SESSION 6 – ASSIGNMENT 6.1

Date: 7th Jan 2019

1. **Import the Titanic Dataset from the following link: https://drive.google.com/file/d/1JTJCjdGuUxzKXYlwOavwovB01k6FWg3r/view?ts=5b42ea10**
   Perform the below operations:
   a. **Pre-process 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 by 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**

```
library(readr)
Titanic3 <- read_csv("G:/DATA ANALYTICS/DATA/titanic3.csv")
View(Titanic3)

#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.
head(Titanic3)
tail(Titanic3)
# check structure, as only character vectors can be split using strsplit function
str(Titanic3$Name)
Titanic3$Name<-as.character(Titanic3$Name)
str(Titanic3$Name)
#telling R to call rbind, on two charecters split by strsplit.
#in strsplit, as the data has many " ", and all breaks in many pieces
# hence, using sub() {and not gsub()}, which replaces only first pattern
# so, sub changes first space in ; and the strsplit splits along ; and then rbind binds along
#colums, which is called by do.call
namessplit<-do.call(rbind,strsplit(sub(" ",";",Titanic3$Name),";"))
```

```
head(namessplit)
#converting the charecters to data frame and naming the columns
namessplit<-data.frame(namessplit)
names(namessplit)<-c("family_name", "name")
head(namessplit)
str(namessplit)

#getting title separated from first name
Title<-do.call(rbind,strsplit(sub(" ",";",namessplit$name),";"))
head(Title)
Title<-data.frame(Title)
names(Title)<-c("title", "first_name")
head(Title)
str(Title)
head(Title)
#merging the rownames in titanic survival data to form new data set
#similar to text to columns in excel
#tried merge function which didnt work as expected, but cbind is simpler and gives right data.
str(Titanic3)
TitanicData<-cbind(namessplit,Titanic3)
head(TitanicData)
View(TitanicData)
str(TitanicData)
TitanicData<-cbind(Title,TitanicData)
head(TitanicData)
View(TitanicData)

# There is one more effective way of doing this, and more efficiently
#in the names, we want only titles, i.e Mr or Ms etc.
# names are like this - Braund Mr. Owen Harris
# from these, we need to remove everything after the "."
subtitles<-gsub("\\..*", "", TitanicData$Name) # "\\." is read as ".", one more . after that
indicates one more charecter after that, and * after . (.*) means all charecters post "."
head(subtitles)
# from subtitles, we need to remove everything before title, including space.
Title<-gsub(".*\\ ", "", subtitles) # putting "." before any charecter, here space represented as
"\\ ", selects one charecter before it, and putting * makes it ALL charecters before it.
head(Title)
#graphical representation of the data in various forms
```

#barplot -No. of passangers by Family name

```
familyname<-table(TitanicData$family_name)
View(familyname)
barplot(familyname,main = "survival as per family name", xlab = "family_name", ylab = "count",col ="red")
```

#barplot -No. of passengers by Title

```
Title<-table(Title)
Title
barplot(Title,xlab = "Title", ylab = "No. of Passangers",
      main = "survival as per Title" , col = c("blue", "red"), las=3)
text(Title, 0,table(Title), pos = 3, srt = 90)
```

```
)
> str(Titanic3)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        1309 obs. of  14 variables:
 $ pclass   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ survived : num  1 1 0 0 0 1 1 0 1 0 ...
 $ name     : chr  "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. Helen Loraine" "Allison,
Mr. Hudson Joshua Creighton" ...
 $ sex      : chr  "female" "male" "female" "male" ...
 $ age      : num  29 1 2 30 25 48 63 39 53 71 ...
 $ sibsp    : num  0 1 1 1 1 0 1 0 2 0 ...
 $ parch    : num  0 2 2 2 2 0 0 0 0 0 ...
 $ ticket   : chr  "24160" "113781" "113781" "113781" ...
 $ fare     : num  211 152 152 152 152 ...
 $ cabin    : chr  "B5" "C22 C26" "C22 C26" "C22 C26" ...
 $ embarked : chr  "S" "S" "S" "S" ...
 $ boat     : chr  "2" "11" NA NA ...
 $ body     : num  NA NA NA 135 NA NA NA NA NA 22 ...
 $ home.dest: chr  "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chester
ville, ON" ...
 - attr(*, "spec")=
  .. cols(
  ..   pclass = col_double(),
  ..   survived = col_double(),
  ..   name = col_character(),
  ..   sex = col_character(),
  ..   age = col_double(),
  ..   sibsp = col_double(),
  ..   parch = col_double(),
  ..   ticket = col_character(),
  ..   fare = col_double(),
  ..   cabin = col_character(),
  ..   embarked = col_character(),
  ..   boat = col_character(),
  ..   body = col_double(),
  ..   home.dest = col_character()
  .. )
>
```

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	2
2	1	1	Allison, Master. Hudson Trevor	male	1	1	2	113781	151.5500	C22 C26	S	11
3	1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.5500	C22 C26	S	NA
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.5500	C22 C26	S	NA
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	151.5500	C22 C26	S	NA
6	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500	E12	S	3
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63	1	0	13502	77.9583	D7	S	10
8	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0.0000	A36	S	NA
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2	0	11769	51.4792	C101	S	D
10	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042	NA	C	NA

