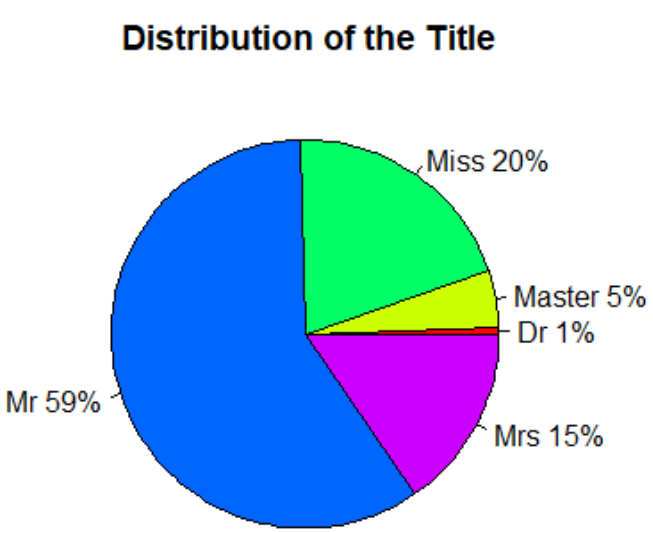
Assignment6.1

Aruna

12 November 2018

```
#assignment 6.2

 # 1. a Pre process  the passanger names to come up with a list of titles
 #      that reperesents families and represent using appropriate
 #      Visualization chart

 #import the readxl package
install.packages("readxl")
 library(readxl)
 #read the titanic 3 table
 titanic3 =read_excel("F:/R Notes/Assignments/titanic3.xls")
 View(titanic3)

d$Title<-regmatches(as.character(titanic3$name),regexpr("\\,[A-z ]{1,20}\\.",
as.character(titanic3$name)))
#d$Title
# check how many titles are present
d$Title<-unlist(lapply(d$Title,FUN=function(x) substr(x, 3, nchar(x)-1)))
table(d$Title)
#Merge the Title table to the new existing titanic tbale
titanic3_title   <- cbind(titanic3, d)
#View(titanic3_title)
titanic3_title$Title=as.factor(titanic3_title$Title)



# group the names
d$Title[which(d$Title %in% c("Mme", "Mlle"))] <- "Miss"
d$Title[which(d$Title %in% c("Lady", "Ms", "the Countess", "Dona"))] <- "Mrs"
d$Title[which(d$Title=="Dr" & d$Sex=="female")] <- "Mrs"
d$Title[which(d$Title=="Dr" & d$Sex=="male")] <- "Mr"
d$Title[which(d$Title %in% c("Capt", "Col", "Don", "Jonkheer", "Major",
"Rev", "Sir"))] <- "Mr"
d$Title<-as.factor(d$Title)

#d$Title
#move the title to new vector
t = table(d$Title)
# to the table we need the percentages
prcnt = round(t/sum(t)*100)
group = c("Dr","Master","Miss","Mr","Mrs")
#use paste function to bind both the lable name and percentage
```

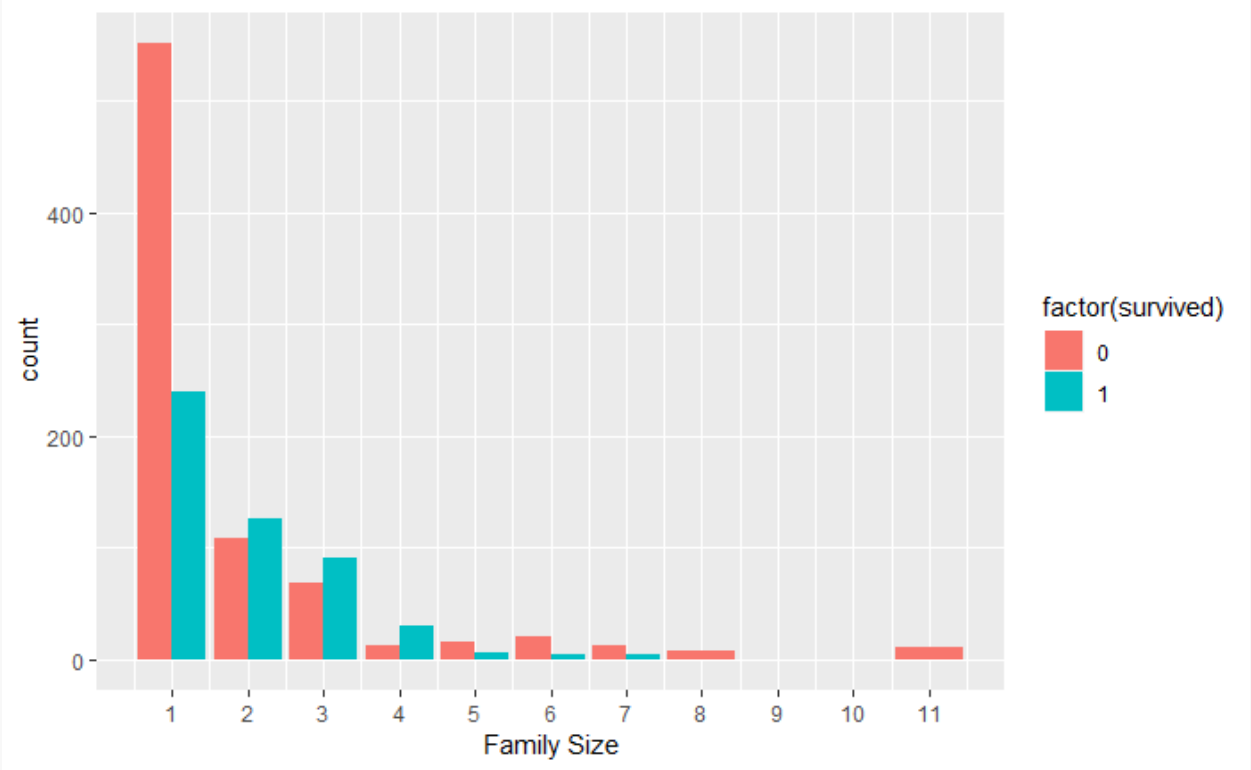
```
grouped = paste(group, prcnt)

# use a seperator and add the symbol %
label = paste(grouped, "%", sep = "")
#draw a pie chart
pie(t, labels = label,
    main = " Distribution of the Title",
    col = rainbow(length(label))
    )
```



