R Notebook

Karen Mazidi

Data exploration in R with the Titanic data set.

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. 
##  $ 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$pclass <- as.factor(df$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
```

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.