Análisis de Datos con R

Sobre mi…

Ana Valdivia

Analista de datos en la UGR

I DATA

Coorganizadora de @DataBeersGRX

DATA BEERS GRX

¿Qué es R?

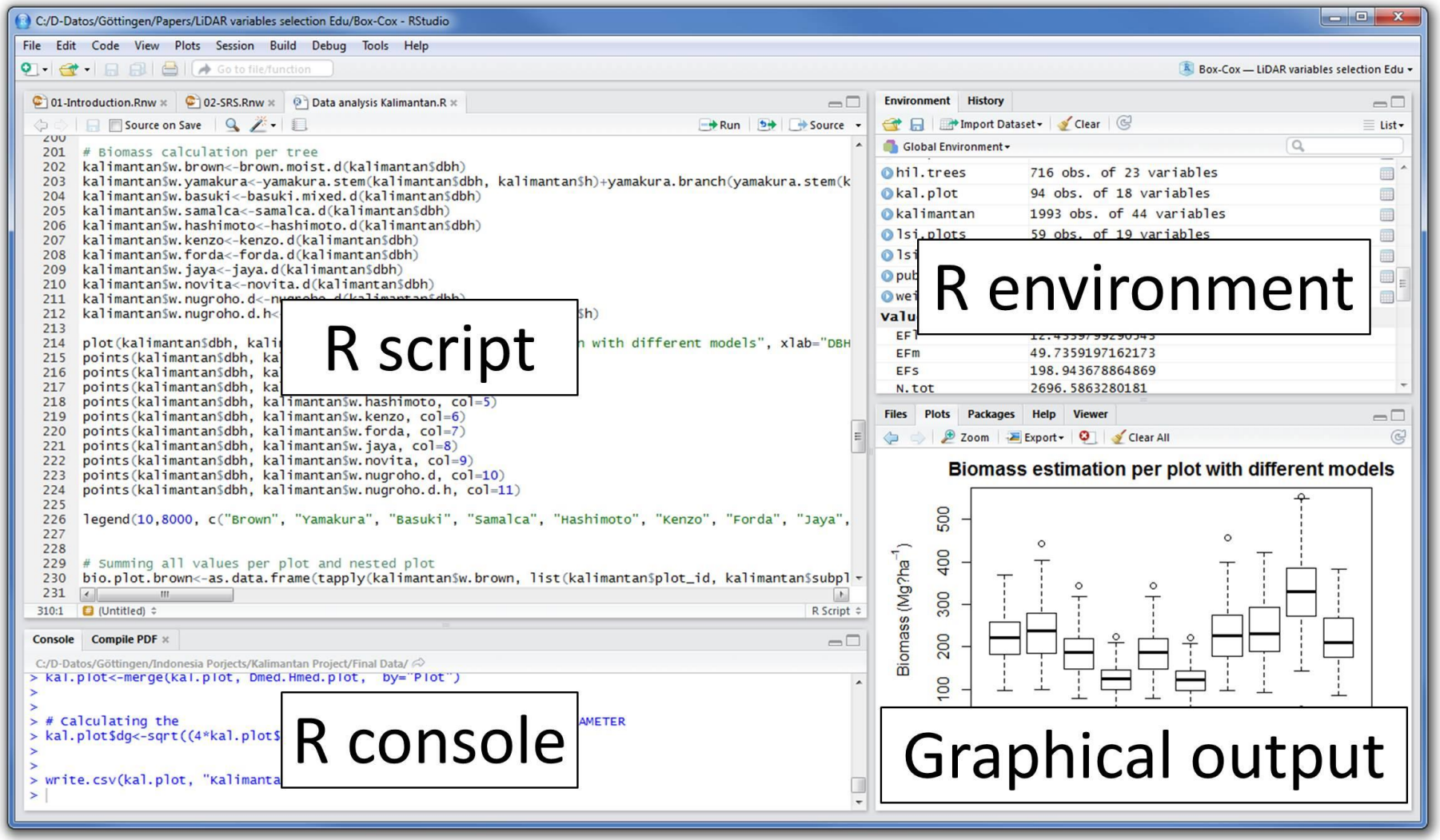
¿Qué es R?