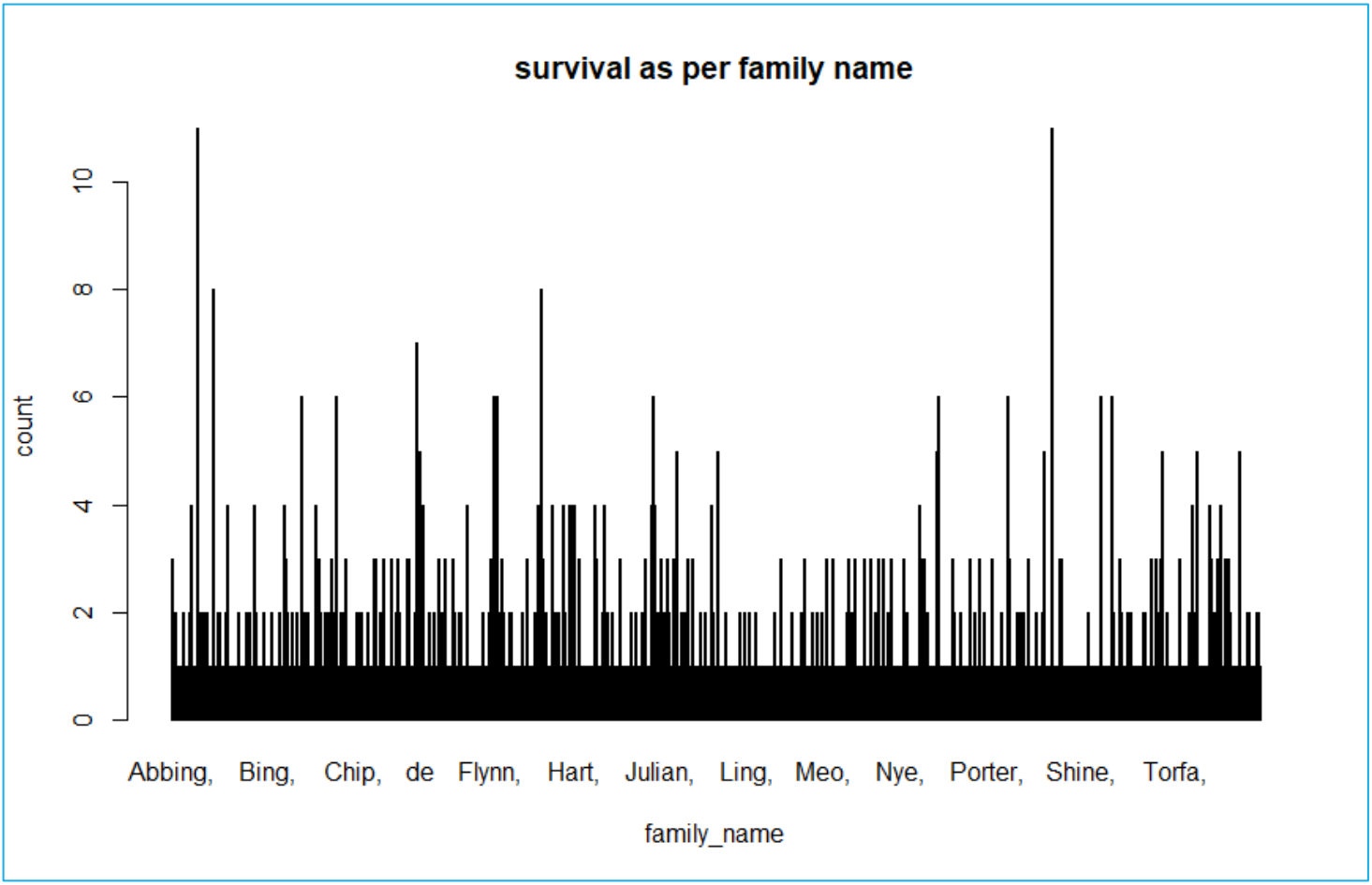
```
  .. )
> View(Titanic3)
> #Perform the following:
> # a. Preprocess the passenger names to come up with a list of titles that represent families
> #and represent using appropriate visualization graph.
> head(Titanic3)
# A tibble: 6 x 14
  pclass survived name                 sex      age sibsp parch ticket  fare cabin embarked boat   body home.dest
   <dbl>    <dbl> <chr>                <chr>  <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <chr>    <chr> <dbl> <chr>
1      1        1 Allen, Miss. Elisabe~ female    29     0     0 24160  211.  B5    S        2        NA St Louis, MO
2      1        1 Allison, Master. Hud~ male       1     1     2 113781 152.  C22 ~ S        11       NA Montreal, PQ / C~
3      1        0 Allison, Miss. Helen~ female     2     1     2 113781 152.  C22 ~ S        NA       NA Montreal, PQ / C~
4      1        0 Allison, Mr. Hudson ~ male      30     1     2 113781 152.  C22 ~ S        NA      135 Montreal, PQ / C~
5      1        0 Allison, Mrs. Hudson~ female    25     1     2 113781 152.  C22 ~ S        NA       NA Montreal, PQ / C~
6      1        1 Anderson, Mr. Harry   male      48     0     0 19952   26.6 E12   S        3        NA New York, NY
> tail(Titanic3)
# A tibble: 6 x 14
  pclass survived name                    sex      age sibsp parch ticket  fare cabin embarked boat   body home.dest
   <dbl>    <dbl> <chr>                   <chr>  <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <chr>    <chr> <dbl> <chr>
1      3        0 Youseff, Mr. Gerious    male      NA     0     0 2627   14.5  NA    C        NA       NA NA
2      3        0 Zabour, Miss. Hileni    female    15     1     0 2665   14.5  NA    C        NA      328 NA
3      3        0 Zabour, Miss. Thamine   female    NA     1     0 2665   14.5  NA    C        NA       NA NA
4      3        0 Zakarian, Mr. Mapriededer male    27     0     0 2656    7.22 NA    C        NA      304 NA
5      3        0 Zakarian, Mr. Ortin     male      27     0     0 2670    7.22 NA    C        NA       NA NA
6      3        0 Zimmerman, Mr. Leo      male      29     0     0 315082  7.88 NA    S        NA       NA NA
> str(Titanic3$name) # check structure, as only charecter vectors can be split using strsplit function
 chr [1:1309] "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. Helen Loraine" ...
> Titanic3$name<-as.character(Titanic3$name)
> str(Titanic3$name)
 chr [1:1309] "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. Helen Loraine" ...
> #telling R to call rbind, on two charecters split by strsplit.
> #in strsplit, as the data has many " ", and all breaks in many pieces
> # hence, using sub() {and not gsub()}, which replaces only first pattern
> # so, sub changes first space in ; and the strsplit splits along ; and then rbind binds along colums, which is called by do
.call
> namessplit<-do.call(rbind,strsplit(sub(" ",";",Titanic3$name),";"))
>
```

