Data Wrangling Project 2

Ankita Hanmiah

November 17, 2018

## Load Data into RStudio

library("readr")  
library("dplyr")  
library("tidyr")  
  
titanic\_original <- read.csv("titanic\_original.csv")  
View(titanic\_original)

## Port of Embarkation

# Replacing Empty Strings with NA

titanic\_original$embarked <- na\_if(titanic\_original$embarked, "")

# Converting Embarked NAs with S

embarkedNA <- which(is.na(titanic\_original$embarked) == TRUE)  
titanic\_original$embarked[embarkedNA] <- "S"

## Age

# 1) Calculating Mean of Age Column and Populating Missing Values

avg\_age <- mean(titanic\_original$age, na.rm = TRUE)  
ageNA <- which(is.na(titanic\_original$age) == TRUE)  
titanic\_original$age[ageNA] <- avg\_age

# 2) Other Ways to Populate Missing Values

1. Average females and males separately and assign based on gender of the passengers

* female <- filter(titanic\_original, sex %in% c("female"))  
  female\_avg <- mean(female$age, na.rm = TRUE)  
  female\_NA <- which(is.na(female$age) == TRUE)  
  male <- filter(titanic\_original, sex %in% c("male"))  
  male\_avg <- mane(male$age, na.rm = TRUE)  
  male\_NA <- which(is.na(male$age) == TRUE)  
  titanic\_original$age[female\_NA] <- female\_avg   
  titanic\_oritinal$age[male\_NA] <- male\_avg
* The ages average is around 24, the female average around 28, and the male around 30. There is not a significantly large difference between these numbers, and using an averaged number still does not accurately account for the ranges of ages that would have been seen otherwise.

1. Choose a sample of 264 random numbers between the minimum and maximum ages of passengers and assign those to the missing values

* max(titanic\_original$age)
* maximum age comes out to 80

random\_age <- sample(0:80, 256, replace = TRUE)  
titanic\_original$age[ageNA] <- random\_age

Randomising the values would account for the range of ages that might be seen but whether or not it is more accurate is debatable. Probably the most accurate way to fill the values would be to look into connections between the variables, between possible family members, use the details that have been given, and try to quantify that into a possible age category.

## Lifeboat

# Replace Missing Values with Dummy Value

boatMV <- na\_if(titanic\_original$boat, "")  
boatNA <- which(is.na(titanic\_original$boat) == TRUE)  
titanic\_original$boat[boatNA] <- "NA"

## Cabin

# 1) Does it make sense to fill missing cabin numbers with a value?

No, because the cabin numbers are a specific value that indicates the specific part of the ship the passenger was assigned to stay. There is no real accurate or even general way to postulate values to meaningfully placehold the missing values.

# 2) What does a missing value here mean?

A large number of the values are missing and this could indicate many things. It could indicate class/importance, as many of the cabin numbers for first class are available but scrolling through third class, the majority are missing. It could also possibly mean the person did not survive or was not accounted for, though looking through some examples, it shows that this does not really hold true.

# Creating New Column for has\_cabin\_number

titanic\_original <- titanic\_original$cabin %>% mutate(has\_cabin\_number = (ifelse(!is.na(cabin), yes = 1, no = 0)))

## Save CSV file of Clean Data

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