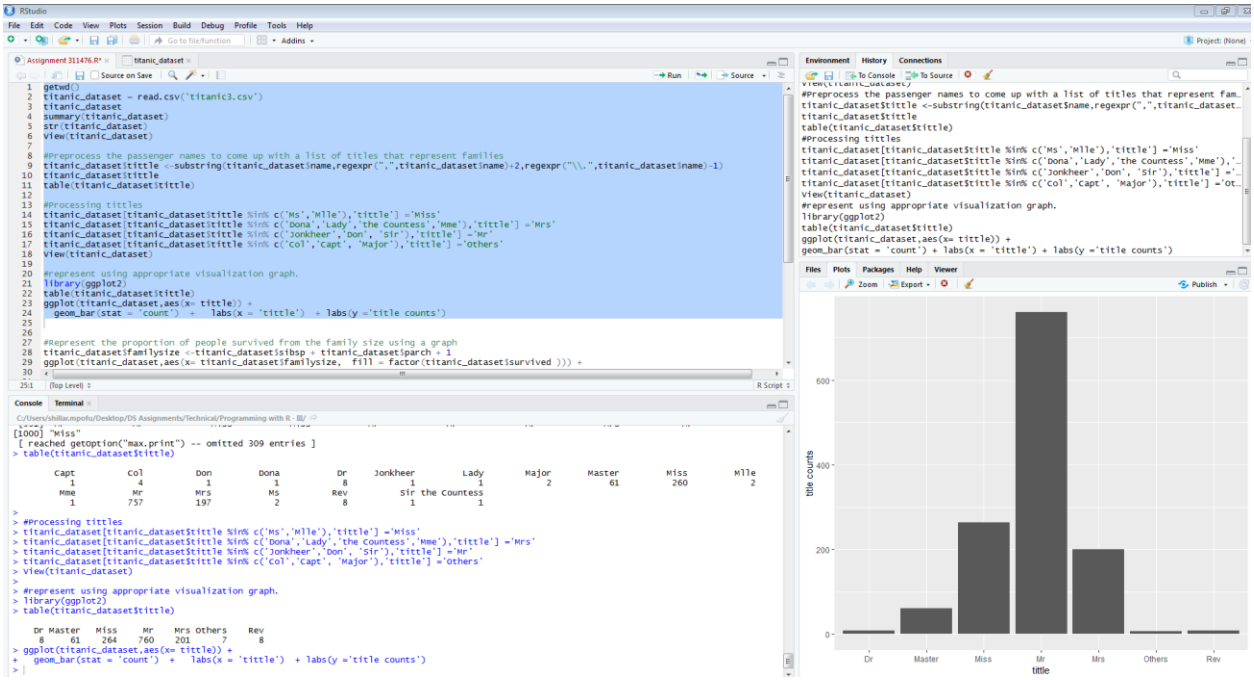
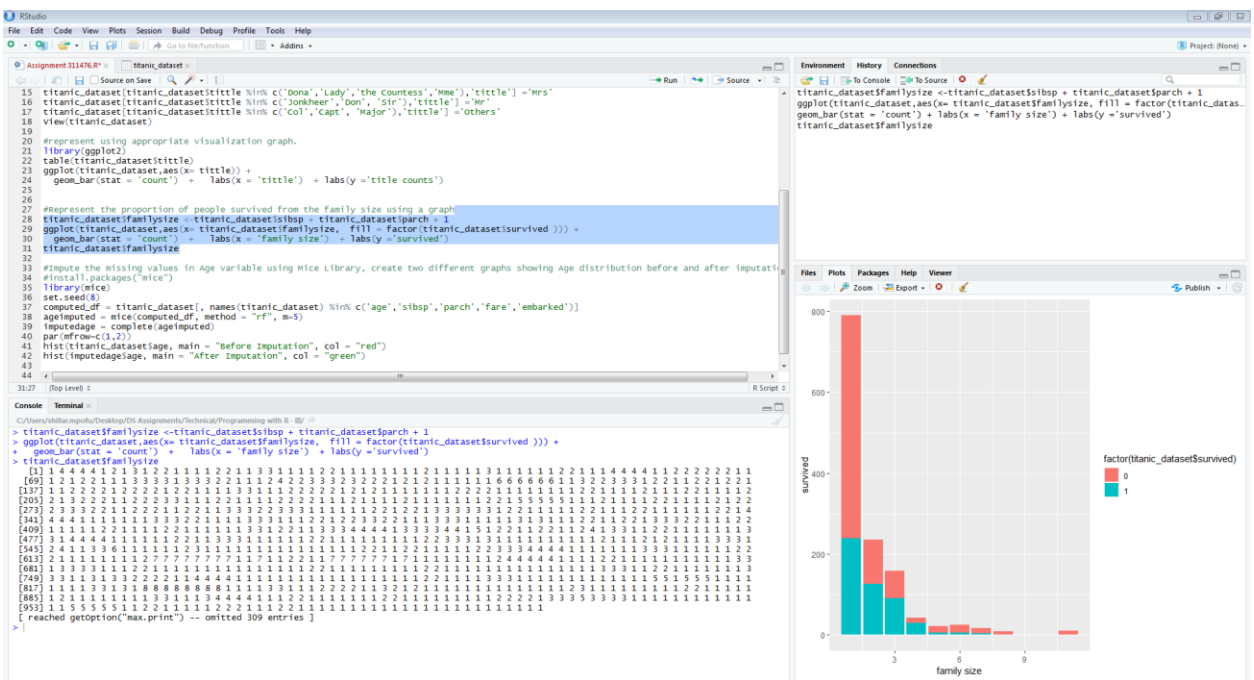
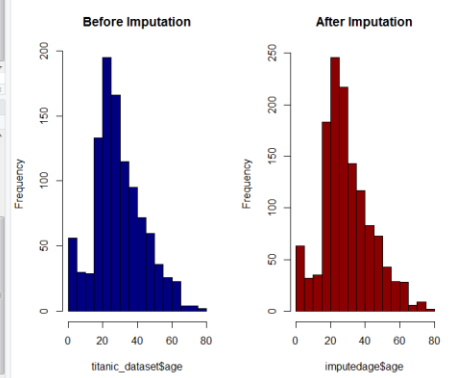



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Assignment 311476.R*
titanic_dataset
Source on Save
Run
Source
titanic_dataset[titanic_dataset$tittle %in% c('Dona','Lady','the Countess','Mme'),'tittle'] ='Mrs'
titanic_dataset[titanic_dataset$tittle %in% c('Jonkheer','Don', 'Sir'),'tittle'] ='Mr'
titanic_dataset[titanic_dataset$tittle %in% c('Col','Capt', 'Major'),'tittle'] ='Others'
View(titanic_dataset)
#represent using appropriate visualization graph.
library(ggplot2)
table(titanic_dataset$tittle)
ggplot(titanic_dataset,aes(x= tittle)) +
geom_bar(stat = 'count') + labs(x = 'tittle') + labs(y ='title counts')
#Represent the proportion of people survived from the family size using a graph
titanic_dataset$familysize <-titanic_dataset$sibsp + titanic_dataset$parch + 1
ggplot(titanic_dataset,aes(x= titanic_dataset$familysize, fill = factor(titanic_dataset$survived))) +
geom_bar(stat = 'count') + labs(x = 'family size') + labs(y ='survived')
titanic_dataset$familysize
#Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputati
#install.packages("mice")
library(mice)
set.seed(8)
computed_df = titanic_dataset[, names(titanic_dataset) %in% c('age','sibsp','parch','fare','embarked')]
ageimputed = mice(computed_df, method = "rf", m=5)
imputedage = complete(ageimputed)
par(mfrow=c(1,2))
hist(titanic_dataset$age, main = "Before Imputation", col = "red")
hist(imputedage$age, main = "After Imputation", col = "green")
Console Terminal
C:/Users/shillar.mpofu/Desktop/DS Assignments/Technical/Programming with R - III/
> titanic_dataset$familysize <-titanic_dataset$sibsp + titanic_dataset$parch + 1
> ggplot(titanic_dataset,aes(x= titanic_dataset$familysize, fill = factor(titanic_dataset$survived))) +
