```r
library(ggplot2)
#install.packages('Rcpp')
#install.packages('dplyr')
library(dplyr)
library(randomForest)
library(Rcpp)
library(dplyr)
```

Step 1: Load the data.

```r
setwd("C:/Users/nwelpulw/Desktop/Udemy/Projects/Titanic")
train<-read.csv('train.csv',stringsAsFactors = FALSE)
test<-read.csv('test.csv',stringsAsFactors = FALSE)
```

train has 12 variables but test has 11 variables and to combbine both datasets,number of column should be same.

So add Survived column with NA value in test dataset.

```r
test$Survived<-NA
```

Combine both datasets.

```r
full<-rbind(train,test)
str(full)
```

```
## '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)"
##  $ 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" ...
```

Feature engineering with Name.

```r
head(full$Name)
```

```
## [1] "Braund, Mr. Owen Harris"
## [2] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
## [3] "Heikkinen, Miss. Laina"
## [4] "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
## [5] "Allen, Mr. William Henry"
## [6] "Moran, Mr. James"
```

Take out titles.

```
strsplit(full$Name,split = '[,.]')[[1]][2]
```

```
## [1] " Mr"
```

```
full$Title<-sapply(full$Name,FUN = function(x){strsplit(x,split = '[,.]')[[1]][2]})
```

There is blank space before title which needs to be removed.

```
full$Title<-sub(" ","",full$Title)

table(full$Title,full$Sex)
```

```
##
##                 female male
##   Capt               0    1
##   Col                0    4
##   Don                0    1
##   Dona               1    0
##   Dr                 1    7
##   Jonkheer           0    1
##   Lady               1    0
##   Major              0    2
##   Master             0   61
##   Miss             260    0
##   Mlle               2    0
##   Mme                1    0
##   Mr                 0  757
##   Mrs              197    0
##   Ms                 2    0
##   Rev                0    8
##   Sir                0    1
##   the Countess       1    0
```

```
Rare_Title<-c('Capt','Col','Don','Dona','Dr','Jonkheer','Lady','Major','Rev','Sir','the Countess')

full$Title[full$Title=='Mlle' | full$Title=='Ms']<-'Miss'
full$Title[full$Title=='Mme']<-'Mrs'
full$Title[full$Title %in% Rare_Title]<-'Rare_Title'

table(full$Title,full$Sex)
```

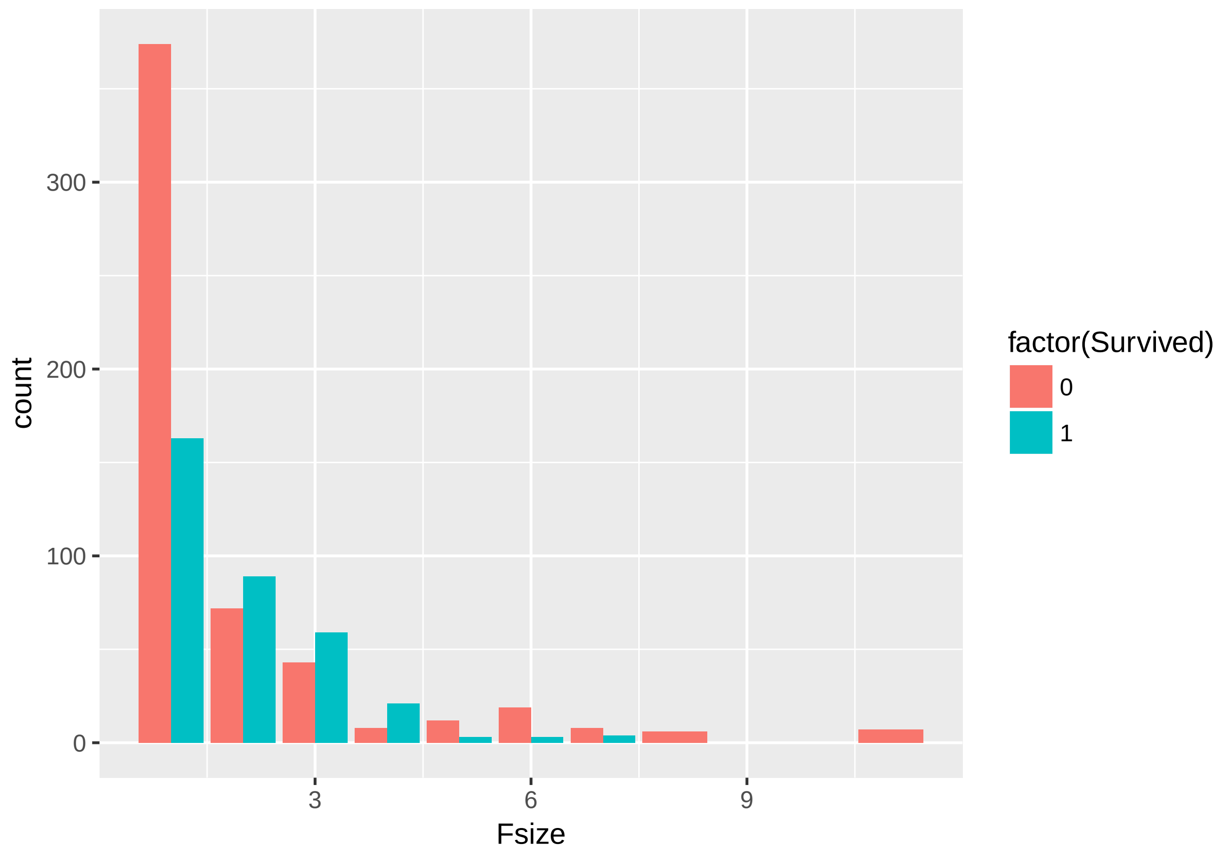
```
##
```

```
##              female male
##   Master          0   61
##   Miss          264    0
##   Mr              0  757
##   Mrs           198    0
##   Rare_Title      4   25
```

