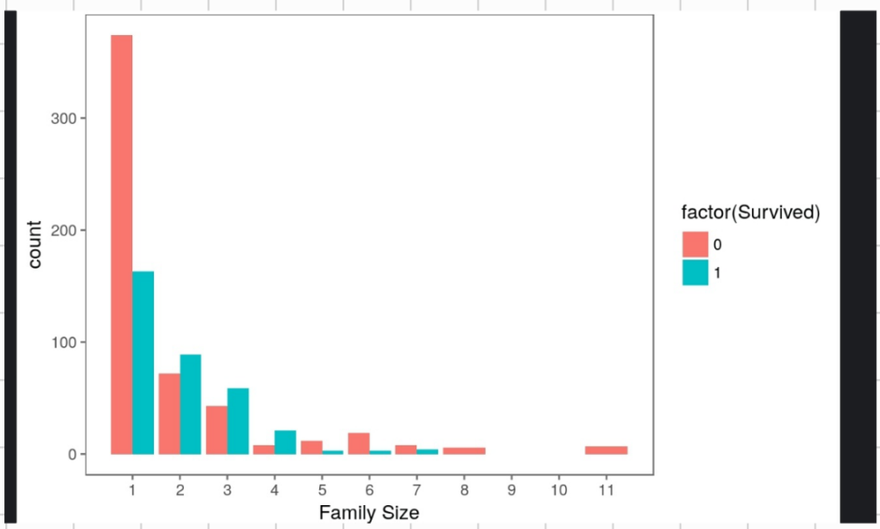
https://www.kaggle.com/code/mrisdal/exploring-survival-on-the-titanic

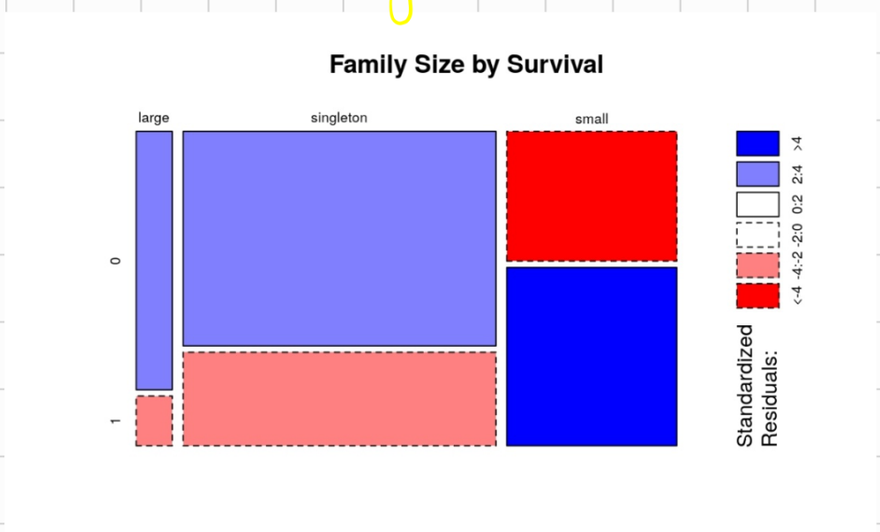
Variable Name	Description
Survived	Survived (1) or died (0)
Pclass	Passenger's class
Name	Passenger's name
Sex	Passenger's sex
Age	Passenger's age
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Fare
Cabin	Cabin
Embarked	Port of embarkation

- extract title and family from name
- create fam size var



fam size < 2 or > 4 less survivability

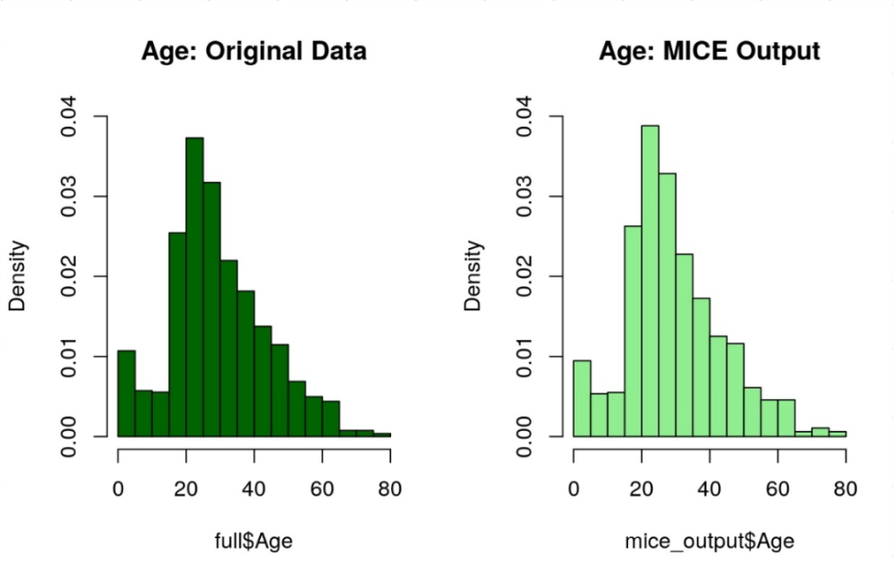
↳ we can create discretized family size feature (3 groups)
↳ singleton, small, large



