Data Vizualization for Titanic Data

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.3

#Read the train data.  
trn <- read.csv("ttn\_train.csv",header=TRUE,stringsAsFactors=FALSE)  
#Check how many rows and columns in the data  
dim(trn)

## [1] 891 12

# Check the data types and what sort of values are there in the data set  
str(trn)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

#Convert the columns such as Pclass,Sex,Sibsp etc to Factor columns as these have very few unique values.  
  
trn$Sex <- as.factor(as.character(trn$Sex))  
trn$Survived <- as.factor(as.character(trn$Survived))  
trn$Pclass <- as.factor(as.character(trn$Pclass))  
trn$SibSp <- as.factor(as.character(trn$SibSp))  
trn$Parch <- as.factor(as.character(trn$Parch))  
  
trn$Embarked <- ifelse(trn$Embarked =="","S",trn$Embarked)  
trn$Embarked <- as.factor(as.character(trn$Embarked))  
summary(trn$Embarked)

## C Q S   
## 168 77 646

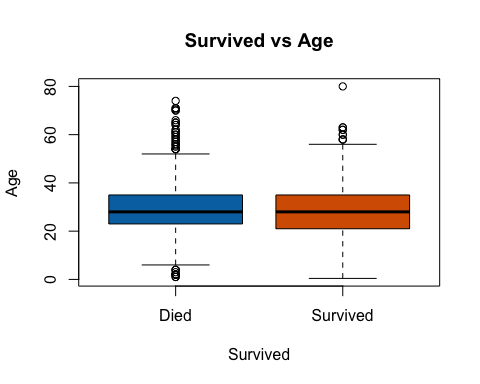
#Only Age has 177 misisng records. I will impute these with median values of age. There are other ways to do as well but I will keep it simple and use median age  
  
med\_age <- median(trn$Age,na.rm=TRUE)  
#impute the data  
trn$Age <- ifelse(is.na(trn$Age),med\_age,trn$Age)  
  
  
summary(trn)

## PassengerId Survived Pclass Name Sex   
## Min. : 1.0 0:549 1:216 Length:891 female:314   
## 1st Qu.:223.5 1:342 2:184 Class :character male :577   
## Median :446.0 3:491 Mode :character   
## Mean :446.0   
## 3rd Qu.:668.5   
## Max. :891.0   
##   
## Age SibSp Parch Ticket Fare   
## Min. : 0.42 0:608 0:678 Length:891 Min. : 0.00   
## 1st Qu.:22.00 1:209 1:118 Class :character 1st Qu.: 7.91   
## Median :28.00 2: 28 2: 80 Mode :character Median : 14.45   
## Mean :29.36 3: 16 3: 5 Mean : 32.20   
## 3rd Qu.:35.00 4: 18 4: 4 3rd Qu.: 31.00   
## Max. :80.00 5: 5 5: 5 Max. :512.33   
## 8: 7 6: 1   
## Cabin Embarked  
## Length:891 C:168   
## Class :character Q: 77   
## Mode :character S:646   
##   
##   
##   
##

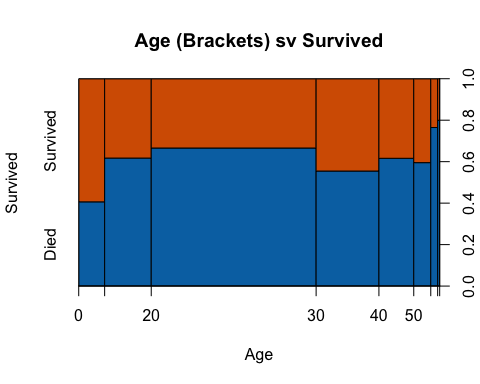
#It provides meadin,mean and other summary stats. I could see that 687 values for cabin are blank andfor Embarked there are 2 such values.  
# I will remove cabin column as well as Ticket columns form the data set  
  
trn$Cabin <- NULL  
trn$Ticket <- NULL  
  
#Now for vizualization Let us make sure that we have proper values rather than 0 and 1  
  
levels(trn$Survived) <- c("Died","Survived")  
levels(trn$Pclass) <- c("First","Second","Third")  
levels(trn$Sex) <- c("Female","Male")  
levels(trn$Embarked) <- c("Cherbourg","Queenstown","Southampton")  
  
