Progamming in R Assignment 3.1

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.

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

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

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Session 23_Script.R | feature+engineering.R | model+evaluation1.R | Session 21_Script.R | principal component.R* | Assignment3.1.R*

```
setwd("C:/Users/ngonidzashe.gomo/Documents/01. Network Performance/Data Science Training/assignments")
getwd()
mytitanic=read.csv("titanic3.csv")
#a. Preprocess the passenger names to come up with a list of titles thatrepresent families
mytitanic$tittle <-substring(mytitanic$name,regexpr(",",mytitanic$name)+2,regexpr("\\.",mytitanic$name)-1)



library(dplyr)

#Processing tittles
mytitanic[mytitanic$tittle %in% c('Mme'),'tittle'] ='Mrs'
mytitanic[mytitanic$tittle %in% c('Sir'),'tittle'] ='Mr'
mytitanic[mytitanic$tittle %in% c('Ms','Mlle'),'tittle'] ='Miss'
mytitanic[mytitanic$tittle %in% c('Lady','Major','Don','Dona','Capt','Col','Jonkheer',"the Countess"),'tittle'] ='Others

#represent using appropriate visualization graph.
library(ggplot2)
table(mytitanic$tittle)
ggplot(mytitanic,aes(x= mytitanic$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.
mytitanic$familysize <-mytitanic$sibsp + mytitanic$parch + 1
ggplot(mytitanic,aes(x= mytitanic$familysize,  fill = factor(mytitanic$survived ))) +
  geom_bar(stat = 'count')  +   labs(x = 'Family Size')  + labs(y ='Survived')
```

23:1 (Top Level) R Script

Console Terminal

~/01. Network Performance/Data Science Training/assignments/

```
> #a. Preprocess the passenger names to come up with a list of titles thatrepresent families
> mytitanic$tittle <-substring(mytitanic$name,regexpr(",",mytitanic$name)+2,regexpr("\\.",mytitanic$name)-1)
>
>
>
> library(dplyr)
>
> #Processing tittles
> mytitanic[mytitanic$tittle %in% c('Mme'),'tittle'] ='Mrs'
> mytitanic[mytitanic$tittle %in% c('Sir'),'tittle'] ='Mr'
> mytitanic[mytitanic$tittle %in% c('Ms','Mlle'),'tittle'] ='Miss'
> mytitanic[mytitanic$tittle %in% c('Lady','Major','Don','Dona','Capt','Col','Jonkheer',"the Countess"),'tittle'] ='Others'
>
> #represent using appropriate visualization graph.
> library(ggplot2)
> table(mytitanic$tittle)

          Dr Master   Miss     Mr    Mrs Others    Rev
           1      8     61    264    758    198     12      8
> ggplot(mytitanic,aes(x= mytitanic$tittle)) +
+   geom_bar(stat = 'count')  +   labs(x = 'Tittle')  + labs(y ='Tittle Counts')
>
```

Environment | History | Connections

Import Dataset | List

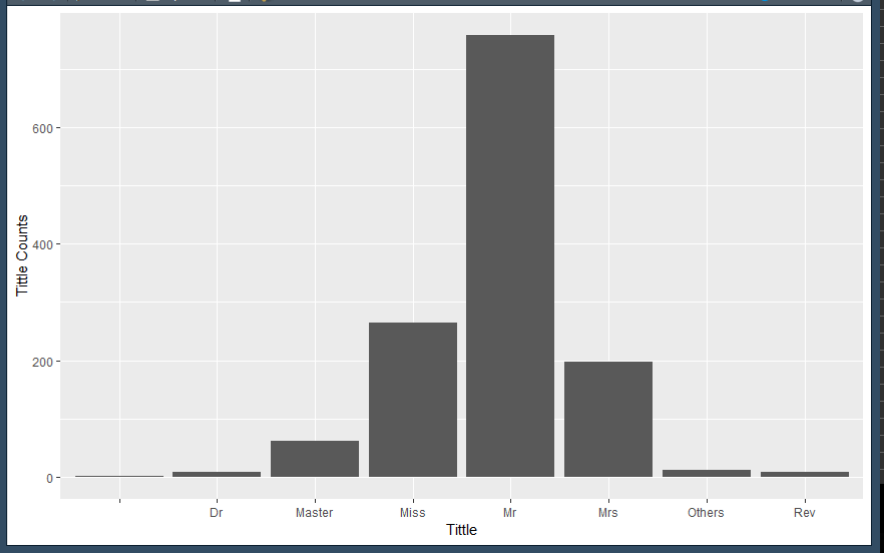
Global Environment

Data

mytitanic 1310 obs. of 15 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Session 23_Script.R | feature+engineering.R | model+evaluation1.R | Session 21_Script.R | principal component.R* | Assignment3.1.R*