Es un lenguaje y entorno para cálculos estadísticos y gráficos.

Es un proyecto GNU!!!!

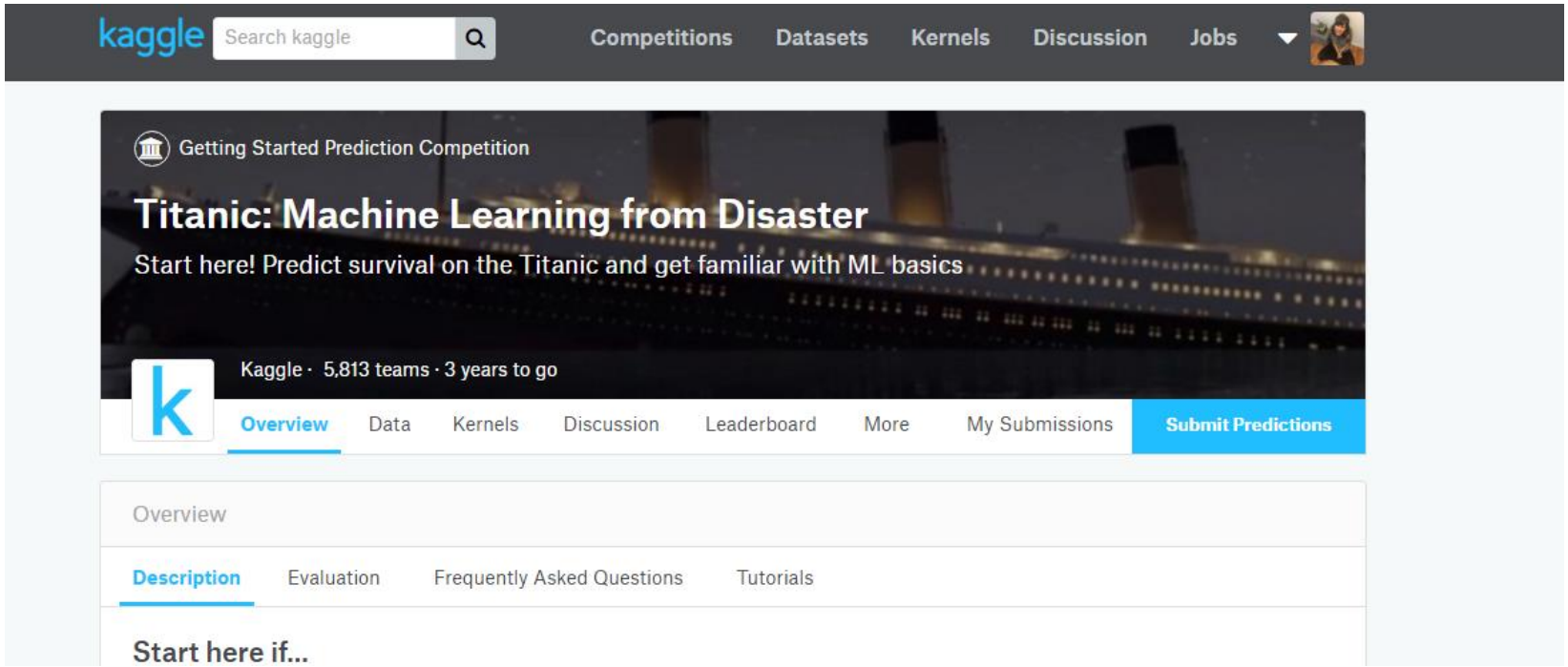
¿Qué es R?

¿Qué es R?