# try summary data again  
  
summary(trn)

## PassengerId Survived Pclass Name   
## Min. : 1.0 Died :549 First :216 Length:891   
## 1st Qu.:223.5 Survived:342 Second:184 Class :character   
## Median :446.0 Third :491 Mode :character   
## Mean :446.0   
## 3rd Qu.:668.5   
## Max. :891.0   
##   
## Sex Age SibSp Parch Fare   
## Female:314 Min. : 0.42 0:608 0:678 Min. : 0.00   
## Male :577 1st Qu.:22.00 1:209 1:118 1st Qu.: 7.91   
## Median :28.00 2: 28 2: 80 Median : 14.45   
## Mean :29.36 3: 16 3: 5 Mean : 32.20   
## 3rd Qu.:35.00 4: 18 4: 4 3rd Qu.: 31.00   
## Max. :80.00 5: 5 5: 5 Max. :512.33   
## 8: 7 6: 1   
## Embarked   
## Cherbourg :168   
## Queenstown : 77   
## Southampton:646   
##   
##   
##   
##

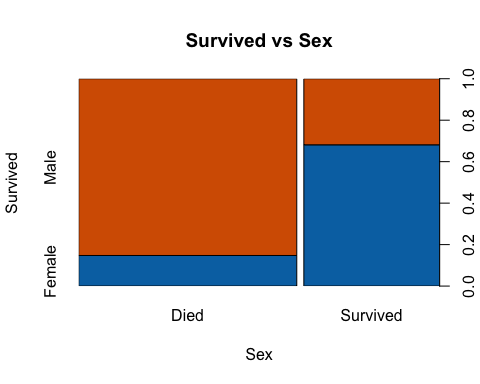
# Now let us look at some of the graphs. In Data analysis we found that age Pclass and sex were quite good columns for model and let us visually see them.  
  
#set the colors for graphs  
colset <- c("#0072B2", "#D55E00" ,"#CC79A7")  
#boxplot  
plot(trn$Age~trn$Survived,col=colset,xlab='Survived',ylab='Age',main="Survived vs Age")



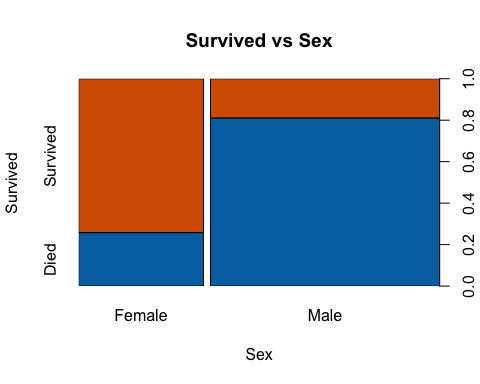
# survived has 25th percentile on lower side as compared to not survived.  
#mosaic plot style where age will be divided into 10 buckets andit will provide hows the survial rate is in these brackeets.  
plot(trn$Survived~trn$Age,col=colset,xlab='Age',ylab='Survived',main="Age (Brackets) sv Survived")



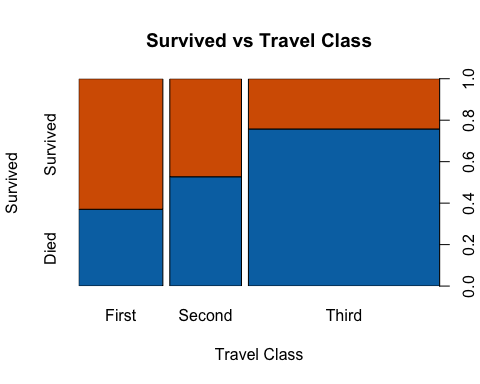
# Here age was divided into intervals of 10's and the breadth of interval corresponds to the number of people.  
  
#As you could see that more people survived in 0-10 an d11-20 as compared to 20-30 and for 30-40 more people survived.  
  
