Data Analayis for Titanic Data

#Read the train data.  
trn <- read.csv("ttn\_train.csv",header=TRUE)  
#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 : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ 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 : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...  
## $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

#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$Embarked <- as.factor(as.character(trn$Embarked))  
  
# Check the data types and what sort of values are there in the data set again to see that needed colums are now converted to factor columns.  
str(trn)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1 4 1 2 ...  
## $ Parch : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 2 3 1 ...  
## $ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...  
## $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

#Now chck whether we have any misisng data  
dim(trn)[1]

## [1] 891

#comple cases will give the data where all columns are not NA if any of the column is NA that record wont be a comple case and thus it is a way to find out any na in whole record.  
sum(complete.cases(trn))

## [1] 714

#Now find the columns which has the missing data.  
  
sapply(lapply(trn,function(x) is.na(x)),sum)

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 177   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 0 0

#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)  
  
#run below again and make sure that it is same as dim(trn)  
sum(complete.cases(trn))

## [1] 891

#Now we have the data Let us look at how the data looks like. I will use summary command to see the data for all the columns  
  
summary(trn)

## PassengerId Survived Pclass   
## Min. : 1.0 0:549 1:216   
## 1st Qu.:223.5 1:342 2:184   
## Median :446.0 3:491   
## Mean :446.0   
## 3rd Qu.:668.5   
## Max. :891.0   
##   
## Name Sex Age   
## Abbing, Mr. Anthony : 1 female:314 Min. : 0.42   
## Abbott, Mr. Rossmore Edward : 1 male :577 1st Qu.:22.00   
## Abbott, Mrs. Stanton (Rosa Hunt) : 1 Median :28.00   
## Abelson, Mr. Samuel : 1 Mean :29.36   
## Abelson, Mrs. Samuel (Hannah Wizosky): 1 3rd Qu.:35.00   
## Adahl, Mr. Mauritz Nils Martin : 1 Max. :80.00   
## (Other) :885   
## SibSp Parch Ticket Fare Cabin   
## 0:608 0:678 1601 : 7 Min. : 0.00 :687   
## 1:209 1:118 347082 : 7 1st Qu.: 7.91 B96 B98 : 4   
## 2: 28 2: 80 CA. 2343: 7 Median : 14.45 C23 C25 C27: 4   
## 3: 16 3: 5 3101295 : 6 Mean : 32.20 G6 : 4   
## 4: 18 4: 4 347088 : 6 3rd Qu.: 31.00 C22 C26 : 3   
## 5: 5 5: 5 CA 2144 : 6 Max. :512.33 D : 3   
## 8: 7 6: 1 (Other) :852 (Other) :186   
## Embarked  
## : 2   
## C:168   
## Q: 77   
## S:644   
##   
##   
##

#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.   
# There are 549 dies and 342survived. Mean age is 29.36 whereas median age is 28. Mean fare was 32 whereas median fare was 14.  
#Most of the people embarked from southampton  
  
str(trn)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 28 54 2 27 14 ...  
## $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1 4 1 2 ...  
## $ Parch : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 2 3 1 ...  
## $ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...  
## $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

#It shows the structire of teh data like data types and column types etc.  
# I will remove cabin column as well as Ticket columns form the data set  
  
trn$Cabin <- NULL  
trn$Ticket <- NULL  
  
#Survived is out columns which defines whether a person was survived or died. 1 means survived and 0 means not.  
  
#Let us use table command and see how will it look against other columns.  
table(trn$Survived)

##   
## 0 1   
## 549 342

# There were 549 survived and 342 did not survive.  
  
# age is continous column i will divide it by 10 and use floor values  
table(floor(trn$Age/10),trn$Survived)

##   
## 0 1  
## 0 24 38  
## 1 61 41  
## 2 268 129  
## 3 94 73  
## 4 55 34  
## 5 28 20  
## 6 13 6  
## 7 6 0  
## 8 0 1

#Survival ratio is good for people less than 10 years old and worst is for people between 20 and 30.  
# Note that the meadin age was 28 and thus we have lots of imputed data and if actual ages were known this woyld have been a bit differnt but overall picture should have been more or less similar.  
  
