Project\_3\_2.R

stevehaunguyen

Sat May 19 14:35:14 2018

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
  
#0. Load the data in RStudio --------------------------------------------  
titanic <- read.csv(file = "titanic\_original.csv")  
  
#1. Put "S" for missing values in embarked column -----------------------  
unique(titanic$embarked)

## [1] S C Q  
## Levels: C Q S

titanic$embarked <- ifelse(titanic$embarked == "S", "S",  
 ifelse (titanic$embarked == "C", "C",  
 ifelse(titanic$embarked == "Q", "Q", "S")))  
  
#2. Age -----------------------------------------------------------------  
unique(titanic$age)

## [1] 29.0000 0.9167 2.0000 30.0000 25.0000 48.0000 63.0000 39.0000  
## [9] 53.0000 71.0000 47.0000 18.0000 24.0000 26.0000 80.0000 NA  
## [17] 50.0000 32.0000 36.0000 37.0000 42.0000 19.0000 35.0000 28.0000  
## [25] 45.0000 40.0000 58.0000 22.0000 41.0000 44.0000 59.0000 60.0000  
## [33] 33.0000 17.0000 11.0000 14.0000 49.0000 76.0000 46.0000 27.0000  
## [41] 64.0000 55.0000 70.0000 38.0000 51.0000 31.0000 4.0000 54.0000  
## [49] 23.0000 43.0000 52.0000 16.0000 32.5000 21.0000 15.0000 65.0000  
## [57] 28.5000 45.5000 56.0000 13.0000 61.0000 34.0000 6.0000 57.0000  
## [65] 62.0000 67.0000 1.0000 12.0000 20.0000 0.8333 8.0000 0.6667  
## [73] 7.0000 3.0000 36.5000 18.5000 5.0000 66.0000 9.0000 0.7500  
## [81] 70.5000 22.5000 0.3333 0.1667 40.5000 10.0000 23.5000 34.5000  
## [89] 20.5000 30.5000 55.5000 38.5000 14.5000 24.5000 60.5000 74.0000  
## [97] 0.4167 11.5000 26.5000

mean\_age <- mean(titanic$age, na.rm = TRUE)  
  
titanic$age <- ifelse(is.na(titanic$age), mean\_age, titanic$age)  
  
#Other ways to populate missing values in age column: Could use median, mode  
  
#Would pick using median over the mean, because the age range could be heavily skewed to one side or the other  
# while the median would encompass the range nicely  
  
#Would not pick mode because the mode probably will not be a good representation for statistical purposes  
  
  
#3. Fill empty slots with a dummy value in the lifeboat column ----------  
  
unique(titanic$boat)

## [1] 2 11 3 10 D 4 9   
## [9] 6 B 8 A 5 7 C 14   
## [17] 5 9 13 1 15 5 7 8 10 12 16   
## [25] 13 15 B C D 15 16 13 15   
## 28 Levels: 1 10 11 12 13 13 15 13 15 B 14 15 15 16 16 2 3 4 5 5 7 ... D

titanic$boat <- ifelse(grepl(pattern = ".+", titanic$boat), sapply(titanic$boat, toString), NA)  
  
  
#4. Create a new column "has\_cabin\_number" ------------------------------  
titanic$has\_cabin\_number <- ifelse(grepl(pattern = ".+", titanic$cabin), 1, 0)  
  
  
write.csv(titanic, "titanic\_clean.csv")