C:/D-Datos/Göttingen/Papers/LiDAR variables selection Edu/Box-Cox - RStudio
File Edit Code View Plots Session Build Debug Tools Help
01-Introduction.Rnw
02-SRS.Rnw
Data analysis Kalimantan.R
Source on Save
Run
Source
Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yamakura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)
legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",
# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
R script
Console
Compile PDF
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
R console
Environment
History
Import Dataset
Clear
Global Environment
hil.trees 716 obs. of 23 variables
kal.plot 94 obs. of 18 variables
kalimantan 1993 obs. of 44 variables
EFm 49.7359197162173
EFs 198.943678864869
N.tot 2696.5863280181
R environment
Files
Plots
Packages
Help
Viewer
Zoom
Export
Clear All
Biomass estimation per plot with different models
Biomass (Mg?ha⁻¹)
100
200
300
400
500
Graphical output

Titanic

Titanic


kaggle
Search kaggle
Competitions
Datasets
Kernels
Discussion
Jobs
Getting Started Prediction Competition
Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Kaggle · 5,813 teams · 3 years to go
Overview
Data
Kernels
Discussion
Leaderboard
More
My Submissions
Submit Predictions
Overview
Description
Evaluation
Frequently Asked Questions
Tutorials
Start here if...

1. Importar el dataset

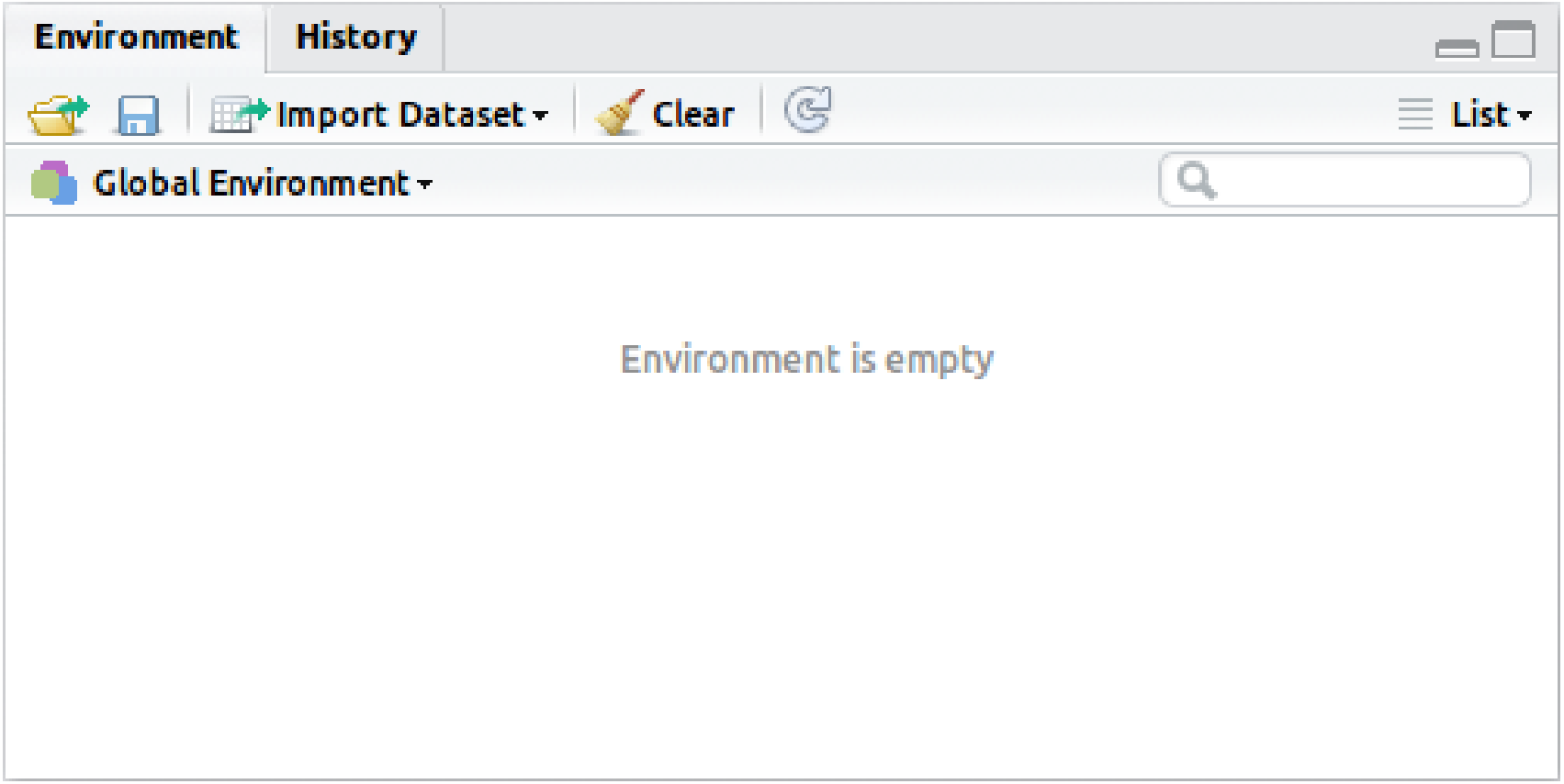
1. Importar el dataset

read.csv()

```
> train <- read.csv("train.csv", stringsAsFactors=FALSE)
```

1. Importar el dataset

Import Dataset



1. Importar el dataset

Tutorial1.R × train ×

891 observations of 12 variables

	PassengerId	Survived	Pclass	Name
1	1	0	3	Braund, Mr. Owen Harris
2	2	1	1	Cumings, Mrs. John Bradley (Florence
3	3	1	3	Heikkinen, Miss. Laina
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May
5	5	0	3	Allen, Mr. William Henry
6	6	0	3	Moran, Mr. James
7	7	0	1	McCarthy, Mr. Timothy J
8	8	0	3	Palsson, Master. Gosta Leonard
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilh
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)
11	11	1	3	Sandstrom, Miss. Marguerite Rut
12	12	1	1	Bonnell, Miss. Elizabeth
13	13	0	3	Saundercock, Mr. William Henry
14	14	0	3	Andersson, Mr. Anders Johan
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfir

2. Analizar estructura del dataset

2. Analizar estructura del dataset

str()

```
> str(train)
'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 16 559 520 629 416 581 ...
 $ 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",..: 525 596 662 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 ...
```

2. Analizar estructura del dataset

table ()

```
> table(train$Survived)
  0   1
549 342
```

prop.table()

```
> prop.table(table(train$Survived))
        0         1
0.6161616 0.3838384
```

2. Analizar estructura del dataset

summary ()

```
> summary(train$Sex)
female   male
   314    577
```

```
> summary(train$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.42   20.12   28.00   29.70   38.00   80.00     177
```

2. Analizar estructura del dataset

creating variables

```
> prop.table(table(train$Sex, train$Survived))
                 0          1
  female 0.09090909 0.26150393
  male   0.52525253 0.12233446
```

```
> test$Survived <- 0
> test$Survived[test$Sex == 'female'] <- 1
```

3. Juega con los datos

Step 3

3. Juega con los datos

Crea nuevas variables

```
> summary(train$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.42   20.12   28.00   29.70   38.00   80.00     177
```

```
> train$Child <- 0
> train$Child[train$Age < 18] <- 1
```

3. Juega con los datos

aggregate()

```
> 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
```

```
> 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
```

3. Juega con los datos

aggregate()

```
> aggregate(Survived ~ Child + Sex, data=train, FUN=function(x) {sum(x)/length(x)})
  Child    Sex  Survived
1     0 female 0.7528958
2     1 female 0.6909091
3     0   male 0.1657033
4     1   male 0.3965517
```

3. Juega con los datos

install.packages()
library()

```
> install.packages('rattle')
> install.packages('rpart.plot')
> install.packages('RColorBrewer')
> library(rattle)
> library(rpart.plot)
> library(RColorBrewer)
```

3. Juega con los datos

caret

The caret Package

The caret package (short for *C*lassification *A*nd *RE*gression *T*raining) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

