titanic

milti leonard

9/18/2017

Project: Data Wrangling Exercise 2: Dealing with missing values

2 - 3 Hours

In this exercise, you'll work with one of the most popular starter data sets in data science, the Titanic data set. This is a data set that records various attributes of passengers on the Titanic, including who survived and who didn't.

Getting started

```
Read the description of the data set on the Kaggle website.

Download the data as an excel file here.
```

Exercise

Using R, you'll be handling missing values in this data set, and creating a new data set. Specifically, these are the tasks you need to do:

0: Load the data in RStudio

Save the data set as a CSV file called titanic_original.csv and load it in RStudio into a data frame.

```
# titanic <- read_xls('~/R/github/springboard\
# exercises/springboard\ exercises/titanic3.xls')
# write.csv(titanic, '~/R/github/springboard\
# exercises/springboard\ exercises/titanic_original.csv')

titanic <- read.csv("~/R/github/springboard exercises/springboard exercises/titanic_original.csv")
```

1: Port of embarkation

The embarked column has some missing values, which are known to correspond to passengers who actually embarked at Southampton. Find the missing values and replace them with S. (Caution: Sometimes a missing value might be read into R as a blank or empty string.)

```
omitted.embarked <- which(is.na(titanic$embarked))
titanic$embarked[omitted.embarked] = "S"
```

2: Age

You'll notice that a lot of the values in the Age column are missing. While there are many ways to fill these missing values, using the mean or median of the rest of the values is quite common in such cases.

```
Calculate the mean of the Age column and use that value to populate the missing values

Think about other ways you could have populated the missing values in the age column. Why would you pick
```

```
omitted.age <- which(is.na(titanic$age))
titanic$age[omitted.age] = mean(titanic$age, na.rm = TRUE)
```

3: Lifeboat

You're interested in looking at the distribution of passengers in different lifeboats, but as we know, many passengers did not make it to a boat :-(This means that there are a lot of missing values in the boat column. Fill these empty slots with a dummy value e.g. the string 'None' or 'NA'. (*Those dummy values already exist as shown)

```
unavailable.boat = which(is.na(titanic$boat))
```

```
# These dummy values already exist as evidenced by the output
# of the following command.

str(titanic$boat)
```

```
##  Factor w/ 27 levels "1","10","11",..: 12 3 NA NA NA 13 2 NA 27 NA ...
```

4: Cabin

You notice that many passengers don't have a cabin number associated with them.

```
Does it make sense to fill missing cabin numbers with a value?

What does a missing value here mean?
```

You have a hunch that the fact that the cabin number is missing might be a useful indicator of survival. Create a new column has_cabin_number which has 1 if there is a cabin number, and 0 otherwise.

5: Submit the project on Github

Include your code, the original data as a CSV file titanic_original.csv, and the cleaned up data as a CSV file called titanic_clean.csv. SUBMIT YOUR PROJECT.