************** Family Size **************************

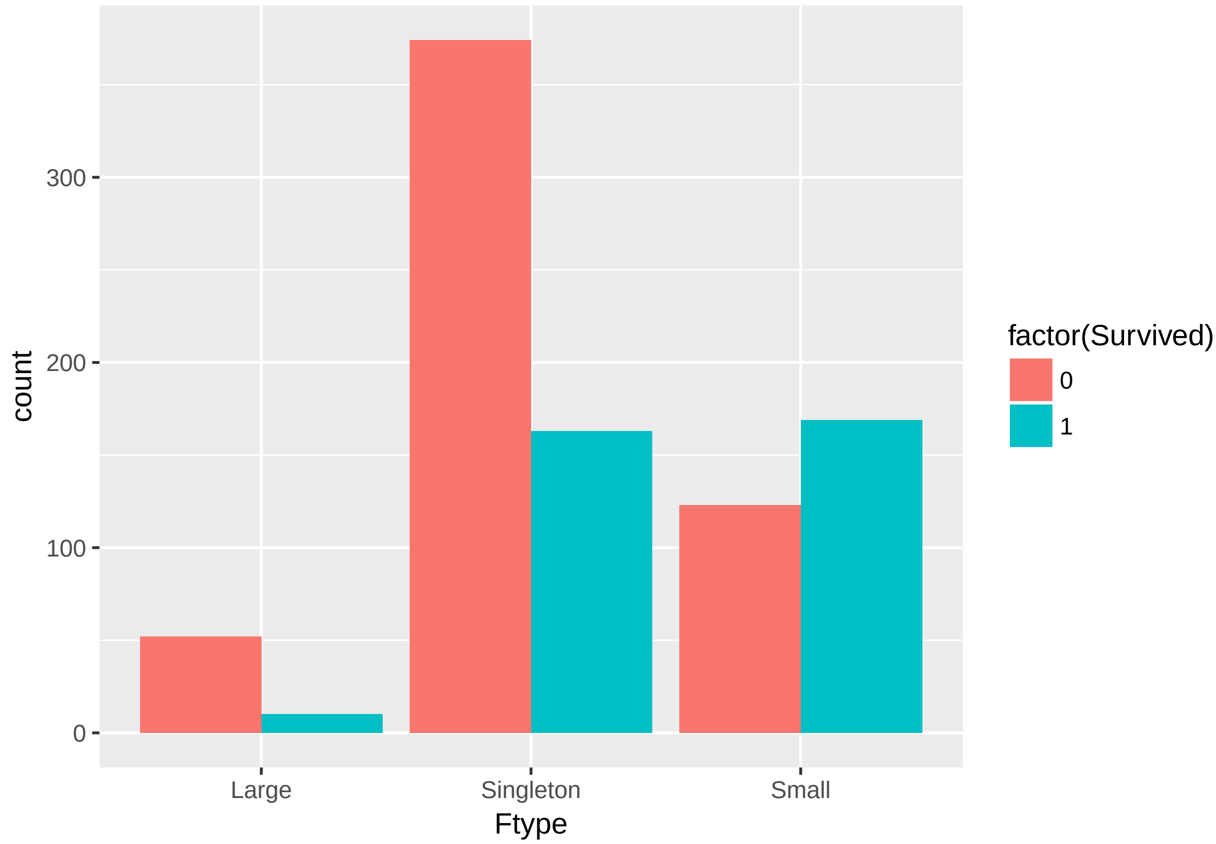
```
full$Fsize<-full$Parch+full$SibSp+1

ggplot()+geom_bar(data = full[1:891,],aes(x=Fsize,fill=factor(Survived)),position = 'dodge')
```




```
full$Ftype[full$Fsize==1]<-'Singleton'
full$Ftype[full$Fsize>1 & full$Fsize<5]<-'Small'
full$Ftype[full$Fsize>4]<-'Large'

ggplot()+geom_bar(data = full[1:891,],aes(x=Ftype,fill=factor(Survived)),position = 'dodge')
```



