SESSION 6

ASSIGNMENT 1

# Import Data Set ; Titanic

TitanicData <- read.csv("c:/hema/Data Analytics/titanicdata.txt", header=FALSE)

View(titanicdata)

str(titanicdata)

psych::describe(titanicdata)

colnames(titanicdata) <- c("PassengerId","Survived","Pclass","Name",

"Sex","Age","SibSp","Parch","Ticket","Fare",

"Cabin","Embarked")

TitanicData <- titanicdata[,-13]

TitanicData$Survived <- as.factor(titanicdata$Survived)

TitanicData$Pclass <- as.factor(titanicdata$Pclass)

TitanicData$SibSp <- as.factor(titanicdata$SibSp)

TitanicData$Parch <- as.factor(titanicdata$Parch)

str(titanicdata)

1. Preprocess the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.

# Convert Name as character

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

# Grab title from passenger names

titanicdata$SubTitle <- gsub("\\..\*", "", titanicdata$Name)

titanicdata$Title <- gsub(".\*\\ ", "", titanicdata$SubTitle)

table(titanicdata$Title) # Count of Titles

# 1. Number of Passangers by Title

Solution:

Title <- barplot(table(titanicdata$Title),

main = "No. of Passangers by Title", xlab = "Title",

ylab = "No. of Passangers", col = "Blue", las =3)

text(Title, 0,table(titanicdata$Title), pos = 3, srt = 90)

OUTPUT:

|  |
| --- |
| > table(titanicdata$Title) # Count of Titles  female male  466 843 |
|  |
| |  | | --- | |  | |

#---------------------------------------------

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

Solution:

x <- table(titanicdata$Survived, titanicdata$Title) # table for survived and died

x # 0 for survived and 1 for died

p <- x[1,] # number of passengers survived

p

Output:

> p

female male

127 682

prop <- round(p\*100/sum(p),1) # proportion of passangers survived

> prop <- round(p\*100/sum(p),1) # proportion of passangers survived

> prop

female male

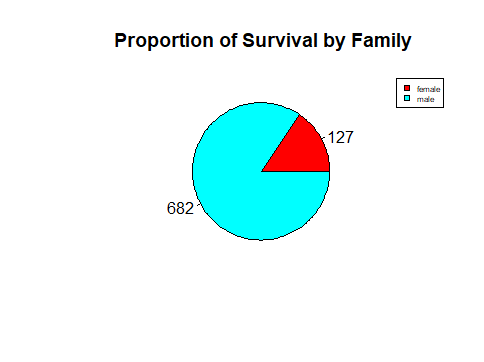
15.7 84.3

# in Pie Chart format

pie\_chart <- pie(p, labels = p, main = " Proportion of Survival by Family",

col = rainbow(length(p)), cex = 1)

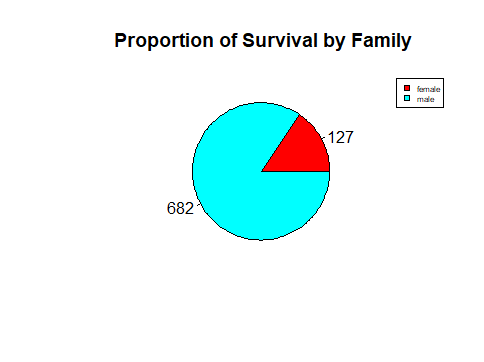
legend("topright", names(p), cex= 0.5, fill = rainbow(length(p)))



pie(prop, labels = prop, main = " Proportion of Survival by Family",

col = rainbow(length(prop)), cex = 1)

legend("topright", names(prop), cex= 0.5, fill = rainbow(length(prop)))



# in barchart format

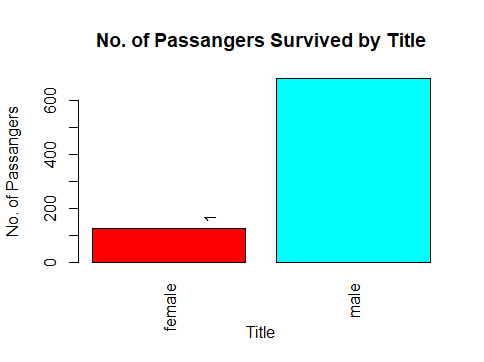
barplot(p, # for number of Passangers

main = "No. of Passangers Survived by Title",

xlab = "Title",

ylab = "No. of Passangers", col = rainbow(length(p)), las =3)

text(p, pos = 3, srt = 90)



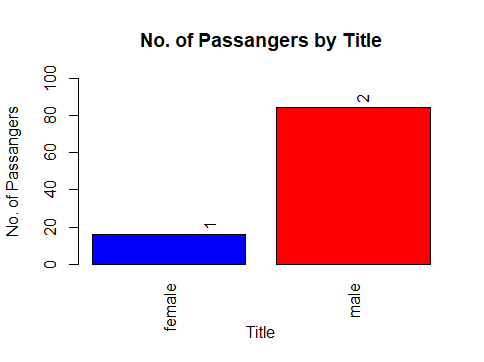
barplot(prop, # for percentage of passangers

main = "No. of Passangers by Title", xlab = "Title",

ylab = "No. of Passangers", col = c("Blue","Red"),

legend = rownames(prop), ylim=c(0, 100), las = 3)

text(prop, pos = 3, srt = 90)



#-----------------------------------------------------------

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

# different graphs showing Age distribution before and after imputation.

library(readr)

TitanicData <- within(TitanicData,

{

agecat <- NA

agecat[Age>=0 & Age<=25] <- "Low"

agecat[Age>=26 & Age<=40] <- "Middle"

agecat[Age>=41] <- "High"

})

head(TitanicData)

# Title and Age Group before imputation

count <- table(TitanicData$agecat, TitanicData$Title)

count

library(ggplot2)

p <- ggplot(data = TitanicData,

mapping = aes(Title, fill = agecat))

p + geom\_bar(position = "stack") + theme(axis.text.x = element\_text(angle = 90)) + labs(title = "Counts of Title with Age Groups")

library(mice)

# All variables shoud be either factor or numeric.

library(dplyr)

str(TitanicData)

dat <- TitanicData[,-13]

str(dat)

dat <- dat %>% mutate(agecat = as.factor(agecat),Title = as.factor(Title)) # convert as factor

str(dat) # Check the data set

# Now the data set is ready for imputation

# using library mice. called earlier

init = mice(dat, maxit=0)

meth = init$method

predM = init$predictorMatrix

# below variable are not required for predicting the age

predM[, c("PassengerId","Name", "Age","Ticket","Cabin", "Embarked")]=0

# specify method for imputing the missing value

meth[c("Age")]="norm"

set.seed(1)

# impute the missing values

imputed = mice(dat, method=meth, predictorMatrix=predM, m=5)

imputed <- complete(imputed)

# check for missings in the imputed dataset

sapply(imputed, function(x) sum(is.na(x)))

# Title and Age Group after imputation

library(ggplot2)

p <- ggplot(data = imputed,

mapping = aes(Title, fill = agecat))

p + geom\_bar(position = "stack")+theme(axis.text.x = element\_text(angle = 90)) + labs(title = "Counts of Title with Age Groups")