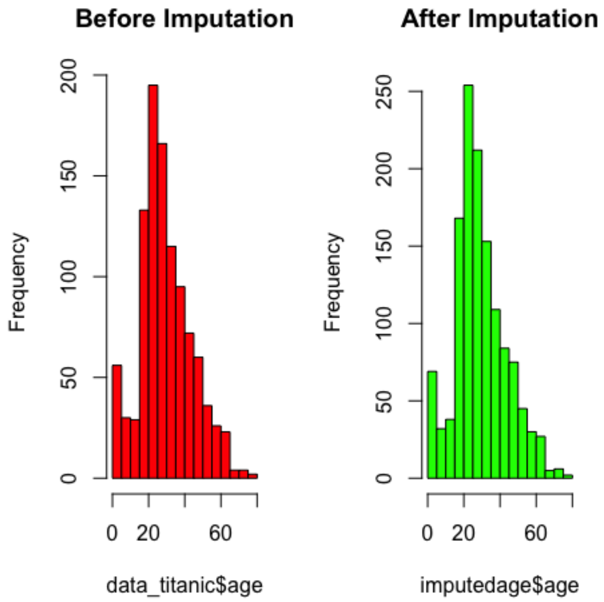
```
> library(readxl)
> data_titanic =read_excel("/Users/gift/work/ACADGILD/Chapters/titanic3.xls")
Warning message:
In read_fun(path = path, sheet_i = sheet, limits = limits, shim = shim,  :
  Coercing text to numeric in M1306 / R1306C13: '328'
> #a. Preprocess the passenger names to come up with a list of titles thatrepresent families
> data_titanic$tittle <-substring(data_titanic$name,regexpr(",",data_titanic$name)+2,regexpr("\\.",data_titanic$name)-1)
> library(dplyr)
> #Processing tittles
> data_titanic[data_titanic$tittle %in% c('Mme'),'tittle'] ='Mrs'
> data_titanic[data_titanic$tittle %in% c('Sir'),'tittle'] ='Mr'
> data_titanic[data_titanic$tittle %in% c('Ms','Mlle'),'tittle'] ='Miss'
> data_titanic[data_titanic$tittle %in% c('Lady','Major','Don','Dona','Capt','Col','Jonkheer',"the Countess"),'tittle'] ='Others'
> #represent using appropriate visualization graph.
> library(ggplot2)
> table(data_titanic$tittle)

    Dr Master   Miss     Mr    Mrs Others    Rev 
     8     61    264    758    198     12      8 
> ggplot(data_titanic,aes(x= data_titanic$tittle)) +
+   geom_bar(stat = 'count')  +   labs(x = 'Tittle')  + labs(y ='Tittle Counts')
> #b. Represent the proportion of people survived from the family size using a graph.
> data_titanic$familysize <-data_titanic$sibsp + data_titanic$parch + 1
> ggplot(data_titanic,aes(x= data_titanic$familysize,  fill = factor(data_titanic$survived ))) +
+   geom_bar(stat = 'count')  +   labs(x = 'Family Size')  + labs(y ='Survived')
> #Impute the missing values in Age variable using Mice Library, create two
> #different graphs showing Age distribution before and after imputation.
> #install.packages("mice")
> library(mice)
> set.seed(8)
> computed_df=data_titanic[, names(data_titanic) %in% c('age','sibsp','parch','fare','embarked')]
> ageimputed = mice(computed_df, 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
Warning message:
Number of logged events: 1 
> imputedage = complete(ageimputed)
> par(mfrow=c(1,2))
> hist(data_titanic$age, main = "Before Imputation", col = "red")
> hist(imputedage$age, main = "After Imputation", col = "green")
> 
> 
> |
> table(data_titanic$tittle)

   Dr Master   Miss     Mr    Mrs Others    Rev
    8     61    264    758    198     12      8
> ggplot(data_titanic,aes(x= data_titanic$tittle)) +
+   geom_bar(stat = 'count')  +   labs(x = 'Tittle')  + labs(y ='Tittle Counts')
>
```



```
> table(data_titanic$tittle)

   Dr Master   Miss     Mr    Mrs Others    Rev
    8     61    264    758    198     12      8
> ggplot(data_titanic,aes(x= data_titanic$tittle)) +
+   geom_bar(stat = 'count')  +   labs(x = 'Tittle')  + labs(y ='Tittle Counts')
```



```
> imputedage = complete(ageimputed)
> par(mfrow=c(1,2))
> hist(data_titanic$age, main = "Before Imputation", col = "red")
> hist(imputedage$age, main = "After Imputation", col = "green")
>
```


Before Imputation
Frequency
data_titanic$age
After Imputation
Frequency
imputedage$age