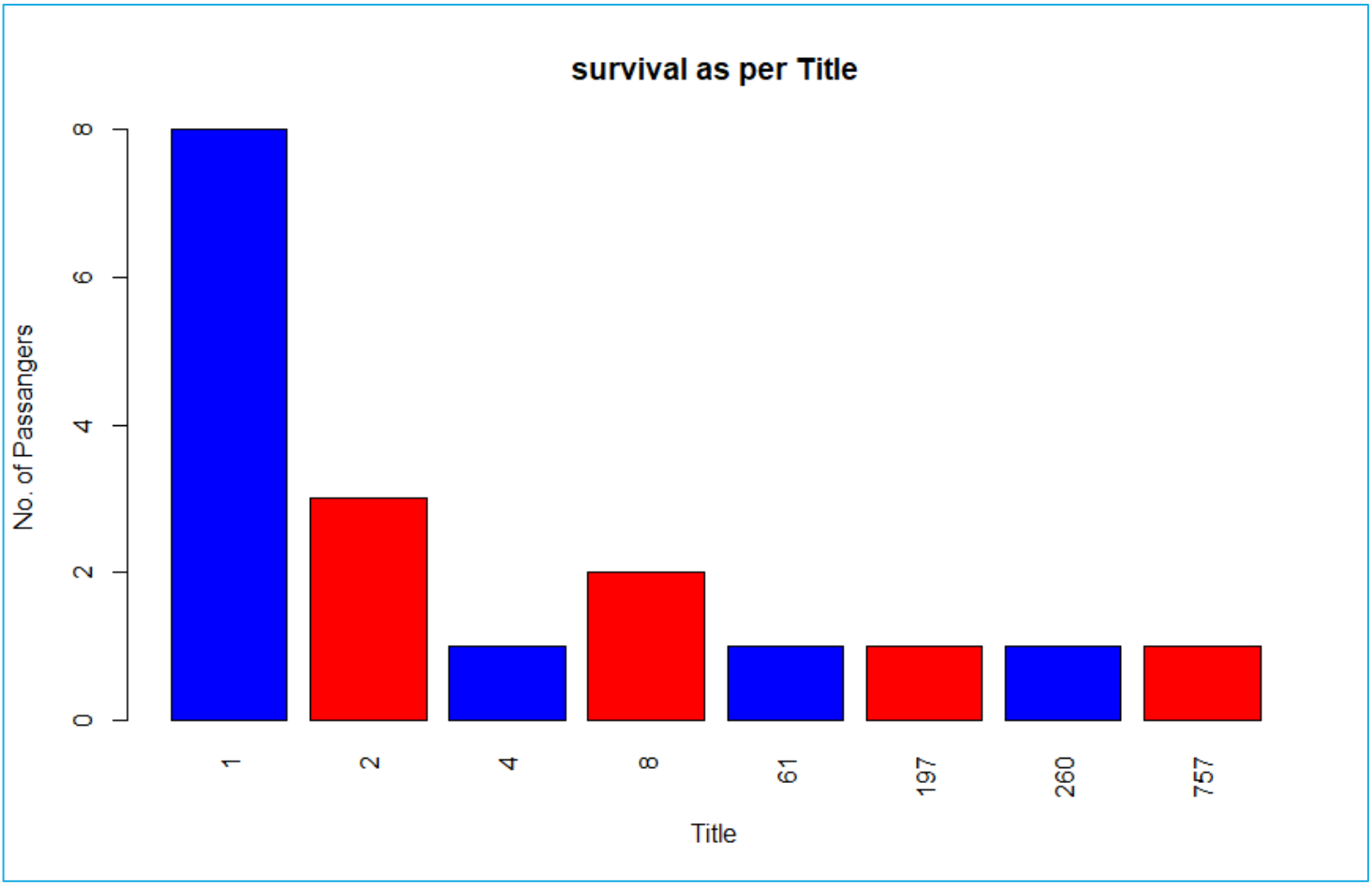
```
Console  Terminal
~/
> # so, sub changes first space in ; and the strsplit splits along ; and then rbind binds along colums, which is called by do
.call
> namessplit<-do.call(rbind,strsplit(sub(" ",";",Titanic3$name),";"))
> head(namessplit)
     [,1]         [,2]
[1,] "Allen,"     "Miss. Elisabeth Walton"
[2,] "Allison,"   "Master. Hudson Trevor"
[3,] "Allison,"   "Miss. Helen Loraine"
[4,] "Allison,"   "Mr. Hudson Joshua Creighton"
[5,] "Allison,"   "Mrs. Hudson J C (Bessie Waldo Daniels)"
[6,] "Anderson," "Mr. Harry"
> #converting the charecters to data frame and naming the columns
> namessplit<-data.frame(namessplit)
> names(namessplit)<-c("family_name", "name")
> head(namessplit)
  family_name                                name
1      Allen,                Miss. Elisabeth Walton
2    Allison,                 Master. Hudson Trevor
3    Allison,                   Miss. Helen Loraine
4    Allison,           Mr. Hudson Joshua Creighton
5    Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
6   Anderson,                             Mr. Harry
> str(namessplit)
'data.frame':   1309 obs. of  2 variables:
 $ family_name: Factor w/ 868 levels "Abbing,","Abbott,",..: 16 17 17 17 17 21 25 25 28 31 ...
 $ name       : Factor w/ 1144 levels "Billiard, Master. James William",..: 159 64 194 575 1019 549 232 864 982 792 ...
> #getting title separated from first name
> Title<-do.call(rbind,strsplit(sub(" ",";",namessplit$name),";"))
> head(Title)
     [,1]      [,2]
[1,] "Miss."   "Elisabeth Walton"
[2,] "Master." "Hudson Trevor"
[3,] "Miss."   "Helen Loraine"
[4,] "Mr."     "Hudson Joshua Creighton"
[5,] "Mrs."    "Hudson J C (Bessie Waldo Daniels)"
[6,] "Mr."     "Harry"
>
```

```
> str(namessplit)
'data.frame':   1309 obs. of  2 variables:
 $ family_name: Factor w/ 868 levels "Abbing,","Abbott,",..: 16 17 17 17 17 21 25 25 28 31 ...
 $ name       : Factor w/ 1144 levels "Billiard, Master. James William",..: 159 64 194 575 1019 549 232 864 982 792 ...
> #getting title separated from first name
> Title<-do.call(rbind,strsplit(sub(" ",";",namessplit$name),";"))
> head(Title)
     [,1]      [,2]
[1,] "Miss."   "Elisabeth Walton"
[2,] "Master." "Hudson Trevor"
[3,] "Miss."   "Helen Loraine"
[4,] "Mr."     "Hudson Joshua Creighton"
[5,] "Mrs."    "Hudson J C (Bessie Waldo Daniels)"
[6,] "Mr."     "Harry"
> Title<-data.frame(Title)
> names(Title)<-c("title", "first_name")
> head(Title)
    title                        first_name
1   Miss.                   Elisabeth Walton
2 Master.                      Hudson Trevor
3   Miss.                      Helen Loraine
4     Mr.            Hudson Joshua Creighton
5    Mrs. Hudson J C (Bessie Waldo Daniels)
6     Mr.                              Harry
> str(Title)
'data.frame':   1309 obs. of  2 variables:
 $ title     : Factor w/ 34 levels "Billiard,","Brito,",..: 18 15 18 21 22 21 18 21 22 21 ...
 $ first_name: Factor w/ 1127 levels "(Ada E Hall)",..: 298 508 465 507 506 456 672 1025 271 909 ...
> head(Title)
    title                        first_name
1   Miss.                   Elisabeth Walton
2 Master.                      Hudson Trevor
3   Miss.                      Helen Loraine
4     Mr.            Hudson Joshua Creighton
5    Mrs. Hudson J C (Bessie Waldo Daniels)
6     Mr.                              Harry
>
```

```
, Mr. Hudson Joshua Creighton" ...
 $ sex        : chr  "female" "male" "female" "male" ...
 $ age        : num  29 1 2 30 25 48 63 39 53 71 ...
 $ sibsp      : num  0 1 1 1 1 0 1 0 2 0 ...
 $ parch      : num  0 2 2 2 2 0 0 0 0 0 ...
 $ ticket     : chr  "24160" "113781" "113781" "113781" ...
 $ fare       : num  211 152 152 152 152 ...
 $ cabin      : chr  "B5" "C22 C26" "C22 C26" "C22 C26" ...
 $ embarked   : chr  "S" "S" "S" "S" ...
 $ boat       : chr  "2" "11" NA NA ...
 $ body       : num  NA NA NA 135 NA NA NA NA NA 22 ...
 $ home.dest  : chr  "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chest
erville, ON" ...
> TitanicData<-cbind(Title,TitanicData)
> head(TitanicData)
    title                        first_name family_name                                name pclass survived
1   Miss.                   Elisabeth Walton      Allen,                Miss. Elisabeth Walton      1        1
2 Master.                      Hudson Trevor    Allison,                 Master. Hudson Trevor      1        1
3   Miss.                      Helen Loraine    Allison,                   Miss. Helen Loraine      1        0
4     Mr.            Hudson Joshua Creighton    Allison,            Mr. Hudson Joshua Creighton      1        0
5    Mrs. Hudson J C (Bessie Waldo Daniels)    Allison, Mrs. Hudson J C (Bessie Waldo Daniels)      1        0
6     Mr.                              Harry   Anderson,                           Mr. Harry      1        1
                                      name    sex age sibsp parch ticket     fare   cabin embarked boat body
1                  Allen, Miss. Elisabeth Walton female  29     0     0  24160 211.3375      B5        S    2   NA
2                 Allison, Master. Hudson Trevor   male   1     1     2 113781 151.5500 C22 C26        S   11   NA
3                   Allison, Miss. Helen Loraine female   2     1     2 113781 151.5500 C22 C26        S <NA>   NA
4           Allison, Mr. Hudson Joshua Creighton   male  30     1     2 113781 151.5500 C22 C26        S <NA>  135
5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female  25     1     2 113781 151.5500 C22 C26        S <NA>   NA
6                         Anderson, Mr. Harry   male  48     0     0  19952  26.5500     E12        S    3   NA
                     home.dest
1                   St Louis, MO
2 Montreal, PQ / Chesterville, ON
3 Montreal, PQ / Chesterville, ON
4 Montreal, PQ / Chesterville, ON
5 Montreal, PQ / Chesterville, ON
6                   New York, NY
>
```

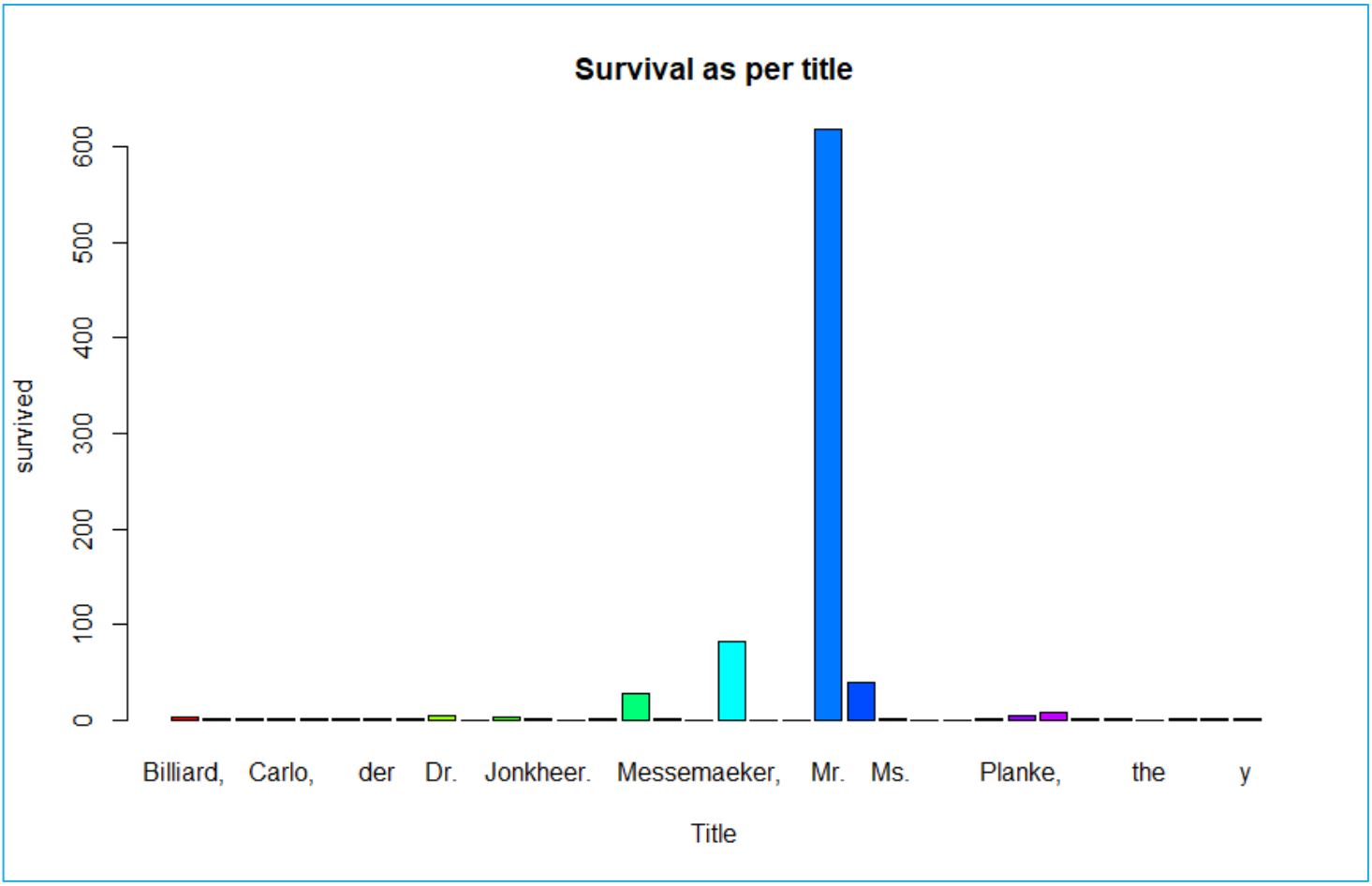
	title	first_name	family_name	name	pclass	survived	name
1	Miss.	Elisabeth Walton	Allen,	Miss. Elisabeth Walton	1	1	Allen, Miss.
2	Master.	Hudson Trevor	Allison,	Master. Hudson Trevor	1	1	Allison, Mas
3	Miss.	Helen Loraine	Allison,	Miss. Helen Loraine	1	0	Allison, Mis
4	Mr.	Hudson Joshua Creighton	Allison,	Mr. Hudson Joshua Creighton	1	0	Allison, Mr.
5	Mrs.	Hudson J C (Bessie Waldo Daniels)	Allison,	Mrs. Hudson J C (Bessie Waldo Daniels)	1	0	Allison, Mrs
6	Mr.	Harry	Anderson,	Mr. Harry	1	1	Anderson, M
7	Miss.	Kornelia Theodosia	Andrews,	Miss. Kornelia Theodosia	1	1	Andrews, M
8	Mr.	Thomas Jr	Andrews,	Mr. Thomas Jr	1	0	Andrews, M
9	Mrs.	Edward Dale (Charlotte Lamson)	Appleton,	Mrs. Edward Dale (Charlotte Lamson)	1	1	Appleton, M
10	Mr.	Ramon	Artagaveytia,	Mr. Ramon	1	0	Artagaveytia

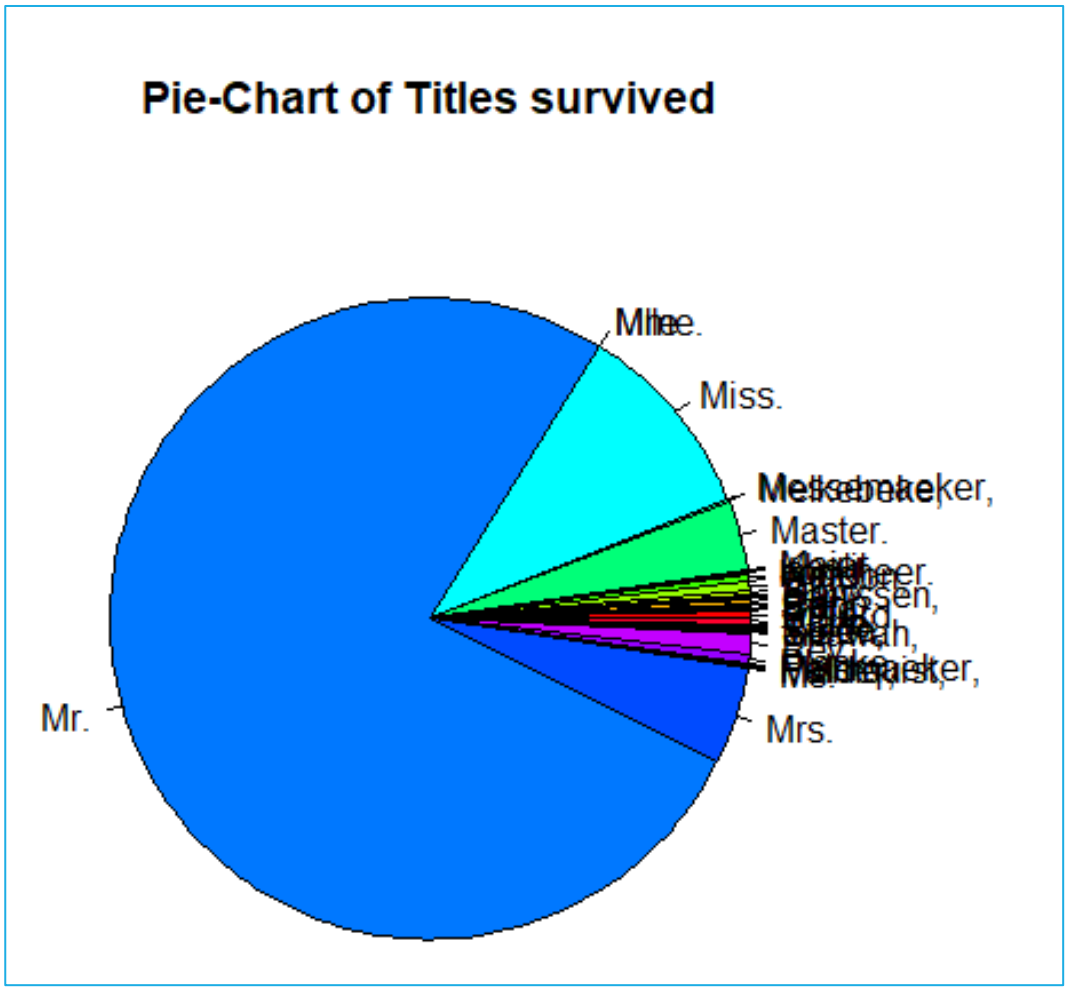


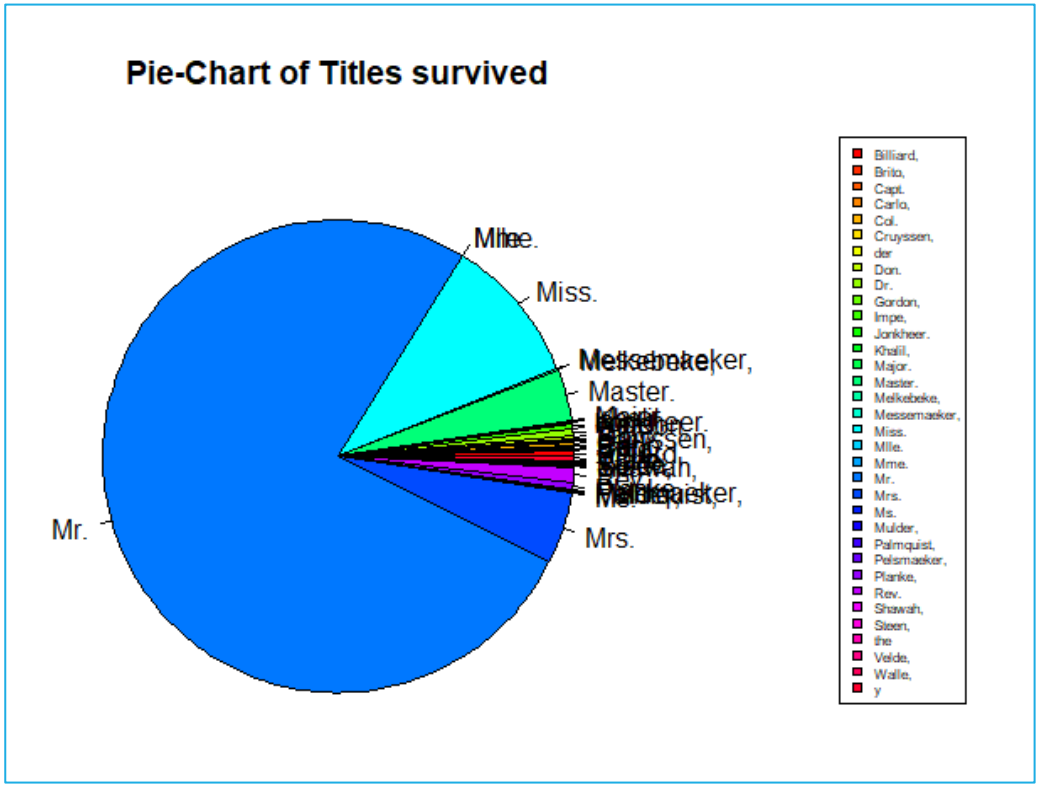


 # b. Represent the proportion of people survived from the family title using a graph.

```
View(TitanicData)
SurvivedTitle<-table(TitanicData$Survived, TitanicData$title)
#survived is 0, first row. we will take only that
p<-SurvivedTitle[1,]

#barplot of survived numbers per title
barplot(p,xlab = "Title", ylab = "survived",
      main= "Survival as per title", col=rainbow(length(p)))
#pie chart showing proportion of survival title wise

pie_chart<-pie(p, main = "Pie-Chart of Titles survived", col = rainbow(length(p)) )
legend("topright", names(p), cex= 0.5, fill = rainbow(length(p)))
```

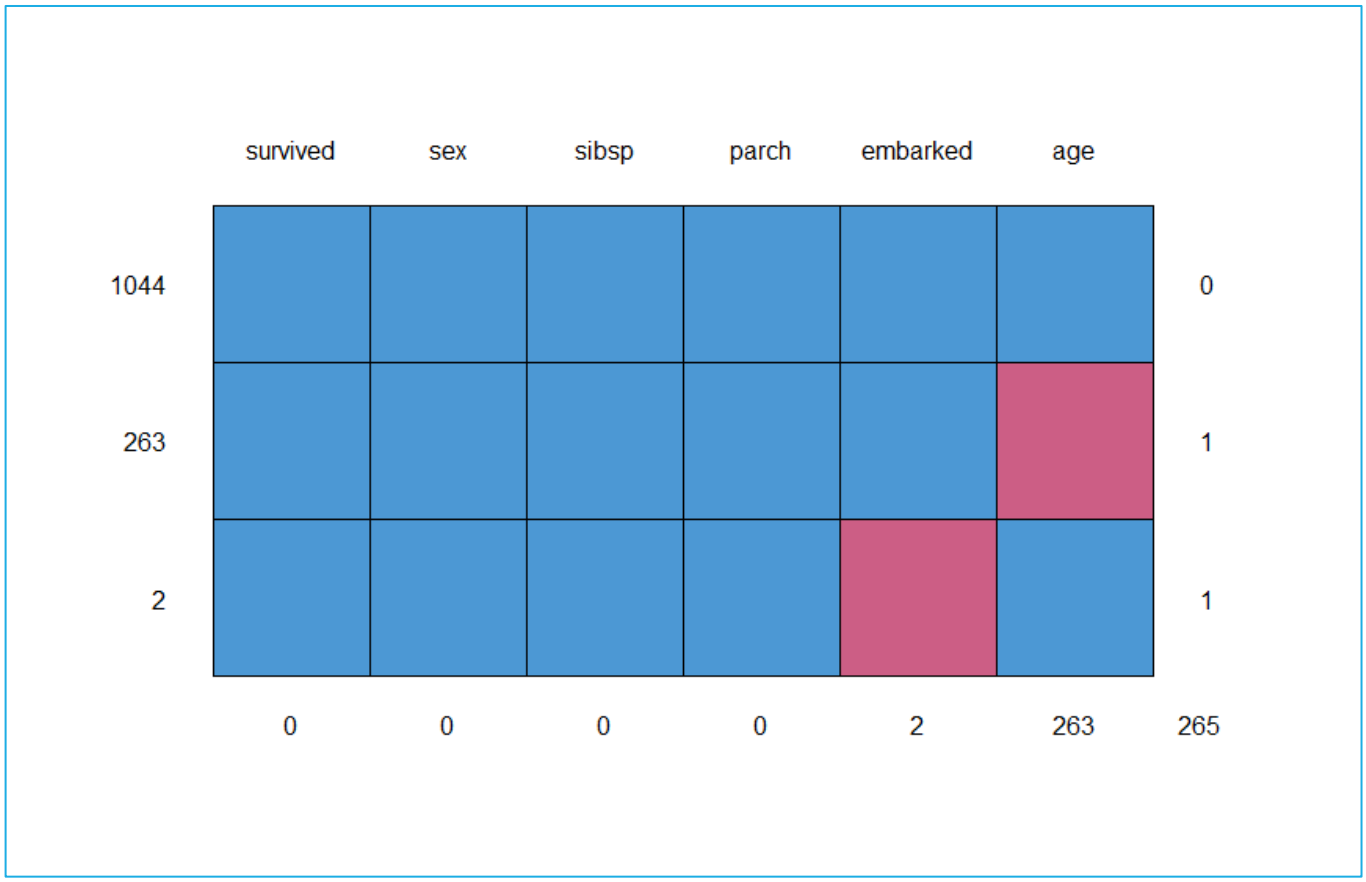

Survival as per title
survived
0
100
200
300
400
500
600
Billiard,
Carlo,
der
Dr.
Jonkheer.
Messemaeker,
Mr.
Ms.
Planke,
the
y
Title


Pie-Chart of Titles survived
Miss.
Master.
Mr.
Mrs.

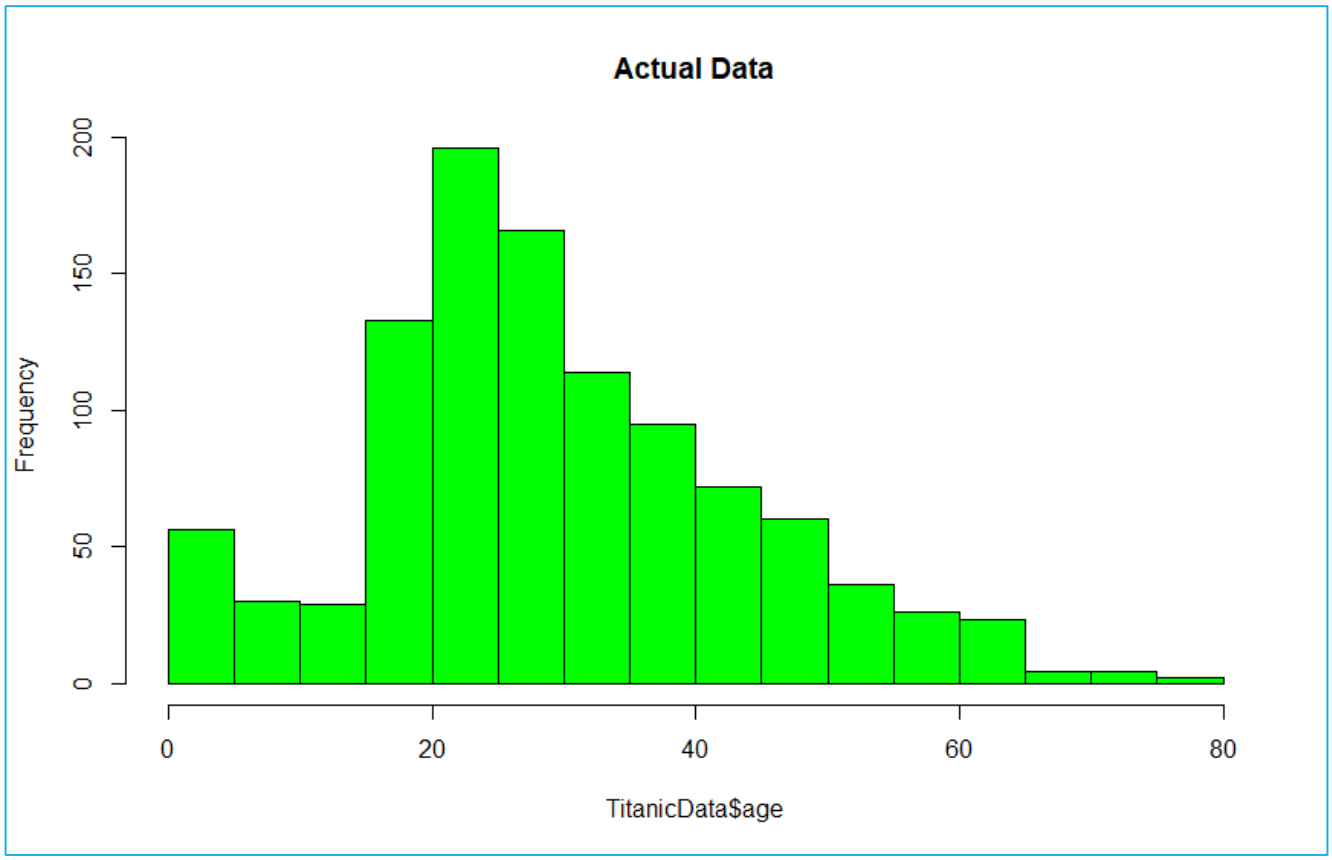

Pie-Chart of Titles survived
Miss.
Master.
Mr.
Mrs.
Billiard,
Brito,
Capt.
Carlo,
Col.
Cruyssen,
der
Don.
Dr.
Gordon,
Impe,
Jonkheer.
Khalil,
Major.
Master.
Melkebeke,
Messemaeker,
Miss.
Mlle.
Mme.
Mr.
Mrs.
Ms.
Mulder,
Palmquist,
Pelsmaeker,
Planke,
Rev.
Shawah,
Steen,
the
Velde,
Walle,
y