Source on Save | Run | Source

```
#represent using appropriate visualization graph.
library(ggplot2)
table(mytitanic$tittle)
ggplot(mytitanic,aes(x= mytitanic$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.
mytitanic$familysize <-mytitanic$sibsp + mytitanic$parch + 1
ggplot(mytitanic,aes(x= mytitanic$familysize,  fill = factor(mytitanic$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=mytitanic[, names(mytitanic) %in% c('age','sibsp','parch','fare','embarked')]
ageimputed = mice(computed_df, method = "rf", m=5)
imputedage = complete(ageimputed)
par(mfrow=c(1,2))
hist(mytitanic$age, main = "Before Imputation", col = "red")
hist(imputedage$age, main = "After Imputation", col = "green")

```

31:80 (Top Level) R Script

Console | Terminal

~/01. Network Performance/Data Science Training/assignments/

```
> #b. Represent the proportion of people survived from the family size using a graph.
> mytitanic$familysize <-mytitanic$sibsp + mytitanic$parch + 1
> ggplot(mytitanic,aes(x= mytitanic$familysize,  fill = factor(mytitanic$survived ))) +
+   geom_bar(stat = 'count')  +   labs(x = 'Family Size')  + labs(y ='Survived')
Warning message:
Removed 1 rows containing non-finite values (stat_count).
>
```

Environment | History | Connections

Import Dataset | List

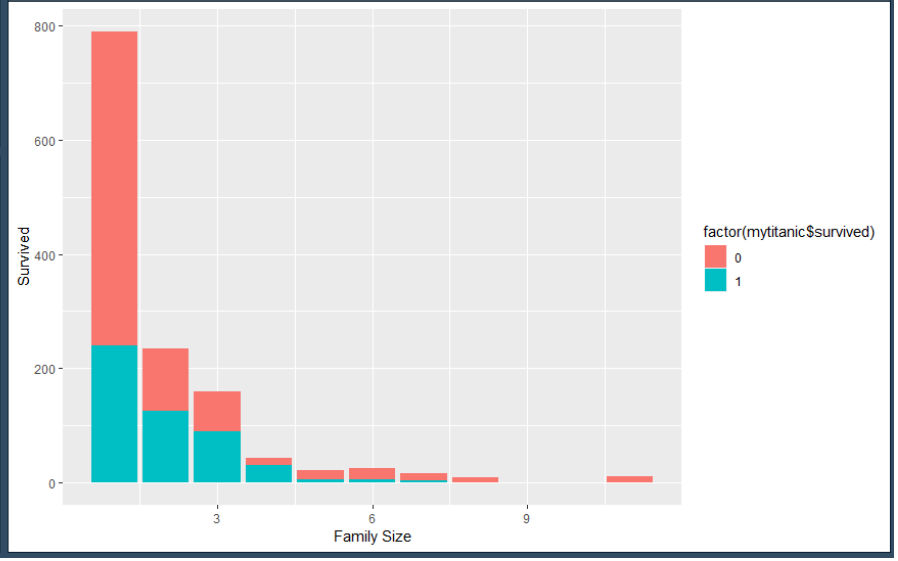
Global Environment

Data

mytitanic 1310 obs. of 16 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Session 23_Script.R | feature+engineering.R | model+evaluation1.R | Session 21_Script.R | principal component.R* | Assignment3.1.R*

```
#represent using appropriate visualization graph.
library(ggplot2)
table(mytitanic$tittle)
ggplot(mytitanic,aes(x= mytitanic$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.
mytitanic$familysize <-mytitanic$sibsp + mytitanic$parch + 1
ggplot(mytitanic,aes(x= mytitanic$familysize,  fill = factor(mytitanic$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=mytitanic[, names(mytitanic) %in% c('age','sibsp','parch','fare','embarked')]
ageimputed = mice(computed_df, method = "rf", m=5)
imputedage = complete(ageimputed)
par(mfrow=c(1,2))
hist(mytitanic$age, main = "Before Imputation", col = "red")
hist(imputedage$age, main = "After Imputation", col = "green")
```

46:1 (Top Level) R Script

Console Terminal

~/01. Network Performance/Data Science Training/assignments/

```
  2   4  age  sibsp  parch  fare
  2   5  age  sibsp  parch  fare
  3   1  age  sibsp  parch  fare
  3   2  age  sibsp  parch  fare
  3   3  age  sibsp  parch  fare
  3   4  age  sibsp  parch  fare
  3   5  age  sibsp  parch  fare
  4   1  age  sibsp  parch  fare
  4   2  age  sibsp  parch  fare
  4   3  age  sibsp  parch  fare
  4   4  age  sibsp  parch  fare
  4   5  age  sibsp  parch  fare
  5   1  age  sibsp  parch  fare
  5   2  age  sibsp  parch  fare
  5   3  age  sibsp  parch  fare
  5   4  age  sibsp  parch  fare
  5   5  age  sibsp  parch  fare
> imputedage = complete(ageimputed)
> par(mfrow=c(1,2))
> hist(mytitanic$age, main = "Before Imputation", col = "red")
> hist(imputedage$age, main = "After Imputation", col = "green")
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data

ageimputed	List of 21
computed_df	1310 obs. of 5 variables
imputedage	1310 obs. of 5 variables
mytitanic	1310 obs. of 16 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish