Titanic

**Packages:**

**library**('ggplot2') *# visualization*

**library**('ggthemes') *# visualization*

**library**('scales') *# visualization*

**library**('dplyr') *# data manipulation*

**library**('mice') *# imputation*

**library**('randomForest') *# classification algorithm*

**commands used:**

>train <- **read.csv**('../input/train.csv', stringsAsFactors = F)

>test <- **read.csv**('../input/test.csv', stringsAsFactors = F)

>full <- **bind\_rows**(train, test) *# bind training & test data*

str(full)

Output:

'data.frame': 1309 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" ...

**Number of survivors and non-survivors:**

> table(train$Survived)

0 1

549 342

**Proportion of survivors and non-survivors:**

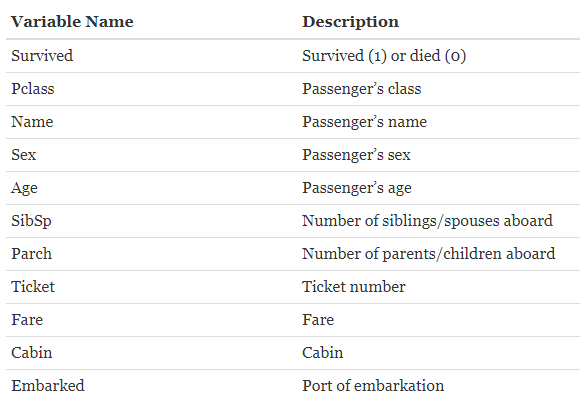
> prop.table(table(train$Survived))

0 1

0.6161616 0.3838384

Ie. 38% of passengers survived disaster in the training set.

**Variables used and their description:**



**Proportion of male and female survivors and non survivors(includes total trai-ning data) :**

> prop.table(table(train$Sex,train$Survived))

0 1

female 0.09090909 0.26150393

male 0.52525253 0.12233446

Explanation: 9% females aboard dead,26% survived

52% males dead and 12% survived

**Proportion of male and female survivors and non survivors(% for male and female separately) :**

> prop.table(table(train$Sex,train$Survived),1)

0 1

female 0.2579618 0.7420382

male 0.8110919 0.1889081

**No. of male and female survivors and non-survivors:**

> table(train$Sex,train$Survived)

0 1

female 81 233

male 468 109

**Finding number of children:**

train$child<-0 #Creating a new variable to find number of children

train$child[train$Age<18]<-1 #Assigning value 1 to represent children

**Finding aggregate number of children survivors:**

> aggregate(Survived ~ Child + Sex, data=train, FUN=sum)

Child Sex Survived

1 0 female 195

2 1 female 38

3 0 male 86

4 1 male 23

The aggregate command takes a formula with the target variable on the left hand side of the tilde symbol and the variables to subset over on the right. We then tell it which dataframe to look at with the data argument, and finally what function to apply to these subsets.

**Aggerate no of children survivors:**

> aggregate(Survived ~ Child + Sex, data=train, FUN=length)

Child Sex Survived

1 0 female 259

2 1 female 55

3 0 male 519

4 1 male 58

**Survival rate wrt to family size:**

We create variable Fsize to based on number of siblings/spouse(s) and number of children/parents.

# Create a family size variable including the passenger themselves

> full$Fsize <- full$SibSp + full$Parch + 1

# Create a family variable

> full$Family <- paste(full$Surname, full$Fsize, sep='\_')

***Plotting to show relationship between Fsize and survival:***

> ggplot(full[1:891,], aes(x = Fsize, fill = factor(Survived))) +

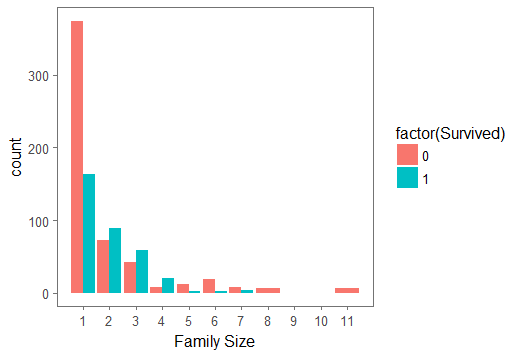
+ geom\_bar(stat='count', position='dodge') +