+ geom_bar(stat = 'count') + labs(x = 'family size') + labs(y ='survived')
> titanic_dataset$familysize
[1] 1 4 4 4 4 1 2 1 3 1 2 2 1 1 1 1 2 2 1 1 1 3 3 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 3 1 1 1 1 1 1 2 2 1 1 1 4 4 4 4 1 1 2 2 2 2 2 1 1
[69] 1 2 1 2 2 1 1 1 3 3 3 3 1 3 3 2 2 1 1 1 2 4 2 2 3 3 3 2 3 2 2 1 2 1 2 1 1 1 1 1 1 6 6 6 6 6 6 1 1 3 2 2 3 3 3 1 2 2 1 1 2 2 1 2 2 1
[137] 1 1 2 2 2 2 1 2 2 2 2 1 2 2 1 1 1 1 3 3 1 1 1 2 2 2 2 1 2 1 2 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 1 1 1 2 2 1 1 1 1 2
[205] 2 1 3 2 2 2 1 1 2 2 2 3 3 1 1 1 2 2 1 1 1 2 2 2 2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 5 5 5 5 5 1 1 1 2 1 1 1 2 2 1 1 1 1 2 1 2 2
[273] 2 3 3 3 2 2 1 1 2 2 2 1 1 2 2 1 1 3 3 3 2 2 3 3 3 1 1 1 1 1 2 2 1 2 2 1 3 3 3 3 3 3 1 2 2 1 1 1 1 2 2 1 1 2 2 1 1 1 1 1 2 2 1 4
[341] 4 4 4 1 1 1 1 1 1 1 3 3 3 2 2 1 1 1 1 3 3 3 1 1 1 2 2 1 2 2 3 3 2 2 1 1 1 3 3 1 1 1 1 1 3 1 3 1 1 1 1 2 1 1 2 2 1 3 3 3 2 2 1 1 1 2 2
[409] 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 1 1 3 3 1 2 2 1 1 3 3 3 4 4 4 4 1 3 3 3 3 4 4 1 5 1 2 2 1 1 2 2 1 1 2 4 1 3 3 1 1 2 2 1 1 1 1 1 1 1 3
[477] 3 1 4 4 4 4 1 1 1 1 1 2 2 1 1 3 3 3 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 2 3 3 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 3 3 3 1
[545] 2 4 1 1 3 3 6 1 1 1 1 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 1 1 1 1 1 2 2 3 3 3 4 4 4 4 1 1 1 1 1 1 1 3 3 3 1 1 1 1 1 2 2
[613] 2 1 1 1 1 1 1 1 1 2 7 7 7 7 7 7 7 7 1 1 7 1 1 2 2 1 1 7 7 7 7 7 7 1 7 1 1 1 1 1 1 1 1 2 4 4 4 4 4 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 3 3
[681] 1 3 3 3 3 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 3 3 3 1 1 2 2 1 1 1 1 1 1 3
[749] 3 3 1 1 3 1 3 3 2 2 2 2 1 1 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 5 5 1 5 5 5 1 1 1 1
