**Data Science Unit 3: Exercise 2 Stephen Golden**

**2016 08 30**

Titanic data set

**Reading in the titanic.csv file**

titanic <- read.csv("L:/MyDocuments/Data Science/Data Wrangling 2/titanic.csv")

**1. Replace missing Southampton with S**

titanic$embarked <- sub("^$", "S", titanic$embarked) [This caused all of the values to be displayed with quotes – Changed class?]

The values were factors. When I inserted the “S”, they became characters. This shows the data type for the “embarked” column

sapply(titanic$embarked, class)

“You can re-factor afterward”

titanic$embarked=as.factor(titanic$embarked)

I did.

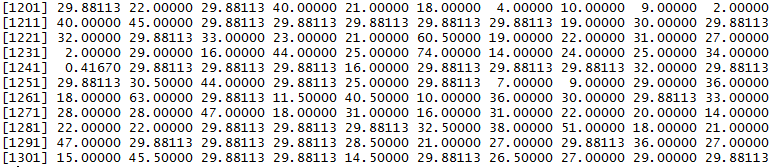
**2. Calculate mean of Age column and place it into missing values**

titanic$age[is.na(titanic$age)] = mean(titanic$age, na.rm=TRUE)

The statement:

mean(titanic$age, na.rm=TRUE)

provides the value 29.88113, so we’ll presume that to be the mean age.



Considering other ways to populate the missing values:

1. Placing zeroes in the missing age values would skew the mean toward the minimum.
2. You could randomly insert age values ranging from the minimum age to the maximum age, but since the passengers were self-selecting, it may bias the results because passengers of a particular age group might be more frequent than a random distribution would produce.
3. We could simply not count the NA entries in the mean of the passenger age and only take the mean of the ages recorded. That seems to me to be as valid as imputing an average age on those missing. It would be interesting to see the mean of those present and the total mean with the average replacing the NA. Wait! That’s what we did! We first computed the mean without the NA entries, then inserted that into the NA places in the data.

The total mean after replacing the NA with the mean of existing values

is identical to the mean of existing values: 29.88113

If issuing

mean(titanic$age, na.rm=TRUE)

before replacing the values is computing the mean of just the values that exist, then issuing mean(titanic$age, na.rm=TRUE) after computing and inserting the mean into the NA values,

titanic$age[is.na(titanic$age)] = mean(titanic$age, na.rm=TRUE),

produces the exact same mean: 29.88113

**3. Place NA in the boat column if there is no boat ID.**

titanic$boat <- sub("^$", "NA", titanic$boat)

Re-factor the values

titanic$boat=as.factor(titanic$boat)

**4. Add a new column to indicate 1 for cabin number and 0 for no cabin number**

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

**5. Write cleaned file to .csv and submit to github**

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