#******** Cabin **************

```
head(full$Cabin[2])
```

```
## [1] "C85"
```

```
strsplit(full$Cabin,NULL)[[1]][1]
```

```
## [1] NA
```

```
full$deck<-sapply(full$Cabin,FUN = function(x){strsplit(x,NULL)[[1]][1]})
```

******* Missing Value : Embarked *****************

```
head(full$Embarked)
```

```
## [1] "S" "C" "S" "S" "S" "Q"
```

```
which(full$Embarked=="")
```

```
## [1]  62 830
```

```
full[c(62,830),c(3,10,16)]
```

```
##     Pclass Fare deck
## 62       1   80    B
## 830      1   80    B
```

```r
full[(full$Pclass==1 & full$deck=="B" & full$Embarked=="C"),c(3,10,16,12)]
```

```
##        Pclass     Fare deck Embarked
## NA         NA       NA <NA>     <NA>
## 32          1 146.5208    B        C
## NA.1       NA       NA <NA>     <NA>
## 55          1  61.9792    B        C
## NA.2       NA       NA <NA>     <NA>
## 119         1 247.5208    B        C
## 140         1  79.2000    B        C
## NA.3       NA       NA <NA>     <NA>
## 195         1  27.7208    B        C
## 196         1 146.5208    B        C
## NA.4       NA       NA <NA>     <NA>
## NA.5       NA       NA <NA>     <NA>
## 292         1  91.0792    B        C
## NA.6       NA       NA <NA>     <NA>
## 300         1 247.5208    B        C
## NA.7       NA       NA <NA>     <NA>
## 312         1 262.3750    B        C
## 330         1  57.9792    B        C
## 370         1  69.3000    B        C
## NA.8       NA       NA <NA>     <NA>
## NA.9       NA       NA <NA>     <NA>
## NA.10      NA       NA <NA>     <NA>
## 485         1  91.0792    B        C
## 488         1  29.7000    B        C
## NA.11      NA       NA <NA>     <NA>
## NA.12      NA       NA <NA>     <NA>
## 524         1  57.9792    B        C
## NA.13      NA       NA <NA>     <NA>
## 540         1  49.5000    B        C
## NA.14      NA       NA <NA>     <NA>
## 588         1  79.2000    B        C
## NA.15      NA       NA <NA>     <NA>
## 633         1  30.5000    B        C
## 642         1  69.3000    B        C
## 680         1 512.3292    B        C
## 738         1 512.3292    B        C
## 743         1 262.3750    B        C
## NA.16      NA       NA <NA>     <NA>
## 790         1  79.2000    B        C
## NA.17      NA       NA <NA>     <NA>
## NA.18      NA       NA <NA>     <NA>
## NA.19      NA       NA <NA>     <NA>
## NA.20      NA       NA <NA>     <NA>
## 916         1 262.3750    B        C
## 918         1  61.9792    B        C
## 951         1 262.3750    B        C
## 956         1 262.3750    B        C
## NA.21      NA       NA <NA>     <NA>
## 1034        1 262.3750    B        C
## 1058        1  50.4958    B        C
## NA.22      NA       NA <NA>     <NA>
```

```
## 1076       1 247.5208    B        C
## NA.23     NA       NA <NA>     <NA>
## NA.24     NA       NA <NA>     <NA>
## NA.25     NA       NA <NA>     <NA>
## 1208       1 146.5208    B        C
## NA.26     NA       NA <NA>     <NA>
## 1235       1 512.3292    B        C
## NA.27     NA       NA <NA>     <NA>
## NA.28     NA       NA <NA>     <NA>
## 1289       1  79.2000    B        C
## NA.29     NA       NA <NA>     <NA>
```

```
#pclass<-train[train$Pclass==1 & train$Embarked!="",c(10,12) ]

#y_pred=lm(formula=Embarked ~ Fare,data =pclass)


full %>% filter(Pclass==1) %>%group_by(Pclass,Embarked)%>% summarise(mfare = median(Fare,na.rm=TRUE),n =
```

```
## # A tibble: 4 x 4
## # Groups:   Pclass [?]
##   Pclass Embarked   mfare     n
##    <int>    <chr>   <dbl> <int>
## 1      1          80.0000     2
## 2      1        C 76.7292   141
## 3      1        Q 90.0000     3
## 4      1        S 52.0000   177
```

```
full$Embarked[c(62,830)]<-'C'
```