+ scale\_x\_continuous(breaks=c(1:11)) +

+ labs(x = 'Family Size') +

+ theme\_few()

Output:



**#Fare Variables**

train$Fare2 <- '30+'

train$Fare2[train$Fare < 30 & train$Fare >= 20] <- '20-30'

train$Fare2[train$Fare < 20 & train$Fare >= 10] <- '10-20'

train$Fare2[train$Fare < 10] <- '<10'

**Using aggregate function to see percentage of survivor based on fare and class:**

> aggregate(Survived ~ Fare2 + Pclass + Sex, data=train, FUN=function(x){sum(x)/length(x)})

Fare2 Pclass Sex Survived

1 20-30 1 female 0.8333333

2 30+ 1 female 0.9772727

3 10-20 2 female 0.9142857

4 20-30 2 female 0.9000000

5 30+ 2 female 1.0000000

6 <10 3 female 0.5937500

7 10-20 3 female 0.5813953

8 20-30 3 female 0.3333333

9 30+ 3 female 0.1250000

10 <10 1 male 0.0000000

11 20-30 1 male 0.4000000

12 30+ 1 male 0.3837209

13 <10 2 male 0.0000000

14 10-20 2 male 0.1587302

15 20-30 2 male 0.1600000

16 30+ 2 male 0.2142857

17 <10 3 male 0.1115385

18 10-20 3 male 0.2368421

19 20-30 3 male 0.1250000

20 30+ 3 male 0.2400000

Explanation: While the majority of males, regardless of class or fare still don’t do so well, we notice that most of the class 3 women who paid more than $20 for their ticket actually also miss out on a lifeboat.

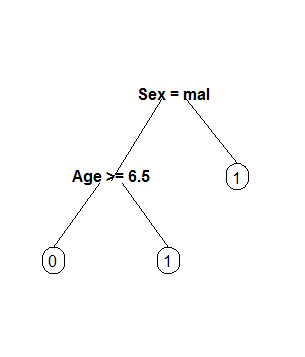
Decision Tree Representation.

Library used:

library(rpart.plot)

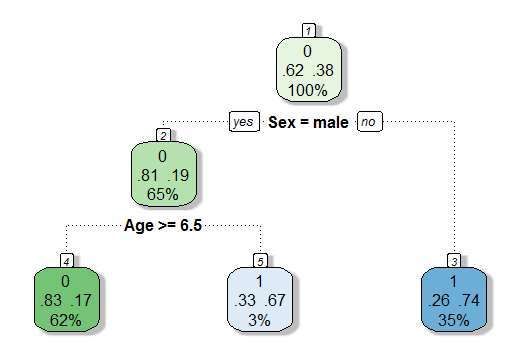
library(RColorBrewer)

> new.fit <- prp(my\_tree,snip=TRUE)$obj



This is a simple decision tree representation that does not give us much information.

**More Representative Decision Tree:**

****

Code used:

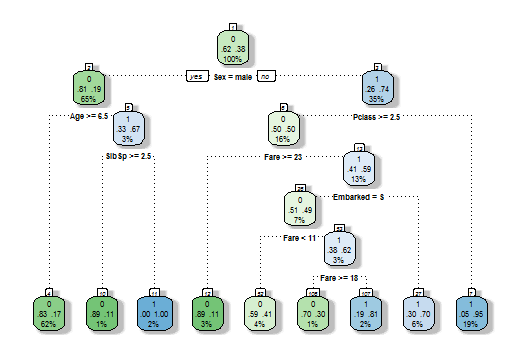
> rpart.plot(binary.model,

+ extra=104, box.palette="GnBu",

+ branch.lty=3, shadow.col="gray", nn=TRUE)

Explanation:

The root node, at the top, shows our [tutorial one](http://trevorstephens.com/kaggle-titanic-tutorial/r-part-1-booting-up) insights, 62% of passengers die, while 38% survive. If the passenger was male, only 19% survive, so the bucket votes that everyone here (65% of passengers) perish.