[817] 1 1 1 1 3 3 1 3 1 8 8 8 8 8 8 8 8 1 1 1 1 3 3 1 1 1 2 2 2 2 1 1 3 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 3 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1
[885] 1 2 1 1 1 1 1 1 1 1 3 3 1 1 1 3 4 4 4 4 1 1 1 2 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 1 3 3 5 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1
[953] 1 1 5 5 5 5 5 1 1 2 2 1 1 1 1 1 2 2 2 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```r
titanic_dataset[titanic_dataset$tittle %in% c('Jonkheer','Don', 'Sir'),'tittle'] ='Mr'
titanic_dataset[titanic_dataset$tittle %in% c('Col','Capt', 'Major'),'tittle'] ='Others'
View(titanic_dataset)

#represent using appropriate visualization graph.
library(ggplot2)
table(titanic_dataset$tittle)
ggplot(titanic_dataset,aes(x= tittle)) +
  geom_bar(stat = 'count')  +   labs(x = 'tittle')  + labs(y ='title counts')


#Represent the proportion of people survived from the family size using a graph
titanic_dataset$familysize <-titanic_dataset$sibsp + titanic_dataset$parch + 1
ggplot(titanic_dataset,aes(x= titanic_dataset$familysize,  fill = factor(titanic_dataset$survived ))) +
  geom_bar(stat = 'count')  +   labs(x = 'family size')  + labs(y ='survived')
titanic_dataset$familysize

#Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputati
#install.packages("mice")
library(mice)
set.seed(8)
computed_df = titanic_dataset[, names(titanic_dataset) %in% c('age','sibsp','parch','fare','embarked')]
ageimputed = mice(computed_df, method = "rf", m=5)
imputedage = complete(ageimputed)
par(mfrow=c(1,2))
hist(titanic_dataset$age, main = "Before Imputation", col = "navy blue")
hist(imputedage$age, main = "After Imputation", col = "dark red")
```

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