Data collection

Ajayi Olabode

15 February 2017

We're working with data on who survived the Titanic.

Source

We're collecting our data from a SQLite database. The titanic3 data was originally pulled in from the PASWR package and is the third major version of that dataset. It contains more features than the basic titanic dataset available in the datasets package.

```
library(DBI)
library(RSQLite)
titanicdb<-dbConnect(SQLite(),dbname="../data-raw/titanic.sqlite")
```

Data

We're using just a single table of data that has already been collated. Here is a quick overview of the data.

```
titanic_all<-dbReadTable(titanicdb, "titanic")
knitr::kable(head(titanic_all))
```

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	e
1st	1	Allen, Miss. Elisabeth Walton	female	29.0000	0	0	24160	211.3375	B5	S
1st	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S
1st	0	Allison, Miss. Helen Loraine	female	2.0000	1	2	113781	151.5500	C22 C26	S
1st	0	Allison, Mr. Hudson Joshua Crei	male	30.0000	1	2	113781	151.5500	C22 C26	S
1st	0	Allison, Mrs. Hudson J C (Bessi	female	25.0000	1	2	113781	151.5500	C22 C26	S
1st	1	Anderson, Mr. Harry	male	48.0000	0	0	19952	26.5500	E12	S

```
knitr::kable(summary(titanic_all))
```

pclass	survived	name	sex	age	sibsp	parch
Length:1309	Min. :0.000	Length:1309	Length:1309	Min. : 0.1667	Min. :0.0000	Min. :0.0
Class :character	1st Qu.:0.000	Class :character	Class :character	1st Qu.:21.0000	1st Qu.:0.0000	1st Qu.:0.0
Mode :character	Median :0.000	Mode :character	Mode :character	Median :28.0000	Median :0.0000	Median :0.
NA	Mean :0.382	NA	NA	Mean :29.8811	Mean :0.4989	Mean :0.3
NA	3rd Qu.:1.000	NA	NA	3rd Qu.:39.0000	3rd Qu.:1.0000	3rd Qu.:0.0
NA	Max. :1.000	NA	NA	Max. :80.0000	Max. :8.0000	Max. :9.0
NA	NA	NA	NA	NA's :263	NA	NA

Defensive stuff

Store a copy (one-off)

```
cache_file<-"../data-raw/rawdatacache.Rdata"
if(!file.exists(cache_file)) {
  titanic_cache<-titanic_all
  save(titanic_cache,file = cache_file)
  rm(titanic_cache)
  }
```

Check for changes

```
load(cache_file)
if(!identical(titanic_all, titanic_cache)) stop("Hey, the data has changed, you should check that out!")
```

We could change things and rerun with:

```
dbWriteTable(titanicdb, "titanic", head(titanic_all), append=TRUE)
```

Restore your original data with your get_data.R file.