***************** Missing Value : Fare *************************

```
summary(full$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   7.896  14.454  33.295  31.275 512.329       1
```

```
which(is.na(full$Fare))
```

```
## [1] 1044
```

```
full[1044,]
```

```
##      PassengerId Survived Pclass               Name  Sex  Age SibSp Parch
## 1044        1044       NA      3 Storey, Mr. Thomas male 60.5     0     0
##      Ticket Fare Cabin Embarked Title Fsize     Ftype deck
## 1044   3701   NA              S    Mr      1 Singleton <NA>
```

```
full %>% filter(Pclass=='3' & Embarked=='S') %>% summarise(median(Fare, na.rm = TRUE))
```

```
##   median(Fare, na.rm = TRUE)
## 1                       8.05
```

```
full$Fare[1044]<-8.05
```

***************** Missing Value : Age *****************

```
library(rpart)

summary(full$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.17   21.00   28.00   29.88   39.00   80.00     263
```

```
pred <- rpart(Age ~Pclass+SibSp+Embarked+Title,data = full[!is.na(full$Age),])

summary(pred)
```

```
## Call:
## rpart(formula = Age ~ Pclass + SibSp + Embarked + Title, data = full[!is.na(full$Age),
##     ])
##   n= 1046
##
##           CP nsplit rel error    xerror       xstd
## 1 0.21028409      0 1.0000000 1.0012796 0.04529407
## 2 0.10512853      1 0.7897159 0.7918756 0.03517463
## 3 0.05220533      2 0.6845874 0.6886374 0.03308297
## 4 0.02716919      3 0.6323821 0.6362894 0.03201778
## 5 0.01816094      4 0.6052129 0.6174335 0.03199064
## 6 0.01056208      5 0.5870519 0.5953919 0.03037402
## 7 0.01000000      6 0.5764899 0.5944076 0.03016726
##
## Variable importance
##    Title   Pclass    SibSp Embarked
##       55       29       10        5
##
## Node number 1: 1046 observations,    complexity param=0.2102841
##   mean=29.88114, MSE=207.5502
##   left son=2 (266 obs) right son=3 (780 obs)
##   Primary splits:
##       Title    splits as  LLRRR,   improve=0.210284100, (0 missing)
##       Pclass   < 1.5 to the right, improve=0.154604900, (0 missing)
##       SibSp    < 2.5 to the right, improve=0.071073330, (0 missing)
##       Embarked splits as  RLL,     improve=0.008481903, (0 missing)
##   Surrogate splits:
##       SibSp    < 2.5 to the right, agree=0.773, adj=0.109, (0 split)
##       Embarked splits as  RLR,     agree=0.748, adj=0.008, (0 split)
##
## Node number 2: 266 observations,    complexity param=0.05220533
##   mean=18.56831, MSE=164.0627
##   left son=4 (53 obs) right son=5 (213 obs)
##   Primary splits:
##       Title    splits as  LR---,   improve=0.25970370, (0 missing)
##       SibSp    < 0.5 to the right, improve=0.21272070, (0 missing)
##       Pclass   < 1.5 to the right, improve=0.19354290, (0 missing)
##       Embarked splits as  RRL,     improve=0.02984813, (0 missing)
##   Surrogate splits:
##       SibSp < 3.5 to the right, agree=0.831, adj=0.151, (0 split)
##
```

```
## Node number 3: 780 observations,    complexity param=0.1051285
##   mean=33.7391, MSE=163.8521
##   left son=6 (562 obs) right son=7 (218 obs)
##   Primary splits:
##       Pclass   < 1.5 to the right, improve=0.178578300, (0 missing)
##       Title    splits as  --LRR,   improve=0.039397110, (0 missing)
##       Embarked splits as  RRL,     improve=0.011405030, (0 missing)
##       SibSp    < 2.5 to the right, improve=0.006958206, (0 missing)
##   Surrogate splits:
##       Embarked splits as  RLL,   agree=0.767, adj=0.165, (0 split)
##       Title    splits as  --LLR, agree=0.731, adj=0.037, (0 split)
##
## Node number 4: 53 observations