#Now check the class  
  
table(trn$Pclass,trn$Survived)

##   
## 0 1  
## 1 80 136  
## 2 97 87  
## 3 372 119

#Survival ratio is quite good in 1st Class which is expectedit is quite good in second class as welland in the 3rd class only 25% people survived. Thus, this looks one of possible candidate which could be used in model to check whether a person will survive or die.  
  
#now chekc based on Sex  
table(trn$Sex,trn$Survived)

##   
## 0 1  
## female 81 233  
## male 468 109

# Female survival ratio is quite high and mails is just 20 %.Thus, this seems to be a great feature as well.  
  
  
table(trn$Embarked,trn$Survived)

##   
## 0 1  
## 0 2  
## C 75 93  
## Q 47 30  
## S 427 217

#Port of emabrkment seems to be a nice columns as well. I will explain later why this might not be useful for final model.  
  
table(trn$SibSp,trn$Survived)

##   
## 0 1  
## 0 398 210  
## 1 97 112  
## 2 15 13  
## 3 12 4  
## 4 15 3  
## 5 5 0  
## 8 7 0

#if there were just 1 or sibling aboard chancec of survival looks better. Thu it could be useful.  
  
table(trn$Parch,trn$Survived)

##   
## 0 1  
## 0 445 233  
## 1 53 65  
## 2 40 40  
## 3 2 3  
## 4 4 0  
## 5 4 1  
## 6 1 0

#if there were just 1 or more parent aboard chances of survival looks better. Thu it could be useful.  
  
  
  
table(trn$Fare,trn$Survived)

