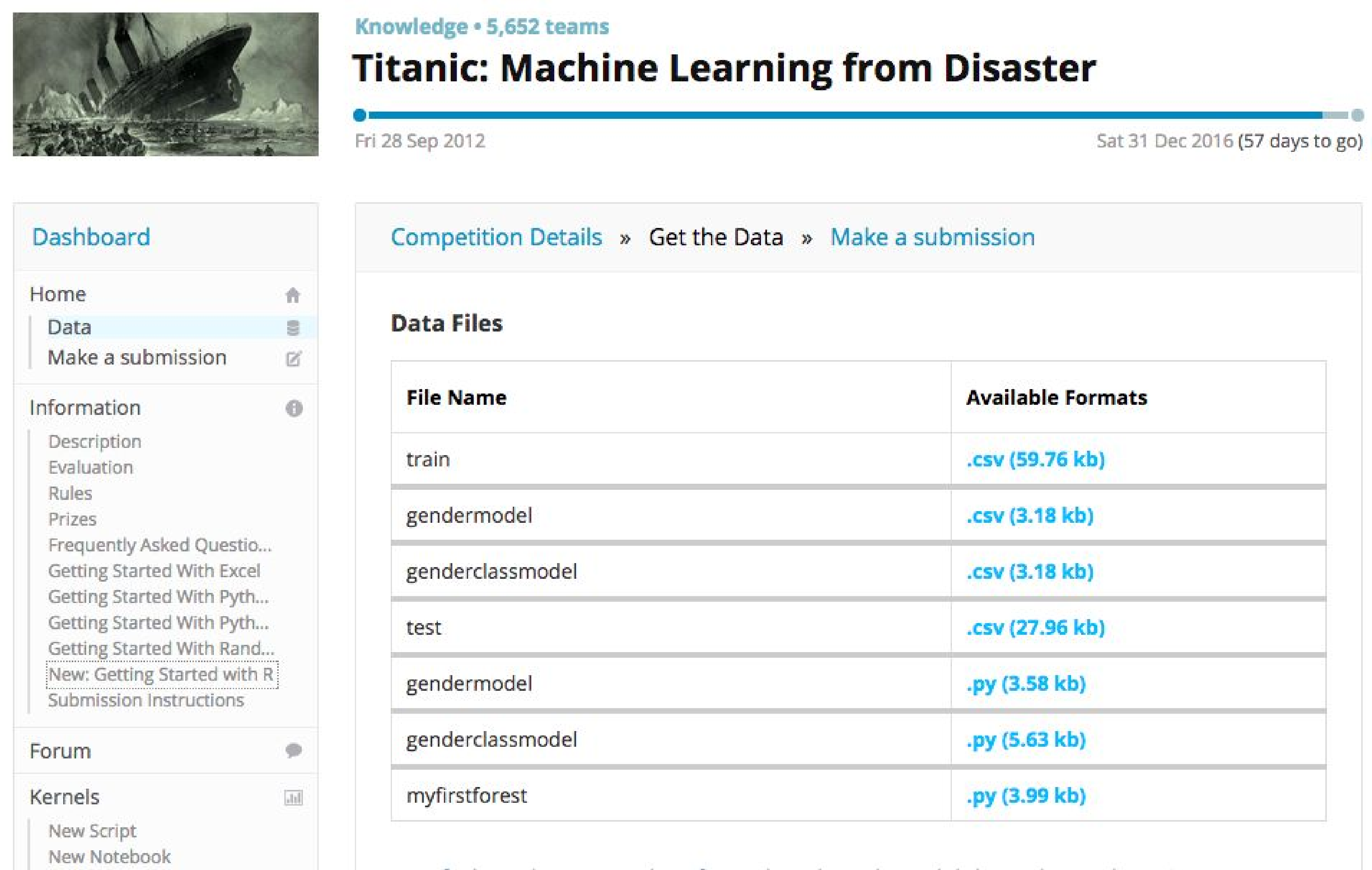
Data Preprocessing

•••

The Dataset

The Titanic Competition from Kaggle

- https://www.kaggle.com/c/titanic/data
- The Task
  - Predict Survival



Knowledge • 5,652 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Sat 31 Dec 2016 (57 days to go)

Dashboard
Home
Data
Make a submission
Information
Description
Evaluation
Rules
Prizes
Frequently Asked Questio...
Getting Started With Excel
Getting Started With Pyth...
Getting Started With Pyth...
Getting Started With Rand...
New: Getting Started with R
Submission Instructions
Forum
Kernels
New Script
New Notebook

Competition Details » Get the Data » Make a submission

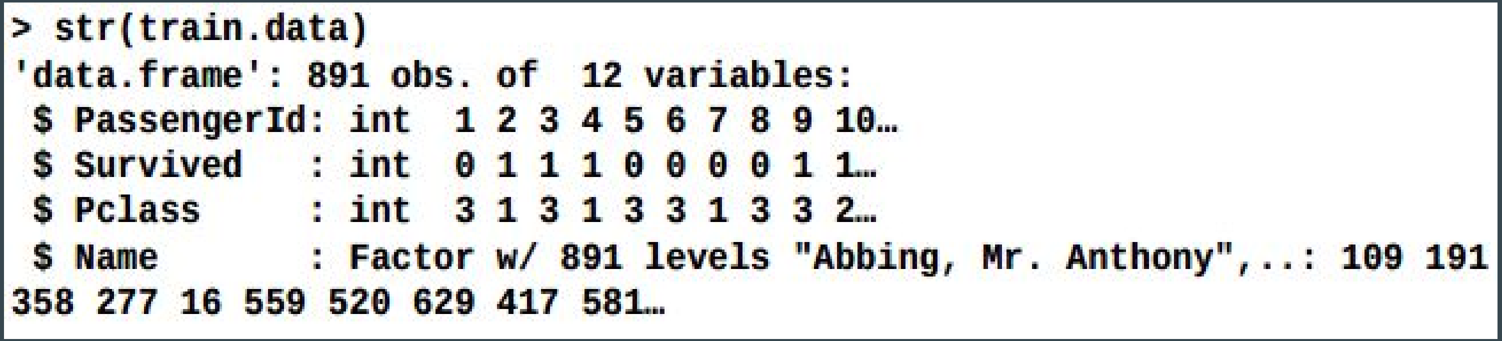
Data Files

File Name	Available Formats
train	.csv (59.76 kb)
gendermodel	.csv (3.18 kb)
genderclassmodel	.csv (3.18 kb)
test	.csv (27.96 kb)
gendermodel	.py (3.58 kb)
genderclassmodel	.py (5.63 kb)
myfirstforest	.py (3.99 kb)

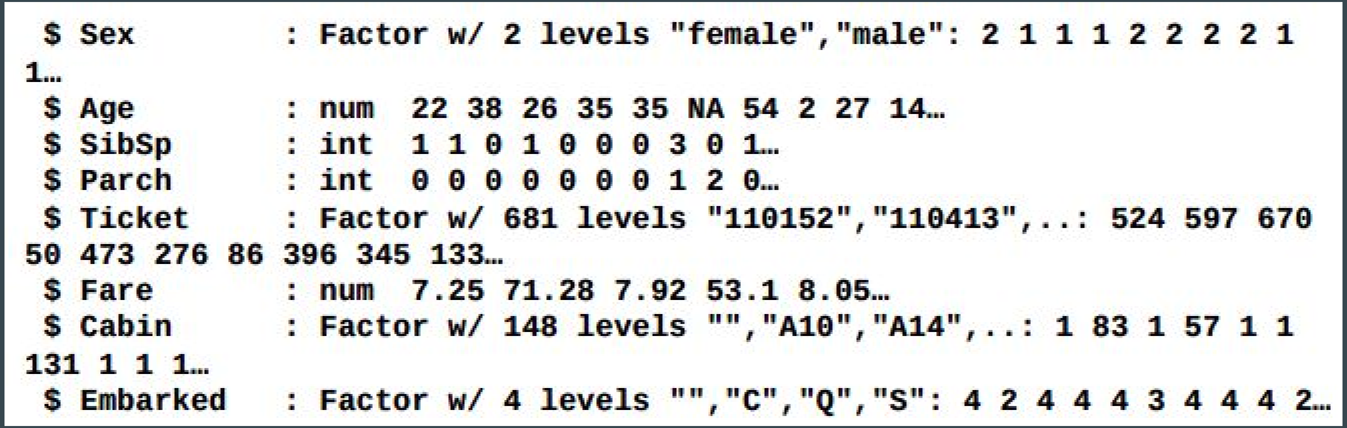
Example: Find Missing Values

Read Data

```
train.data = read.csv("train.csv", na.strings=c("NA", ""))

str(train.data)
```



```
> str(train.data)
'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…
```

Factor to Strings

```
train.data$Survived = factor(train.data$Survived)

train.data$Pclass = factor(train.data$Pclass)
```

Find Missing Values

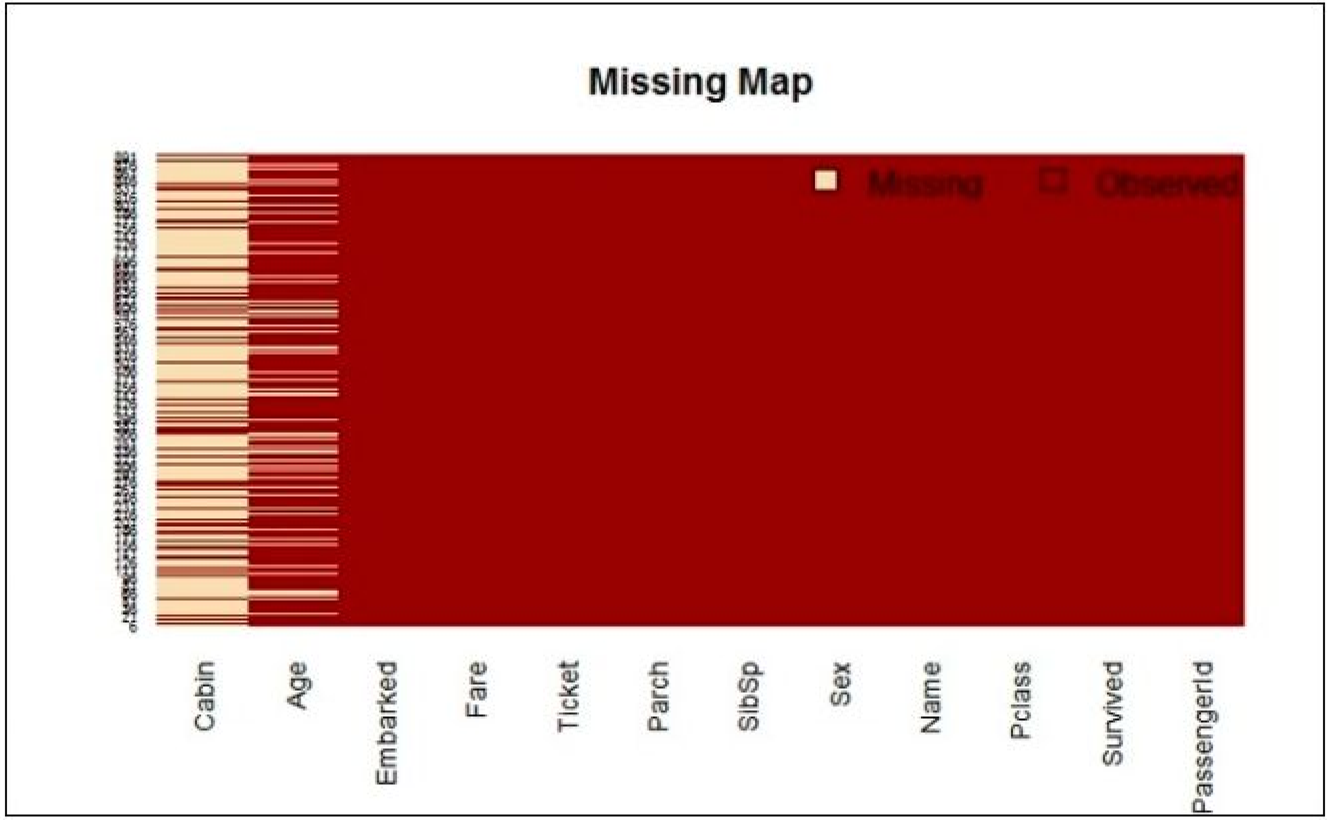
```
is.na(train.data$Age)

sum(is.na(train.data$Age) == TRUE)

sum(is.na(train.data$Age) == TRUE) / length(train.data$Age)

sapply(train.data, function(df) {

    sum(is.na(df)==TRUE)/ length(df)

})
```

Find Missing Values

```
library(YaleToolkit)

whatis(train)

whatis(test)
```

Find Missing Values

```
install.packages("Amelia")

require(Amelia)

missmap(train.data, main="Missing Map")
```


Missing Map
Missing
Observed
Cabin
Age
Embarked
Fare
Ticket
Parch
SibSp
Sex
Name
Pclass
Survived
PassengerId

Existing Guides

Walkthrough

1

- https://www.kaggle.com/benhamner/titanic/random-forest-benchmark-r

Walkthrough

2

- https://github.com/wehrley/wehrley.github.io/blob/master/SOUPTONUTS.md

Walkthrough

3

- http://trevorstephens.com/kaggle-titanic-tutorial/r-part-1-booting-up/
- http://trevorstephens.com/kaggle-titanic-tutorial/r-part-2-the-gender-class-model/
- http://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/
- http://trevorstephens.com/kaggle-titanic-tutorial/r-part-4-feature-engineering/
- http://trevorstephens.com/kaggle-titanic-tutorial/r-part-5-random-forests/