**Assignment 3\_4**

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.

**Answer:**

#creating new dataframe with the separate col for lastname and firstname from titanic df

install.packages("tidyr")

library(tidyr)

install.packages("dplyr")

library(dplyr)

titan\_train=separate(titanic\_train, Name,c("lastname","firstname"), ", ")

#to get the df with

library(sqldf)

titan\_lastname=sqldf("select lastname, count(PassengerId) from titan\_train where parch>0 or sibsp>0 group by lastname")

#convert to df

titan\_lastname\_df=as.data.frame(titan\_lastname)

#draw word cloud of list of titles having families

library(wordcloud)

library("tm")

set.seed(1234)

wordcloud(words = titan\_lastname\_df$lastname, freq = titan\_lastname\_df$`count(PassengerId)`, min.freq = 1, max.words=Inf,scale=c(2,.3), random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))

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

**Answer:**

titan\_lastname=sqldf("select lastname, count(PassengerId), sum(Survived) from titan\_train where parch>0 or sibsp>0 group by lastname")

barplot(table(titan\_lastname\_df$`sum(Survived)`, titan\_lastname\_df$`count(PassengerId)`), legend.text = TRUE,beside = TRUE )

c. Impute the missing values in Age variable using Mice Library, create two different

graphs showing Age distribution before and after imputation.