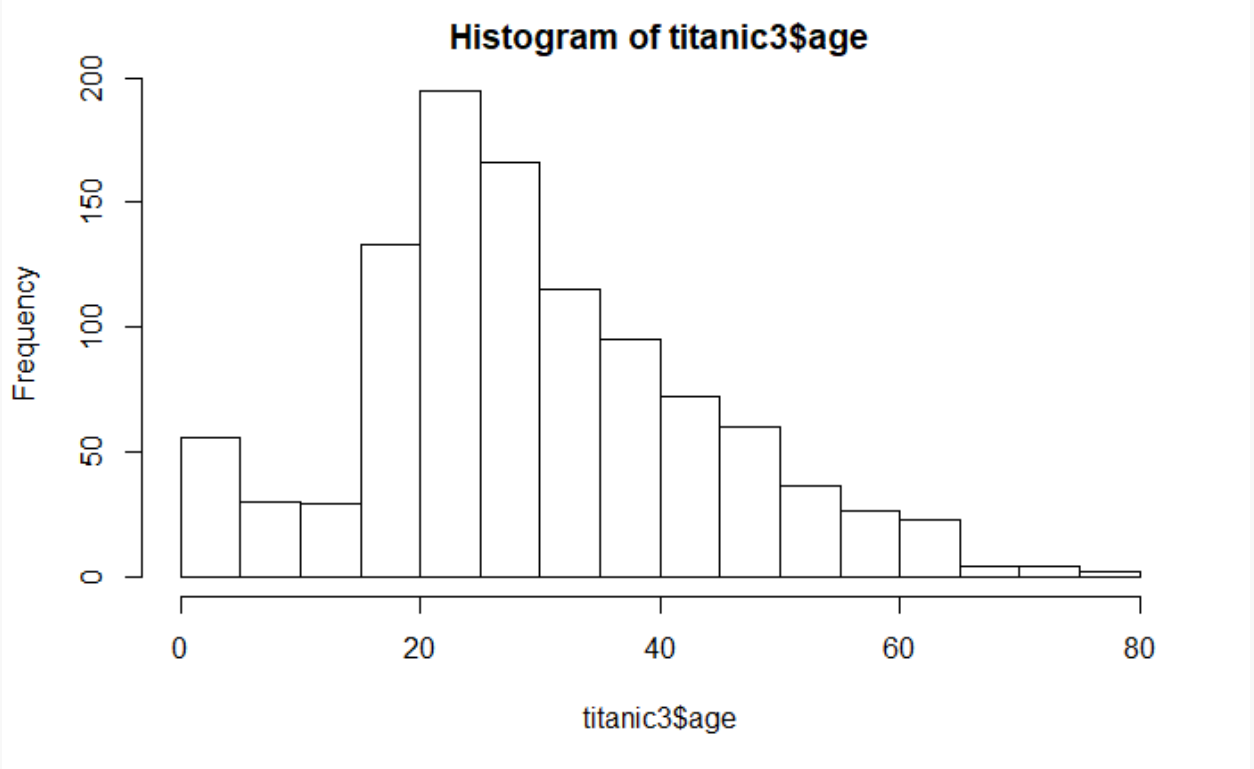

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

```
# 1. b. Represent  the proportion of people survived by family size
#       using a graph


# Create a family size variable including the passenger themselves
titanic3_title$Fsize <- titanic3$sibsp + titanic3$parch + 1

# Create a new variable that shows the family name and the family size
titanic3_title$Family <- paste(titanic3_title$Title, titanic3_title$Fsize,
sep = "_")
#View(titanic3_title)

#install.packages("ggplot2")
library(ggplot2)
library(scales) # visualization
```

```
library(dplyr) # data manipulation
library(mice) # imputation


ggplot(titanic3_title[1:1310,], aes(x = Fsize, fill = factor(survived))) +
  geom_bar(stat='count', position='dodge') +
  scale_x_continuous(breaks=c(1:11)) +
  labs(x = 'Family Size')
```




```
#1. c Impute the missing values in Age Variable using Mice Library
#     create two different graphs showing Age Distribution before
#     and after imputation
hist(titanic3$age)
# how many rows has the age filled
sum(complete.cases(titanic3$age))
#number of rows missing the age parameter
sum(!complete.cases(titanic3$age))
summary(titanic3$age)
# Imputation is still not taught
```


Histogram of titanic3$age
Frequency
0
50
100
150
200
0
20
40
60
80
titanic3$age

R Markdown