Data\_Wrangling\_2

Joe Marco

July 25, 2017

### Step 1:

Load Librarys for packages related to data manipulation (i used supressmessages to hide the packages loading so it was not super messy, this is optional)

suppressMessages(library(tidyverse))  
suppressMessages(library(magrittr))

### Step 2:

* Locate the CSV file (view it if you like to see what it looks like) in the below location and save it.
* <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls>

### Step 3:

* Load CSV file, and give it the name you wish to proceed with. In my case I am calling this data frame “titanic\_original”. You may need to open the excel sheet(xls) and save it as a CSV if needed. I named my file titanic\_original.csv and saved it. (ENSURE THE FILE IS SAVED IN THE SAME DIRECTORY AS R CODE).
* After naming the file, you can take an optional step to view the data to see what you want to manipulate using View() -- this is optional
* I suppressed the messages to the r markdown looks clean and less cluttered

suppressMessages (titanic\_original <- read\_csv("titanic\_original.csv"))  
#View(titanic\_original)

### Step 4:

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.)

titanic\_original <- titanic\_original %>%   
 mutate(embarked = if\_else(is.na(embarked), 'S', embarked))  
any(is.na(titanic\_original$embarked))

## [1] FALSE

#tibble will display to show you what the data frame looks like

### Step 5:

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 any of those over the mean (or not)?

#one way to place mean(average) of age if age column is blank   
  
mean\_age <- mean(titanic\_original$age[!is.na(titanic\_original$age)])  
  
titanic\_original <- titanic\_original %>%   
 mutate(age = if\_else(is.na(age), mean\_age, age))  
  
  
#other ways to polulate could have been to calculate the mean per gender. This way you would get a mean for males, a mean for females, then place the appropriate mean with the appropriate gender. Thus giving you a more distinguished gender seperated view of the data

### Step 6:

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\_original <- titanic\_original %>%   
 mutate(boat = if\_else(is.na(boat), 'NA', boat))

### Step 7:

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.

titanic\_clean <- titanic\_original %>%   
 mutate(has\_cabin\_number = ifelse(cabin == "",0,1))

* write to a new file called titanic\_clean.csv

write.csv(titanic\_clean, file = "titanic\_clean.csv")