##   mean=5.482642, MSE=16.99177
##
## Node number 5: 213 observations,    complexity param=0.02716919
##   mean=21.82437, MSE=147.4482
##   left son=10 (152 obs) right son=11 (61 obs)
##   Primary splits:
##       Pclass   < 1.5 to the right, improve=0.18780720, (0 missing)
##       SibSp    < 0.5 to the right, improve=0.14555750, (0 missing)
##       Embarked splits as  RRL,     improve=0.02453456, (0 missing)
##   Surrogate splits:
##       Embarked splits as  RLL, agree=0.775, adj=0.213, (0 split)
##
## Node number 6: 562 observations,    complexity param=0.01056208
##   mean=30.37011, MSE=116.7829
##   left son=12 (361 obs) right son=13 (201 obs)
##   Primary splits:
##       Pclass   < 2.5 to the right, improve=0.03493722, (0 missing)
##       Title    splits as  --LRR,   improve=0.02300209, (0 missing)
##       Embarked splits as  LRL,     improve=0.01586441, (0 missing)
##       SibSp    < 1.5 to the right, improve=0.01297640, (0 missing)
##   Surrogate splits:
##       Title splits as  --LRR, agree=0.669, adj=0.075, (0 split)
##
## Node number 7: 218 observations
##   mean=42.42431, MSE=180.5023
##
## Node number 10: 152 observations,    complexity param=0.01816094
##   mean=18.49072, MSE=115.3497
##   left son=20 (53 obs) right son=21 (99 obs)
##   Primary splits:
##       SibSp    < 0.5 to the right, improve=0.22487070, (0 missing)
##       Embarked splits as  LRR,     improve=0.06437730, (0 missing)
##       Pclass   < 2.5 to the right, improve=0.02326302, (0 missing)
##   Surrogate splits:
##       Embarked splits as  LRR, agree=0.678, adj=0.075, (0 split)
##
## Node number 11: 61 observations
##   mean=30.13115, MSE=130.7369
##
## Node number 12: 361 observations
##   mean=28.86288, MSE=100.2727
```

```
##
## Node number 13: 201 observations
##   mean=33.07711, MSE=135.0276
##
## Node number 20: 53 observations
##   mean=11.53, MSE=90.38458
##
## Node number 21: 99 observations
##   mean=22.21717, MSE=88.88971
```

```
y_pred<-predict(pred,newdata =full[is.na(full$Age),] )

summary(y_pred)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.483  28.863  28.863  28.427  28.863  42.424
```

```
full$Age[is.na(full$Age)]<-predict(pred,newdata =full[is.na(full$Age),] )
```

**************** Random Forest Model ****************

```
full$Sex<-factor(full$Sex)
full$Embarked<-factor(full$Embarked)
full$Title<-factor(full$Title)
full$Ftype<-factor(full$Ftype)

train <- full[1:891,]
test <- full[892:1309,]

set.seed(123)
#train$Sex<-factor(train$Sex)
##train$Embarked<-factor(train$Embarked)
#train$Title<-factor(train$Title)
#train$Ftype<-factor(train$Ftype)

#test$Sex<-factor(test$Sex)
#test$Embarked<-factor(test$Embarked)
#test$Title<-factor(test$Title)
#test$Ftype<-factor(test$Ftype)

#rf_model<-randomForest(factor(Survived)~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked+Title+Fsize+Ftype,dat

rf_model<-randomForest(factor(Survived)~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked+Title+Fsize+Ftype,data

rf_model  #740 right
```

```
##
## Call:
##  randomForest(formula = factor(Survived) ~ Pclass + Sex + Age +      SibSp + Parch + Fare + Embarked
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
```

```
##         OOB estimate of  error rate: 17.06%
## Confusion matrix:
##     0   1 class.error
## 0 493  56   0.1020036
## 1  96 246   0.2807018
```

```r
pred<-predict(rf_model,test)

#solution <- data.frame(PassengerID = test$PassengerId, Survived = pred)

# Write the solution to file
#write.csv(solution, file = 'titanic_2.csv', row.names = F)
```