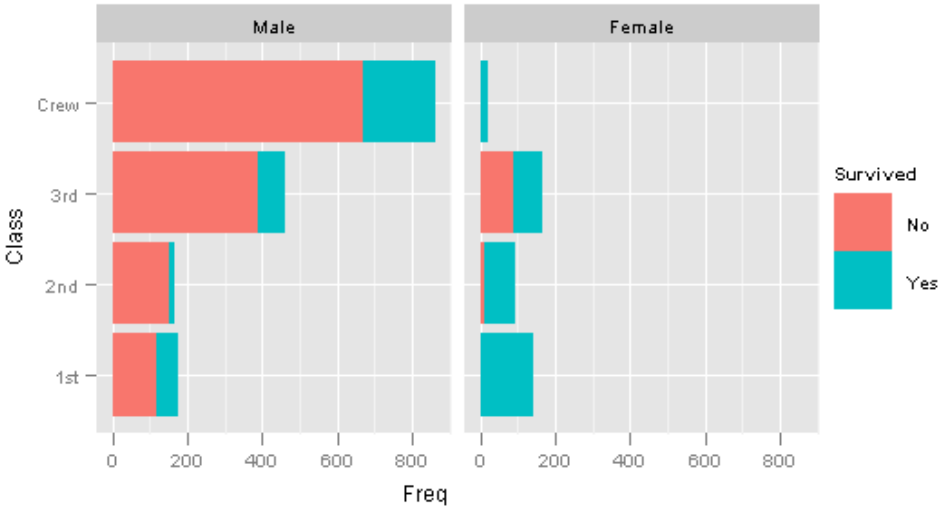
- data splitting

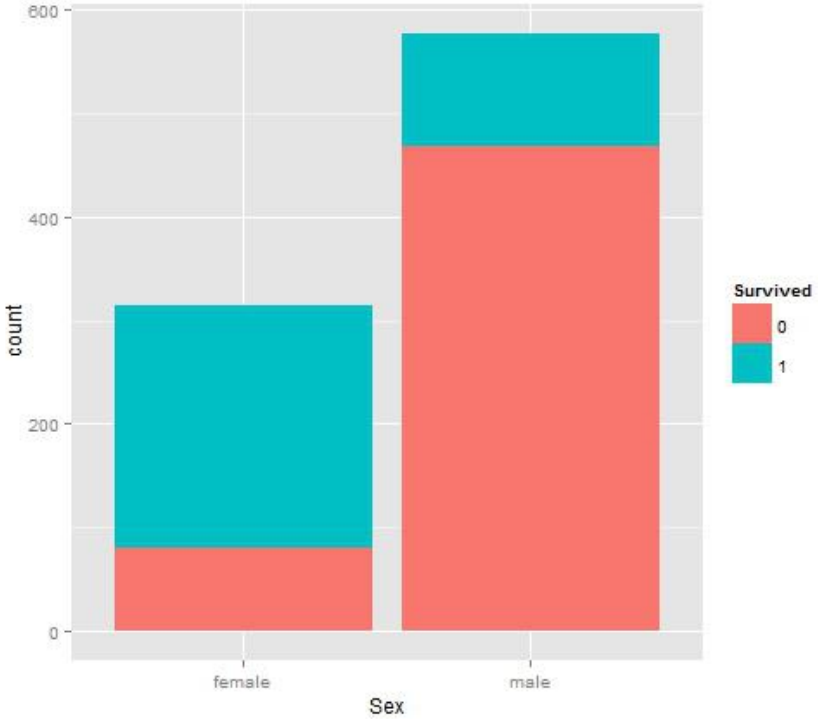
General Topics

- Front Page
- Visualizations
- Pre-Processing
- Data Splitting
- Variable Importance
- Model Performance
- Parallel Processing

3. Juega con los datos

ggplot2





3. Juega con los datos… ¡Pero no naufragues!

Fuentes

Titanic: Getting Started With R

http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/

Trevor Stephens

Regular Data Scientist,
Occasional Blogger.

San Francisco, CA

Website

Twitter

LinkedIn

Github

bugs or typos, or have any suggestions on making the tutorial
follow, please send me a direct message through Twitter. All c
on my Github repository.

I will be dividing this series of tutorials into five parts:

- Part 1: Booting Up R
- Part 2: The Gender-Class Model
- Part 3: Decision Trees
- Part 4: Feature Engineering
- Part 5: Random Forests

So go ahead and get started with part 1

Si quieres saber más…

Titanic: Getting Started With R

http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/

Coursera – Machine Learning

https://www.coursera.org/learn/machine-learning

Si quieres saber más…

@ana_valdi

avaldivia@ugr.es