Springboard Data Wrangling Ex 2: Titanic Survival

Emma Freeman

August 20, 2016

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.csv("titanic3_original.csv", stringsAsFactors = FALSE)
```

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.)

```
any(is.na(titanic$embarked))
```

```
## [1] FALSE
```

```
which(titanic$embarked == "")
```

```
## [1]  169  285 1310
```

```
titanic$embarked[titanic$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.

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

Think about other ways you could have populated the missing values in the age column. Why would you pick any of those over the mean (or not)?

The missing values may not be missing at random and could be predictors, for example it may be more likely that age data are missing for children or the elderly. We could create a random dataset from the original values and randomly sample from the new dataset, or use regression substition to estimate the missing values from the other values. If the data are truly missing at random, using the mean would work.

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'

```
titanic$boat[titanic$boat == ""] <- "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.

Missing cabin numbers might not be random; that is, a missing cabin number may indicate something important related to likelihood of survival. Perhaps it was more likely that cabin number was missing for lower class passengers or children who did not know their cabin number.

Create a new column has_cabin_number which has 1 if there is a cabin number, and 0 otherwise.

```
titanic$has_cabin_number <- ifelse(grepl(".+", titanic$cabin), 1, 0)

write.csv(titanic, 'titanic_clean.csv')
```