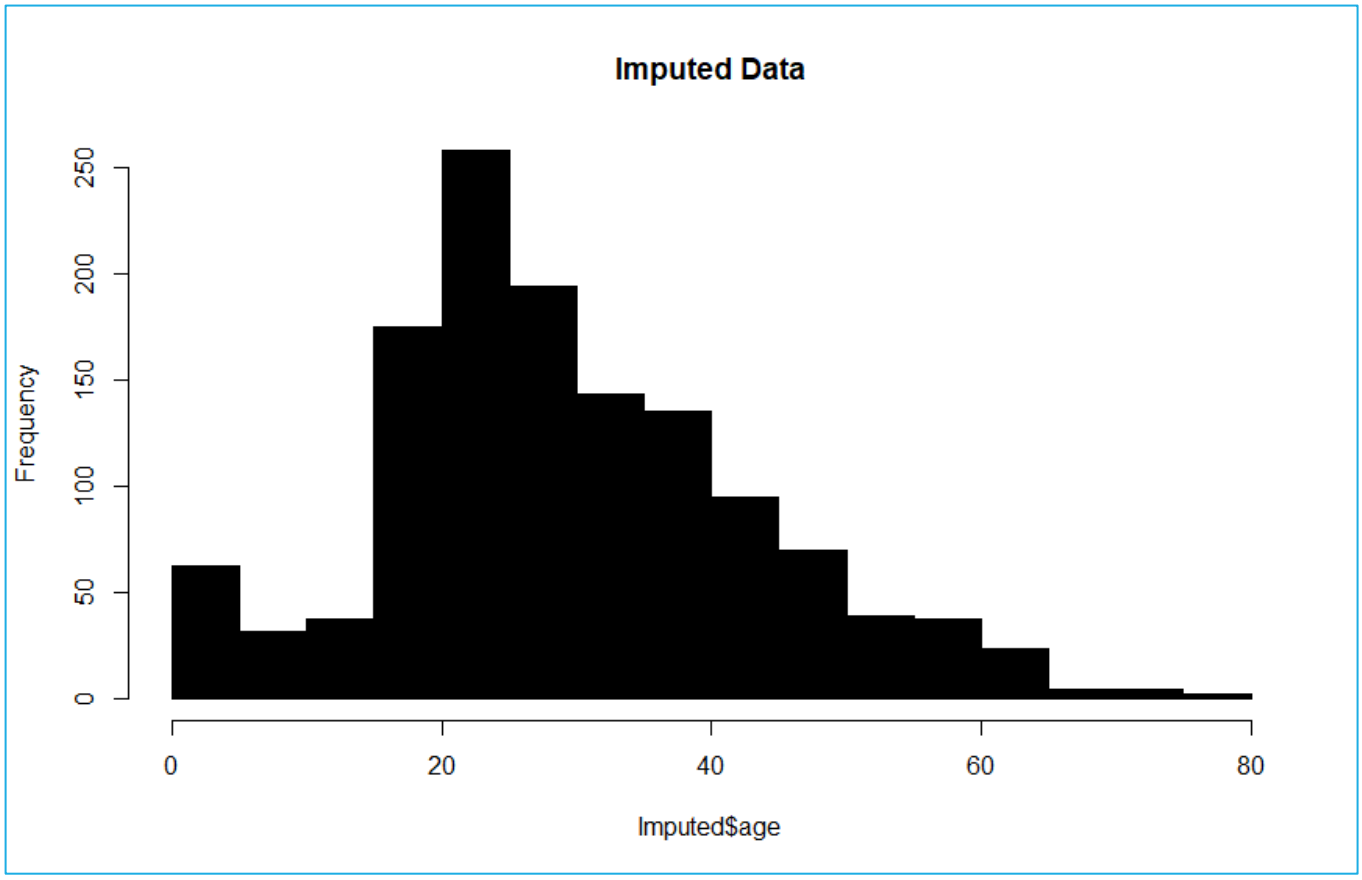
 # c. Impute the missing values in Age variable using Mice Library, create two different #graphs showing Age distribution before and after imputation.

```
library(mice)
sum(is.na(TitanicData$age))
str(TitanicData)

#Removing columns 1,2,3,4,5,7,12,13,14,16,17,18

mini_data <- TitanicData[-c(1,2,3,4,5,7,12,13,14,16,17,18)]
View(mini_data)
md.pattern(mini_data)

library(dplyr)
mini_data <- mini_data %>%
  mutate(
    survived = as.factor(survived),
    sex = as.factor(sex),
    age = as.numeric(age),
    sibsp = as.factor(sibsp),
    parch = as.factor(parch),
    embarked = as.factor(embarked)
  )
str(mini_data)
mice_data <- mice(mini_data, m=5, maxit=50,seed=500)
summary(mini_data)
Imputed=complete(mice_data,5)
hist(TitanicData$age,  main='Actual Data',col="green")
hist(Imputed$age, main='Imputed Data',col="black")
```


survived
sex
sibsp
parch
embarked
age
1044
263
2
0
1
1
0 0 0 0 2 263 265

MD. PATTERN


Actual Data
Frequency
0
50
100
150
200
0
20
40
60
80
TitanicData$age


Imputed Data
Frequency
0
50
100
150
200
250
0
20
40
60
80
Imputed$age

Console Terminal ×

~/

```
iter imp variable
 1   1  age  embarked
 1   2  age  embarked
 1   3  age  embarked
 1   4  age  embarked
 1   5  age  embarked
 2   1  age  embarked
 2   2  age  embarked
 2   3  age  embarked
 2   4  age  embarked
 2   5  age  embarked
 3   1  age  embarked
 3   2  age  embarked
 3   3  age  embarked
 3   4  age  embarked
 3   5  age  embarked
 4   1  age  embarked
 4   2  age  embarked
 4   3  age  embarked
 4   4  age  embarked
 4   5  age  embarked
 5   1  age  embarked
 5   2  age  embarked
 5   3  age  embarked
 5   4  age  embarked
 5   5  age  embarked
 6   1  age  embarked
 6   2  age  embarked
 6   3  age  embarked
 6   4  age  embarked
 6   5  age  embarked
 7   1  age  embarked
 7   2  age  embarked
 7   3  age  embarked
 7   4  age  embarked
 7   5  age  embarked
 8   1  age  embarked
```