

Data Exploration with the Titanic Data

Karen Mazidi

Load the data

Next we use the `read.csv()` function to read a csv in a subdirectory called data. Once you read in the data you will see that it has 1310 observations of 14 variables. We run the `str()` structure function to get a peek at the data.

```
df <- read.csv("data/titanic3.csv", na.strings="NA", stringsAsFactors=FALSE, header=TRUE)
str(df)
```

```
## 'data.frame':    1310 obs. of  14 variables:
## $ pclass      : int   1 1 1 1 1 1 1 1 1 1 ...
## $ survived    : int   1 1 0 0 0 1 1 0 1 0 ...
## $ name        : chr   "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. L
## $ sex         : chr   "female" "male" "female" "male" ...
## $ age         : num   29 0.917 2 30 25 ...
## $ sibsp       : int   0 1 1 1 1 0 1 0 2 0 ...
## $ parch       : int   0 2 2 2 2 0 0 0 0 0 ...
## $ ticket      : chr   "24160" "113781" "113781" "113781" ...
## $ fare        : num   211 152 152 152 152 ...
## $ cabin       : chr   "B5" "C22 C26" "C22 C26" "C22 C26" ...
## $ embarked    : chr   "S" "S" "S" "S" ...
## $ boat        : chr   "2" "11" "" "" ...
## $ body        : int   NA NA NA 135 NA NA NA NA NA 22 ...
## $ home.dest    : chr   "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON"
```

Data cleaning

The `read.csv()` function is a bit aggressive about making things factors. Generally if the column contains character data, it tries to make it a factor. Sometimes this makes sense, sometimes it does not.

We can change a column to a factor with `as.factor()` or change a column to integer with `as.integer()` as shown next.

```
df$survived <- as.factor(df$survived)
df$pclass <- as.factor(df$pclass)
df$sex <- factor(df$sex, levels=c("male", "female"))
```

Factors

Factors are stored internally as integer vectors but also have a character representation for human readability. We can use `contrasts()` to find out more about a factor column.

The contrasts for `pclass` shows that we need 2 variables to encode 3 classes. The base case will be class 1. R will create 2 dummy variables for classes 2 and 3. We will see the importance of these when we get to machine learning.

```
contrasts(df$pclass)
```

```
##      2 3
## 1 0 0
```

```
## 2 1 0
## 3 0 1
```

```
contrasts(df$sex)
```

```
##          female
## male          0
## female        1
```

More exploration

The head() and tail() functions let us look at the first or last few rows.

```
head(df)
```

```
##   pclass survived          name    sex
## 1      1         1    Allen, Miss. Elisabeth Walton female
## 2      1         1  Allison, Master. Hudson Trevor   male
## 3      1         0    Allison, Miss. Helen Loraine female
## 4      1         0  Allison, Mr. Hudson Joshua Creighton male
## 5      1         0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1         1  Anderson, Mr. Harry             male
##      age sibsp parch ticket   fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375    B5         S     2   NA
## 2  0.9167     1     2  113781 151.5500 C22 C26         S    11   NA
## 3  2.0000     1     2  113781 151.5500 C22 C26         S     NA
## 4 30.0000     1     2  113781 151.5500 C22 C26         S    135
## 5 25.0000     1     2  113781 151.5500 C22 C26         S     NA
## 6 48.0000     0     0  19952  26.5500   E12         S     3   NA
##      home.dest
## 1           St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6           New York, NY
```

```
tail(df, n=10)
```

```
##   pclass survived          name    sex  age
## 1301     3         1 Yasbeck, Mrs. Antoni (Selini Alexander) female 15.0
## 1302     3         0    Youseff, Mr. Gerious   male 45.5
## 1303     3         0    Yousif, Mr. Wazli    male  NA
## 1304     3         0  Yousseff, Mr. Gerious   male  NA
## 1305     3         0    Zabour, Miss. Hileni female 14.5
## 1306     3         0    Zabour, Miss. Thamine female  NA
## 1307     3         0  Zakarian, Mr. Mapriededer   male 26.5
## 1308     3         0    Zakarian, Mr. Ortin    male 27.0
## 1309     3         0    Zimmerman, Mr. Leo    male 29.0
## 1310    <NA>     <NA>                <NA>    NA
##      sibsp parch ticket   fare cabin embarked boat body home.dest
## 1301     1     0  2659 14.4542    C     NA
## 1302     0     0  2628  7.2250    C    312
## 1303     0     0  2647  7.2250    C     NA
## 1304     0     0  2627 14.4583    C     NA
## 1305     1     0  2665 14.4542    C    328
```

```
## 1306      1      0  2665 14.4542      C      NA
## 1307      0      0  2656  7.2250      C    304
## 1308      0      0  2670  7.2250      C      NA
## 1309      0      0 315082  7.8750      S      NA
## 1310     NA     NA      NA      NA      NA
```

The `summary()` function can summarize an entire data set or individual columns.

```
summary(df)
```

```
##   pclass   survived      name      sex      age
## 1   :323    0   :809  Length:1310   male :843  Min.   : 0.1667
## 2   :277    1   :500  Class :character female:466 1st Qu.:21.0000
## 3   :709   NA's: 1   Mode  :character  NA's  : 1   Median :28.0000
## NA's: 1                                     Mean   :29.8811
##                                           3rd Qu.:39.0000
##                                           Max.   :80.0000
##                                           NA's   :264
##      sibsp      parch      ticket      fare
## Min.   :0.0000  Min.   :0.000  Length:1310  Min.   : 0.000
## 1st Qu.:0.0000  1st Qu.:0.000  Class :character 1st Qu.:  7.896
## Median :0.0000  Median :0.000  Mode  :character Median : 14.454
## Mean    :0.4989  Mean    :0.385                Mean   : 33.295
## 3rd Qu.:1.0000  3rd Qu.:0.000                3rd Qu.: 31.275
## Max.    :8.0000  Max.    :9.000                Max.   :512.329
## NA's    :1      NA's    :1                NA's   :2
##      cabin      embarked      boat      body
## Length:1310    Length:1310    Length:1310  Min.   : 1.0
## Class :character Class :character Class :character 1st Qu.: 72.0
## Mode  :character Mode  :character Mode  :character Median :155.0
##                                           Mean   :160.8
##                                           3rd Qu.:256.0
##                                           Max.   :328.0
##                                           NA's   :1189
##      home.dest
## Length:1310
## Class :character
## Mode  :character
##
##
##
##
```

```
summary(df$pclass)
```

```
##      1      2      3 NA's
## 323  277  709      1
```

The `names()` function is helpful if you forget the column names.

```
names(df)
```

```
## [1] "pclass" "survived" "name" "sex" "age"
## [6] "sibsp" "parch" "ticket" "fare" "cabin"
## [11] "embarked" "boat" "body" "home.dest"
```

```
summary(df$age)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.      NA's
##  0.1667 21.0000 28.0000 29.8811 39.0000 80.0000      264
```

That's all for now. We will revisit the Titanic data later when we explore classification algorithms: learning how to predict who survived and who didn't based on demographic data in the file.