**# load train data**

train = read.csv("titanic-train.csv")

attach(train)

install.packages("ggplot2")

library(ggplot2)

**# check for empty data**

is.na(train)

**# get the number of missing records**

missing = sapply(train,function(x)sum(is.na(x)))

missing

**# as per the above code 177 people's ages are missing**

> missing

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket

0 0 0 0 0 177 0 0 0

Fare Cabin Embarked

0 0 0

**# replacing the missing ages with mean**

train$Age = ifelse(is.na(train$Age),

ave(train$Age, FUN = function(x) mean(x, na.rm = TRUE)),

train$Age)

> summary(train)

PassengerId Survived Pclass Name Sex

Min. : 1.0 Min. :0.0000 Min. :1.000 Abbing, Mr. Anthony : 1 female:314

1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Abbott, Mr. Rossmore Edward : 1 male :577

Median :446.0 Median :0.0000 Median :3.000 Abbott, Mrs. Stanton (Rosa Hunt) : 1

Mean :446.0 Mean :0.3838 Mean :2.309 Abelson, Mr. Samuel : 1

3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000 Abelson, Mrs. Samuel (Hannah Wizosky): 1

Max. :891.0 Max. :1.0000 Max. :3.000 Adahl, Mr. Mauritz Nils Martin : 1

(Other) :885

Age SibSp Parch Ticket Fare Cabin Embarked

Min. : 0.42 Min. :0.000 Min. :0.0000 1601 : 7 Min. : 0.00 :687 : 2

1st Qu.:22.00 1st Qu.:0.000 1st Qu.:0.0000 347082 : 7 1st Qu.: 7.91 B96 B98 : 4 C:168

Median :29.70 Median :0.000 Median :0.0000 CA. 2343: 7 Median : 14.45 C23 C25 C27: 4 Q: 77

Mean :29.70 Mean :0.523 Mean :0.3816 3101295 : 6 Mean : 32.20 G6 : 4 S:644

3rd Qu.:35.00 3rd Qu.:1.000 3rd Qu.:0.0000 347088 : 6 3rd Qu.: 31.00 C22 C26 : 3

Max. :80.00 Max. :8.000 Max. :6.0000 CA 2144 : 6 Max. :512.33 D : 3

(Other) :852 (Other) :186 summary(train)

**# box plot for checking outliers and as per the plot Age above 70 are outlier**

boxplot(train$Age,main = "outliers",xlab = "Age", col = "blue")



**# checking correlation**

cor(train[c(2,6)])

pairs(~factor(train$Survived)+train$Age)

pairs(~factor(train$Survived)+train$Sex)

plot(train$Sex,train$Age,xlab = "Sex", ylab = "Age")







**# plot the graph to identify the percentage of relation between sex and survival. As per the Plot 75% of female survived**

**# whereas only 20 percentage of male survived.**

install.packages("ggplot2")

library(ggplot2)

ggplot()+

geom\_bar(data=train, aes(x=factor(train$Sex),fill = factor(train$Survived)),

position = "fill")+

scale\_x\_discrete("Sex")+

scale\_y\_continuous("Percent")+

guides(fill = guide\_legend(title = "Survived"))+

scale\_fill\_manual(values = c("blue","yellow"))



**# Bucketing the Age**

AgeBucket = cut(Age, breaks = c(0, 20, 40, 60, 80))

AgeBucket

**# Plot Age vs Survived**

ggplot()+

geom\_bar(data=train, aes(x=factor(AgeBucket),fill = factor(train$Survived)),

position = "fill")+

scale\_x\_discrete("Age")+

scale\_y\_continuous("Percent")+

guides(fill = guide\_legend(title = "Survived"))+

scale\_fill\_manual(values = c("blue","yellow"))



**# asper the plot the highest percent of survival is for the age less than 20**

**# decision tree**

library(rpart)

library(rpart.plot)

tree = rpart(train$Survived~AgeBucket+Sex+Fare+SibSp+Parch, data = train,method = "class", cp = 0.001)

rpart.plot(tree)

