**NAME: SANDEEP PATIL**

**Email: saashasandeep@gmail.com**

**DATA ANALYTICS WITH R, EXCEL and TABLEAU**

**Session 6 – Assignment – 6.1**

1. Import the Titanic Dataset from the link Titanic Data Set.

**Answer:** TitanicData<-read.csv("G:\\LEARNING\\ACADGILD DATA ANALYTICS\\DATA SETS SANDEEP\\TitanicData.txt",header=F,sep="," )

colnames(TitanicData)<-c("PassengerId",

"Survived",

"Pclass",

"Name",

"Sex",

"Age",

"SibSp",

"Parch",

"Ticket",

"Fare",

"Cabin",

"Embarked")

View(TitanicData)

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(TitanicData)

tail(TitanicData)

str(TitanicData$Name) # check structure, as only charecter vectors can be split using strsplit function

TitanicData$Name<-as.character(TitanicData$Name)

str(TitanicData$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(" ",";",TitanicData$Name),";"))

head(namessplit)

#converting the charecters to data frame and naming the columns

namessplit<-data.frame(namessplit)

names(namessplit)<-c("family name", "first name")

head(namessplit)

str(namessplit)

#getting title separated from first name

Title<-(do.call(rbind,strsplit(TitanicData$Name)," "))[,2]

table(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(TitanicData)

TitanicData<-cbind(namessplit,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`)

barplot(familyname,main = "survival as per family name", xlab = "family name", ylab = "count",col ="red")

#barplot -No. of passangers 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)

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

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 chaart 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)))

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

graphs showing Age distribution before and after imputation.

library(titanic)  
sum([is.na](http://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")