Assignment 6.1

Name : Y Vasudev

Batch : DA with R , Excel and Tableau

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.

a)

pass\_titles <- ""

for(i in 1:891){

pass\_titles <- c(pass\_titles,pass\_names[[i]][2])

}

pass\_titles <- pass\_titles[-1]

titles <- ""

for( i in 1:891){

temp <- strsplit(pass\_titles[i],"\\.")

titles <- c(titles,temp)

}

titles[1] <- NULL

title\_category <- ""

for(i in 1:891){

title\_category <- c(title\_category,titles[[i]][1])

}

title\_category <- title\_category[-1]

title\_category <- as.factor(title\_category)

levels(title\_category)

#check for summary

summary(title\_category)

others <- 0

for(i in 1:17){

if(i == 8 || i == 9 || i == 12 || i == 13 )

{

next()

}

else{

others <- others + summary(title\_category)[i]

}

}

#grouping

# We have grouped small categories such as Capt , Col ,Don.....etc into 'others'

library(plotrix)

title\_category <- c(summary(title\_category)[8],summary(title\_category)[9],summary(title\_category)[12],summary(title\_category)[13],others)

labelname <- c("Master","Miss","Mr.","Mrs.","Others")

pie3D(title\_category,labels = labelname,main="List of Titles")

b)

pass\_names <- ""

for( i in 1:891){

temp <- strsplit(titanic\_train$Name[i],",")

pass\_names <- c(pass\_names,temp)

}

pass\_names[[1]] <- NULL

pass\_surnames <- ""

for(i in 1:891){

pass\_surnames <- c(pass\_surnames,pass\_names[[i]][1])

}

pass\_surnames <- pass\_surnames[-1]

class(pass\_surnames)

pass\_surnames <- as.factor(pass\_surnames)

pass\_surnames\_list <- levels(pass\_surnames)

pass\_surnames\_list\_count <- rep(0,times=667)

count <- 0

for(i in 1:667){

n1 <- pass\_surnames\_list[i]

for(j in 1:891){

n2 <- pass\_surnames[j]

if(n1 == n2)

count <- count + 1

}

pass\_surnames\_list\_count[i] <- count

count <- 0

}

pass\_surnames\_list\_count <- as.factor(pass\_surnames\_list\_count)

family\_size <- levels(pass\_surnames\_list\_count)

#check for summary

#summary(pass\_surnames\_list\_count)

family\_size\_count <- table(pass\_surnames\_list\_count)

titanic\_survived <- data.frame(titanic\_train$Survived,pass\_surnames)

count<- rep(0,times=891)

titanic\_survived <- data.frame(titanic\_survived,count)

colnames(titanic\_survived) <- c("survived","surnames","family\_size")

for(i in 1:891){

n <- titanic\_survived$surnames[i]

for(j in 1:667){

if(n == pass\_surnames\_list[j]){

titanic\_survived$family\_size[i] <- pass\_surnames\_list\_count[j]

break

}

}

}

s1 <- table(titanic\_survived$survived,titanic\_survived$family\_size)

barplot(s1,xlab="Family Size",col=c("red","green"),legend=c("Not Survived","Survived"),beside = T,main = "Proportion of People Survived")

c)

library(titanic)

sum(is.na(titanic\_train$Age))

install.packages("mice")

library(mice)

md.pattern(titanic\_train)

# We found there are total 177 missing values in AGE attribute of titanic\_train dataset

mice\_imputes = mice(titanic\_train, m=5, maxit = 40)

Imputed\_data=complete(mice\_imputes,5)

hist(titanic\_train$Age, freq=T, main='Original Data of age ',

col='green')

hist(Imputed\_data$Age, freq=T, main='Imputed Data of age',

col="red")