For every attribute. Give Explanation

**No. of Null values in age:**

> sum(is.na(full$Age))

[1] 263

**Predicting Null values for age:**

> Agefit <- rpart(Age ~ Pclass + Sex + SibSp + Parch + Fare + Embarked + Title + Fsize,

+ data=full[!is.na(full$Age),],

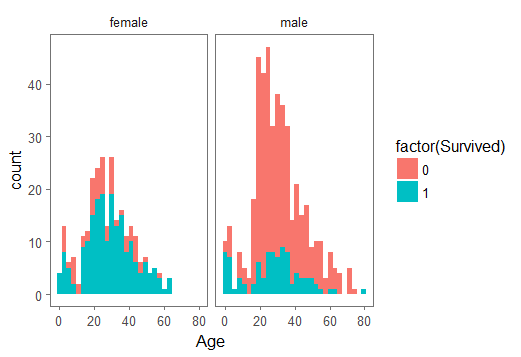
+ method="anova")

> full$Age[is.na(full$Age)] <- predict(Agefit, full[is.na(full$Age),])

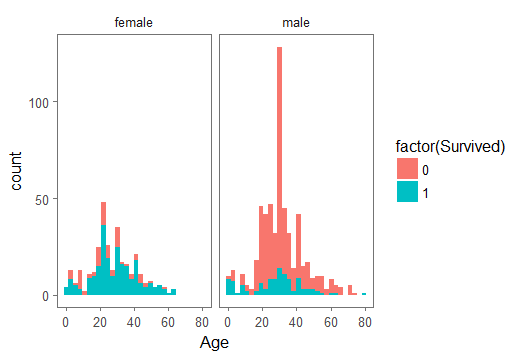
> sum(is.na(full$Age))

[1] 0

**Before Prediction of age.**

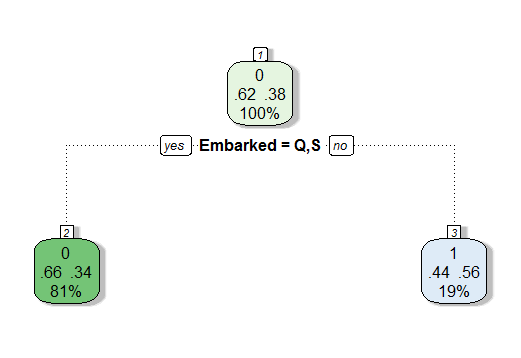
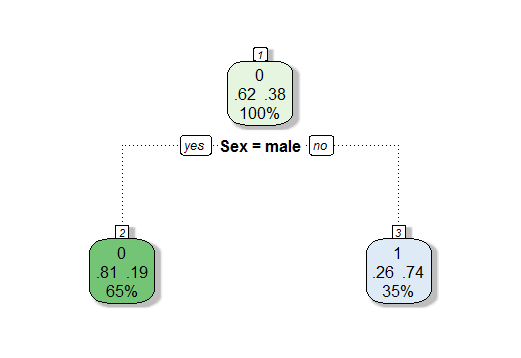
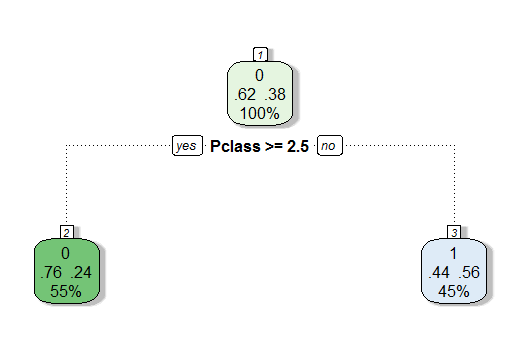
****

**After Prediction of age.**

****

**Random Forest;**

**3 decision trees.**

****