**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\_original <- read.csv("C:/Users/icarvalho/Desktop/Datacamp/Titanic/titanic\_original.csv")

View(titanic\_original)

**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.)

gsub(pattern = "^$", replacement = "s", x = titanic\_original$embarked)

**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.

1. Calculate the mean of the Age column and use that value to populate the missing values
2. 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)?

titanic\_original$age[which(is.na(titanic\_original$age))] = mean(titanic\_original$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'

titanic\_original$boat = sub("^$", "NA", titanic\_original$boat)

**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\_numberwhich has 1 if there is a cabin number, and 0 otherwise.

titanic\_original$has\_cabin\_number <- ifelse(titanic\_original$cabin == "", 0, 1)

**6: 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

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