##   
## 0 1  
## 0 14 1  
## 4.0125 1 0  
## 5 1 0  
## 6.2375 1 0  
## 6.4375 1 0  
## 6.45 1 0  
## 6.4958 2 0  
## 6.75 2 0  
## 6.8583 1 0  
## 6.95 1 0  
## 6.975 1 1  
## 7.0458 1 0  
## 7.05 7 0  
## 7.0542 2 0  
## 7.125 4 0  
## 7.1417 0 1  
## 7.225 9 3  
## 7.2292 11 4  
## 7.25 12 1  
## 7.3125 1 0  
## 7.4958 2 1  
## 7.5208 1 0  
## 7.55 3 1  
## 7.6292 1 0  
## 7.65 3 1  
## 7.725 1 0  
## 7.7292 1 0  
## 7.7333 2 2  
## 7.7375 1 1  
## 7.7417 1 0  
## 7.75 22 12  
## 7.775 13 3  
## 7.7875 0 1  
## 7.7958 4 2  
## 7.8 1 0  
## 7.8292 1 1  
## 7.8542 10 3  
## 7.875 1 0  
## 7.8792 0 4  
## 7.8875 1 0  
## 7.8958 37 1  
## 7.925 10 8  
## 8.0292 0 1  
## 8.05 38 5  
## 8.1125 0 1  
## 8.1375 1 0  
## 8.1583 1 0  
## 8.3 1 0  
## 8.3625 1 0  
## 8.4042 1 0  
## 8.4333 1 0  
## 8.4583 1 0  
## 8.5167 0 1  
## 8.6542 1 0  
## 8.6625 12 1  
## 8.6833 0 1  
## 8.7125 1 0  
## 8.85 1 0  
## 9 2 0  
## 9.2167 1 0  
## 9.225 2 0  
## 9.35 1 1  
## 9.475 1 0  
## 9.4833 1 0  
## 9.5 7 2  
## 9.5875 1 1  
## 9.825 2 0  
## 9.8375 1 0  
## 9.8417 0 1  
## 9.8458 1 0  
## 10.1708 1 0  
## 10.4625 2 0  
## 10.5 15 9  
## 10.5167 1 0  
## 11.1333 0 3  
## 11.2417 0 2  
## 11.5 4 0  
## 12 0 1  
## 12.275 1 0  
## 12.2875 0 1  
## 12.35 1 2  
## 12.475 0 4  
## 12.525 1 0  
## 12.65 0 1  
## 12.875 1 0  
## 13 26 16  
## 13.4167 0 1  
## 13.5 3 1  
## 13.7917 0 1  
## 13.8583 0 1  
## 13.8625 0 1  
## 14 1 0  
## 14.1083 1 0  
## 14.4 2 0  
## 14.4542 6 1  
## 14.4583 3 0  
## 14.5 5 2  
## 15 1 0  
## 15.0458 1 0  
## 15.05 1 0  
## 15.1 1 0  
## 15.2458 2 3  
## 15.5 5 3  
## 15.55 1 0  
## 15.7417 0 2  
## 15.75 0 1  
## 15.85 2 2  
## 15.9 0 2  
## 16 0 1  
## 16.1 7 2  
## 16.7 0 2  
## 17.4 0 1  
## 17.8 2 0  
## 18 3 0  
## 18.75 0 3  
## 18.7875 1 1  
## 19.2583 0 4  
## 19.5 0 2  
## 19.9667 2 0  
## 20.2125 2 0  
## 20.25 1 1  
## 20.525 1 2  
## 20.575 1 1  
## 21 4 2  
## 21.075 4 0  
## 21.6792 1 0  
## 22.025 0 1  
## 22.3583 0 2  
## 22.525 1 0  
## 23 0 4  
## 23.25 0 2  
## 23.45 2 0  
## 24 1 1  
## 24.15 7 1  
## 25.4667 4 0  
## 25.5875 1 0  
## 25.925 1 0  
## 25.9292 0 2  
## 26 16 15  
## 26.25 2 4  
## 26.2833 0 1  
## 26.2875 0 3  
## 26.3875 0 1  
## 26.55 7 8  
## 27 1 1  
## 27.7208 4 1  
## 27.75 2 2  
## 27.9 6 0  
## 28.5 1 0  
## 28.7125 1 0  
## 29 0 2  
## 29.125 5 0  
## 29.7 2 1  
## 30 1 5  
## 30.0708 1 1  
## 30.5 1 4  
## 30.6958 2 0  
## 31 1 2  
## 31.275 7 0  
## 31.3875 1 3  
## 32.3208 1 0  
## 32.5 0 1  
## 33 1 2  
## 33.5 1 0  
## 34.0208 1 0  
## 34.375 4 0  
## 34.6542 1 0  
## 35 1 0  
## 35.5 1 3  
## 36.75 1 1  
## 37.0042 1 1  
## 38.5 1 0  
## 39 1 3  
## 39.4 0 1  
## 39.6 1 1  
## 39.6875 6 0  
## 40.125 1 0  
## 41.5792 1 2  
## 42.4 1 0  
## 46.9 6 0  
## 47.1 1 0  
## 49.5 0 1  
## 49.5042 1 1  
## 50 1 0  
## 50.4958 1 0  
## 51.4792 0 1  
## 51.8625 1 1  
## 52 4 3  
## 52.5542 0 3  
## 53.1 2 3  
## 55 0 2  
## 55.4417 0 1  
## 55.9 1 1  
## 56.4958 2 5  
## 56.9292 0 2  
## 57 0 2  
## 57.9792 0 2  
## 59.4 0 1  
## 61.175 1 0  
## 61.3792 1 0  
## 61.9792 1 0  
## 63.3583 0 1  
## 65 0 2  
## 66.6 1 1  
## 69.3 0 2  
## 69.55 7 0  
## 71 1 1  
## 71.2833 0 1  
## 73.5 5 0  
## 75.25 0 1  
## 76.2917 0 1  
## 76.7292 0 3  
## 77.2875 2 0  
## 77.9583 0 3  
## 78.2667 0 2  
## 78.85 1 1  
## 79.2 2 2  
## 79.65 1 2  
## 80 0 2  
## 81.8583 0 1  
## 82.1708 1 1  
## 83.1583 0 3  
## 83.475 1 1  
## 86.5 0 3  
## 89.1042 0 2  
## 90 1 3  
## 91.0792 0 2  
## 93.5 0 2  
## 106.425 1 1  
## 108.9 1 1  
## 110.8833 1 3  
## 113.275 1 2  
## 120 0 4  
## 133.65 0 2  
## 134.5 0 2  
## 135.6333 1 2  
## 146.5208 0 2  
## 151.55 2 2  
## 153.4625 1 2  
## 164.8667 0 2  
## 211.3375 0 3  
## 211.5 1 0  
## 221.7792 1 0  
## 227.525 1 3  
## 247.5208 1 1  
## 262.375 0 2  
## 263 2 2  
## 512.3292 0 3

# There is no obvious trend here but looks like more the fare chances are better.