Data wrangling document 2

1. 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.)

----> I was able to find N/A ‘s in the embarked column [ 169, 285, 1310] but I wasn’t able to replace the value to S as specified in the instructions . I was getting a repeated error [Error: object 's' not found] .I tried to replace it with a numeric value it would let me , so I replaced all N/A’s with number “1” . Here is the code I used :

View(titanic\_original)

> heart<- titanic\_original$embarked

> is.na(heart)

which(is.na(heart))

[1] 169 285 1310

> heart[which(is.na(heart))] <- 1

> heart

titanic\_original$embarked<-heart

**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)?

-----> I tried filling the values using mean, median and SD functions. But I would for median function as it is the closest value to age and it rounds off the value to zero value. Here is the code :

Mean

totalage<-titanic\_original$age

> totalage

totalage[is.na(totalage)] // populate NA values in the age column

totalage[!is.na(totalage)] // populate Non NA values in the age column

mean(totalage[!is.na(totalage)])// populate mean from values not including NA

median(totalage[!is.na(totalage)]) // same as above for calculating median

sd(totalage[!is.na(totalage)])// same as above calculating standard deviation

totalage1<-totalage

totalage2<-totalage

totalage1[is.na(totalage1)] <- mean(totalage1[!is.na(totalage1

+ )])

> totalage1 // for finding the missing values using mean func

**Median**

> totalage2[is.na(totalage2)] <- median(totalage2[!is.na(totalage2)])

> totalage2 // finding the missing values using median func

**Standard deviation**

titanic\_original$age<-totalage2

> totalage3<-totalage

> totalage3[is.na(totalage3)] <- sd(totalage3[!is.na(totalage3)])

> totalage3 // finding the missing values using SD func

**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'*

star<-titanic\_original$boat

> star[star==""] <- 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.

wello<-titanic\_original$has\_cabin\_number

> !is.na(wello)

which(!is.na(wello))

wello[which(!is.na(wello))] <- 1

> wello

titanic\_original$has\_cabin\_number<-wello

mello<-titanic\_original$has\_cabin\_number

> is.na(mello)

which(is.na(mello))

mello[which(is.na(mello))] <- 0

mello

titanic\_original$has\_cabin\_number<-mello