#Now let us have a look at sex columns  
plot(trn$Survived,trn$Sex,col=colset,xlab='Sex',ylab='Survived',main="Survived vs Sex")



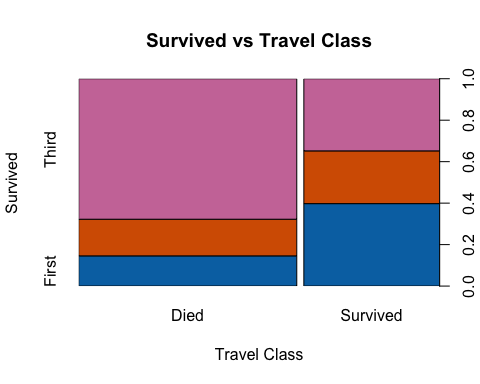
#This is another way to look at the graph  
plot(trn$Survived~trn$Sex,col=colset,xlab='Sex',ylab='Survived',main="Survived vs Sex")



# This clearly shows the distinction and as you could see that the survival for females are much higher than males.  
  
#Now let us have a look at Pclass columns  
plot(trn$Survived~trn$Pclass,col=colset,xlab='Travel Class',ylab='Survived',main="Survived vs Travel Class")



plot(trn$Survived,trn$Pclass,col=colset,xlab='Travel Class',ylab='Survived',main="Survived vs Travel Class")



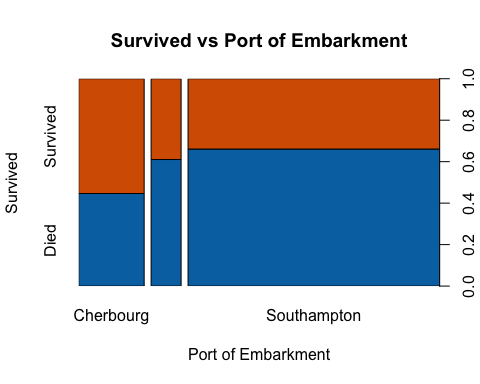
# It clearly shows that people in 3rd class had very less chance of survival as compared to 2nd and 1st class. In fact chances of survival in First class are highest.  
  
#try sibsp columns. I have defined just 3 colors  
plot(trn$Survived~trn$SibSp,col=colset,xlab='Siblings & Sposues',ylab='Survived',main="Survived vs Number of Siblings & Spouses")



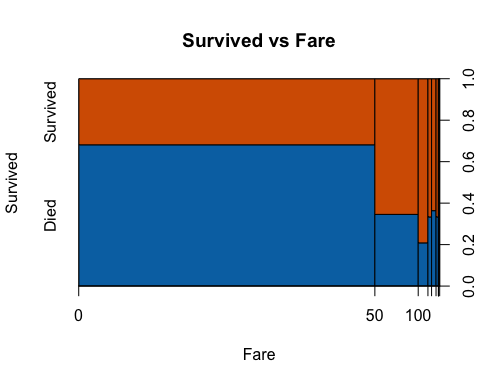
# here you could see that the chances of survival when you have 1 or 2 siblings and spuses were high. The breadth of these graphs corresponds to the count and thus you could see that most of the people were without sibling and spouses on aboard.  
  
#try Parch columns  
plot(trn$Survived~trn$Parch,col=colset,xlab='Parch',ylab='Survived',main="Survived vs Number of Parents & Childs")



# here you could see that the chances of survival when you have 1 or 2 Parch and spuses were high. The breadth of these graphs corresponds to the count and thus you could see that most of the people were without sibling and spouses on aboard.  
  
#Now let us see for portof embarkment  
  
#try Embarked columns  
plot(trn$Survived~trn$Embarked,col=colset,xlab='Port of Embarkment',ylab='Survived',main="Survived vs Port of Embarkment")



# Here it looks like that the chances of survival are higher if you boarded from Cherbourg and very less chances when you boarded from Southampton.  
  