- create Deck feature from cabin

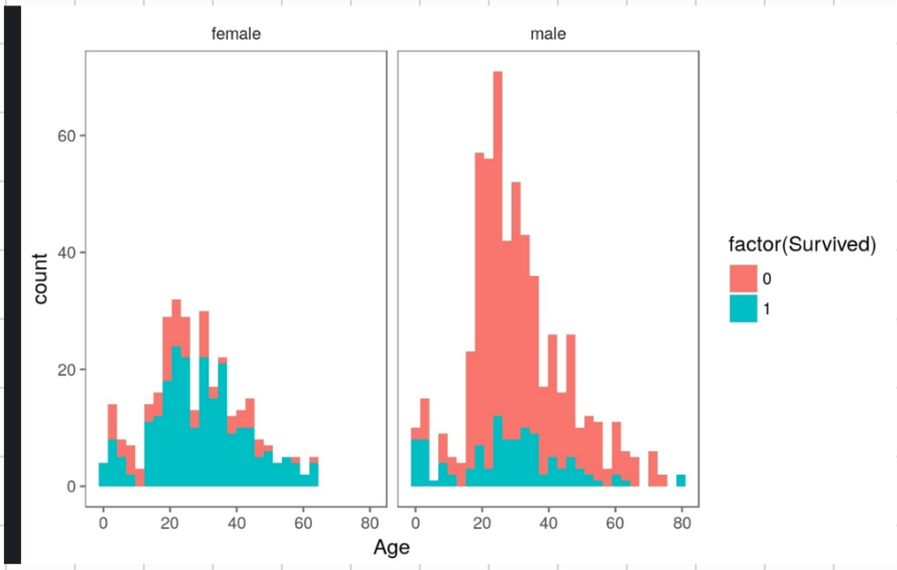
→ Missing values

- since small dataset better not to delete any samples or features
- for Age, lots of missing data. To address, we can create model that predicts age based on other features



→ make sure our Age predictions don't corrupt the initial Age distribution

- create new feature from age:
  - child ⇒ age < 18
  - mother ⇒ female ⊕ >18 ⊕ children > 0 ⊕ not Miss title



```r
# Create the column child, and indicate whether child or adult
full$Child[full$Age < 18] <- 'Child'
full$Child[full$Age >= 18] <- 'Adult'

# Show counts
table(full$Child, full$Survived)
```

child feature

```
##
##            0   1
##   Adult 484 274
##   Child  65  68
```

```r
# Adding Mother variable
full$Mother <- 'Not Mother'
full$Mother[full$Sex == 'female' & full$Parch > 0 & full$Age > 18 & full$Title
!= 'Miss'] <- 'Mother'

# Show counts
table(full$Mother, full$Survived)
```

mother feature

```
##
##                0   1
##   Mother      16  39
##   Not Mother 533 303
```

```r
# Finish by factorizing our two new factor variables
full$Child  <- factor(full$Child)
full$Mother <- factor(full$Mother)
```

→ Prediction

4.2 Building the model

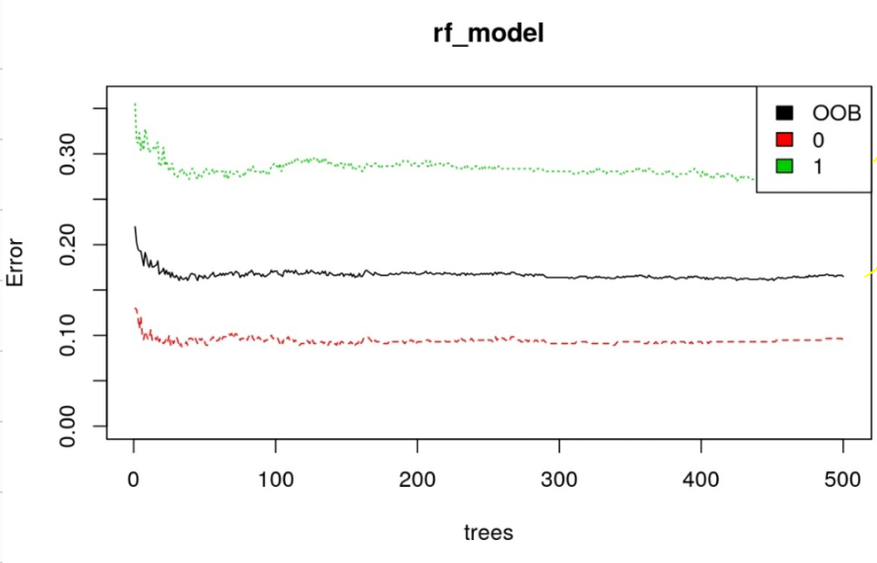
We then build our model using `randomForest` on the training set.

```r
# Set a random seed
set.seed(754)

# Build the model (note: not all possible variables are used)
rf_model <- randomForest(factor(Survived) ~ Pclass + Sex + Age + SibSp + Parch
+
                                            Fare + Embarked + Title +
                                            FsizeD + Child + Mother,
                                            data = train)

# Show model error
plot(rf_model, ylim=c(0,0.36))
legend('topright', colnames(rf_model$err.rate), col=1:3, fill=1:3)
```

Random Forest



error rate in survived preds

overall error rate

error rate in died preds

• feature importance