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

library(xlsx)

library(rJava)

titanic3 <- read\_excel("E:/D/desktop/Acadgild/re89070dataanalyticswithrexceltableauneedt/titanic3.xls")

str(titanic3)

head(titanic3)

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

titanic3$Title <- gsub('(.\*, )|(\\..\*)', '', titanic3$name)

table(titanic3$sex, titanic3$Title)

rare\_title <- c('Dona', 'Lady', 'the Countess','Capt', 'Col', 'Don', 'Dr', 'Major', 'Rev', 'Sir', 'Jonkheer')

titanic3$Title[titanic3$Title == 'Mlle'] <- 'Miss'

titanic3$Title[titanic3$Title == 'Ms'] <- 'Miss'

titanic3$Title[titanic3$Title == 'Mme'] <- 'Mrs'

titanic3$Title[titanic3$Title %in% rare\_title] <- 'Rare Title'

table(titanic3$sex, titanic3$Title)

titanic3$Surname <- sapply(titanic3$name, function(x) strsplit(x, split = '[,.]')[[1]][1])

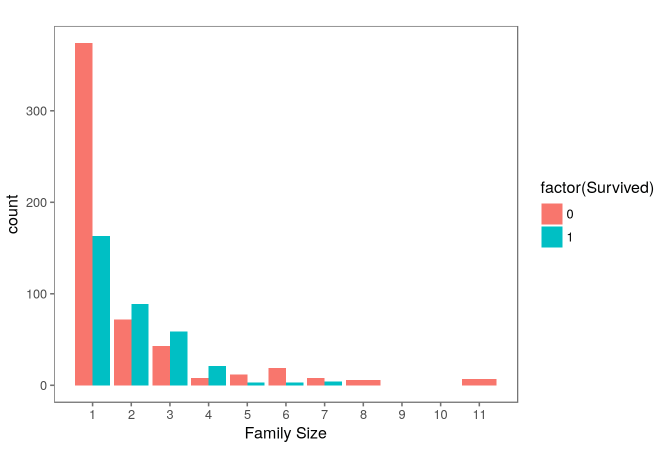
paste("we have",nlevels(factor(titanic3$Surname)), "unique families")

titanic3$Familysize <- titanic3$sibsp + titanic3$parch + 1

titanic3$Family <- paste(titanic3$Surname, titanic3$Familysize, sep='\_')

ggplot(titanic3, aes(x = Familysize, fill = factor(survived))) + geom\_bar(stat='count', position='dodge') +

scale\_x\_continuous(breaks=c(1:11)) + labs(x = 'Family Size') + theme\_few()



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

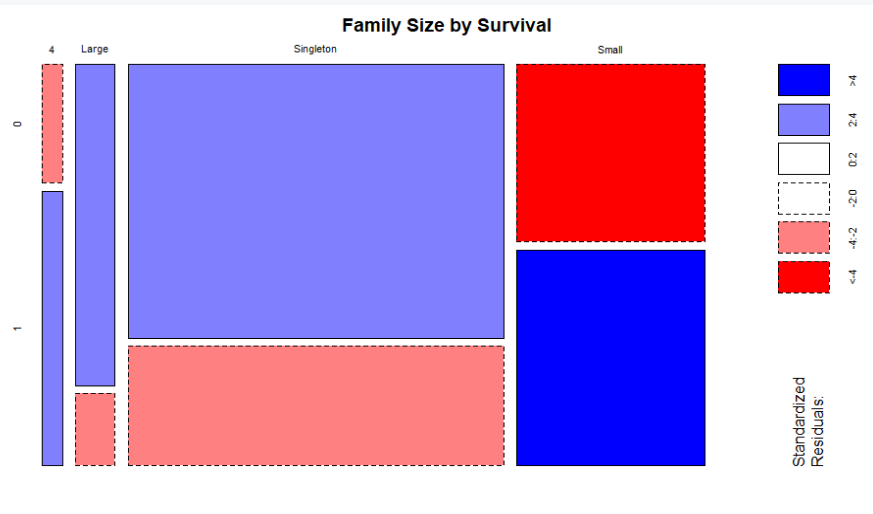
titanic3$fs <- titanic3$Familysize

titanic3$fs[titanic3$Familysize == 1] <- 'Singleton'

titanic3$fs[titanic3$Familysize < 5 & titanic3$Familysize >1 ] <- 'Small'

titanic3$fs[titanic3$Familysize > 4] <- 'Large'

mosaicplot(table(titanic3$fs, titanic3$survived), main='Family Size by Survival', shade=TRUE)



#c. Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputation.

library(mice)

mice\_mod <- mice(titanic3[, !names(titanic3) %in% c('passengerId','name','ticket','cabin','Family','Surname','survived')], method='rf')

mice\_output <- complete(mice\_mod)

par(mfrow=c(1,2))

hist(titanic3$age, freq=F, main='Age: Original Data', xlab = "AGE", col= "lightblue")

hist(mice\_output$age, freq=F, main='Age: MICE Output', xlab = "AGE", col='lightpink')