#try the fare  
plot(trn$Survived~trn$Fare,col=colset,xlab='Fare',ylab='Survived',main="Survived vs Fare")



#As the fare amount increases there are high chances of survival and this looks like correlared with the Pclass column which we already know that explains the survival.  
  
# Now let us take an example where it is possible that there were more females say in Pclass 1 and that is why the success rate is high or vice versa and this is the same case for any combination of the columns. Now we will look at these interactions as well.  
  
mosaicplot(~ Sex + Pclass + Survived, data = trn, color = colset, xlab= "Sex", ylab ="Travel Class",main="Mosaic Plot for Sex, Travel Class and Survived" )



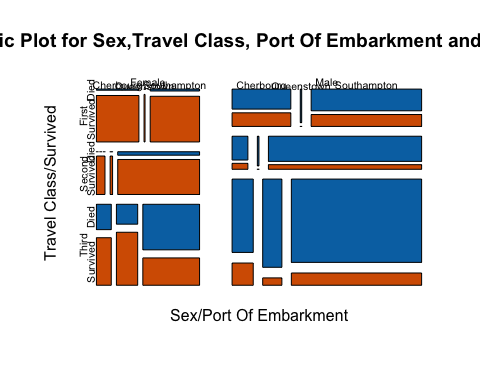
# This is an intersting graph.It shows that if you were a Female from First and Second class chances were much higher as compared to anything else. Although, males from class 1 and 2 has much higher rate of survival as compared to people from 3rd class.  
  
# Females in 1st class had very high rate of survivall and in 3rd class this rate is around 50%.  
  
#Males in 1st class has very high rate but it is not better than the female in 3rd class and thus it looks like that sex was more dominant feature here.  
  
#sex embarked  
mosaicplot(~ Sex + Embarked + Survived , data = trn, color = colset,xlab= "Sex", ylab ="Port Of Embarkment",main="Mosaic Plot for Sex, Port Of Embarkment and Survived" )



#Here you could see that if you were Male from the Queenstown then survival chances were quite less as compared to males from other places..  
  
#Travle class embarked  
mosaicplot(~ Pclass + Embarked + Survived , data = trn, color = colset,xlab= "Travel Class", ylab ="Port Of Embarkment",main="Mosaic Plot for Travel Class, Port Of Embarkment and Survived" )



#Now put all of these together and see how it will work  
  
mosaicplot(~ Sex + Pclass + Embarked + Survived , data = trn, color = colset,xlab= "Sex/Port Of Embarkment", ylab ="Travel Class/Survived",main="Mosaic Plot for Sex,Travel Class, Port Of Embarkment and Survived" )



# There are some intersting facts here. E.g. Males from Queenstown had very less chance of survival irrespective of class. In fact only surived looks from 3rd class.But there are just 1,1 people in these classes. Also, for males port of embarkement did not matter in the first and second class. Howvere, there is a difference for 3rd class.  
  
#For females howevere, the port of emabarkment had differnece for second class and 3rd class both.  
  
table(trn$Pclass,trn$Sex,trn$Embarked)

## , , = Cherbourg  
##   
##   
## Female Male  
## First 43 42  
## Second 7 10  
## Third 23 43  
##   
## , , = Queenstown  
##   
##   
## Female Male  
## First 1 1  
## Second 2 1  
## Third 33 39  
##   
## , , = Southampton  
##   
##   
## Female Male  
## First 50 79  
## Second 67 97  
## Third 88 265

mosaicplot(~ Sex + Parch + Survived , data = trn, color = colset,xlab= "Sex", ylab ="Number of Parents & Childs",main="Mosaic Plot for Sex,Number of Parents & Childs and Survived" )



#for males it looks like that as the number of Parch increased the chances of survival were much better for females it was other way around.  
  
mosaicplot(~ Sex + SibSp + Survived , data = trn, color = colset,xlab= "Sex", ylab ="Number of Siblings & Spouses",main="Mosaic Plot for Sex,Number of Siblings & Spouses and Survived" )



#It shows the same trend as Parent and Child except that there is an exception for femals when value is 3 and I am sure that it is just a small number and only 0,1,2 will have serious impact on how good this feature is.