Data Wrangling Exercise 2

Shafeeq Ur Rahaman

7/17/2017

library(tidyverse)  
library(magrittr)

#### Requirement over here is to clean the dataset provided in order to make it easier to visualize and analyze.

# Step 0: Load the data in RStudio

#### Loading the CSV file that needs to cleaned

titanic\_original <- read\_csv("titanic\_original.csv", col\_names = TRUE)  
titanic\_original

## # A tibble: 1,309 x 14  
## pclass survived name sex  
## <int> <int> <chr> <chr>  
## 1 1 1 Allen, Miss. Elisabeth Walton female  
## 2 1 1 Allison, Master. Hudson Trevor male  
## 3 1 0 Allison, Miss. Helen Loraine female  
## 4 1 0 Allison, Mr. Hudson Joshua Creighton male  
## 5 1 0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female  
## 6 1 1 Anderson, Mr. Harry male  
## 7 1 1 Andrews, Miss. Kornelia Theodosia female  
## 8 1 0 Andrews, Mr. Thomas Jr male  
## 9 1 1 Appleton, Mrs. Edward Dale (Charlotte Lamson) female  
## 10 1 0 Artagaveytia, Mr. Ramon male  
## # ... with 1,299 more rows, and 10 more variables: age <dbl>, sibsp <int>,  
## # parch <int>, ticket <chr>, fare <dbl>, cabin <chr>, embarked <chr>,  
## # boat <chr>, body <int>, home.dest <chr>

# Step 1: Port of embarkation

#### Replacing the missing values such as blank and empty string

titanic\_original %<>% mutate(embarked = if\_else(is.na(embarked), 'S', embarked))  
titanic\_original %>% select(embarked) %>% unique()

## # A tibble: 3 x 1  
## embarked  
## <chr>  
## 1 S  
## 2 C  
## 3 Q

# Step 2: Age

#### Replacing the missing values such as blank and empty string by calculating the mean of the age

#titanic\_original %<>% mutate(age = if\_else(is.na(age), mean(age, na.rm = TRUE), age))  
titanic\_groupby\_average <- titanic\_original %>%   
 group\_by(pclass, sex) %>%   
 summarise(meanage=mean(age, na.rm=TRUE))  
  
titanic\_original <- titanic\_original %>%   
 left\_join(titanic\_groupby\_average, by=c("pclass","sex")) %>%  
 mutate(age=if\_else(is.na(age), meanage, age)) %>%   
 select(-meanage)  
  
any(is.na(titanic\_original$age))

## [1] FALSE

# Step 3: Lifeboat

#### Replacing the missing values such as blank and empty string by NA. But as read\_csv function was used to load the dataset all missing values were already replaced by 'NA'

any(is.na(titanic\_original$cabin))

## [1] TRUE

any(titanic\_original$cabin=="", na.rm = TRUE)

## [1] FALSE

# Step 4: Cabin

#### Create a tracking column has\_cabin\_number so as can be used as a indicator of survival

titanic\_original$has\_cabin\_number <- if\_else(is.na(titanic\_original$cabin), 0, 1)  
titanic\_original

## # A tibble: 1,309 x 15  
## pclass survived name sex  
## <int> <int> <chr> <chr>  
## 1 1 1 Allen, Miss. Elisabeth Walton female  
## 2 1 1 Allison, Master. Hudson Trevor male  
## 3 1 0 Allison, Miss. Helen Loraine female  
## 4 1 0 Allison, Mr. Hudson Joshua Creighton male  
## 5 1 0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female  
## 6 1 1 Anderson, Mr. Harry male  
## 7 1 1 Andrews, Miss. Kornelia Theodosia female  
## 8 1 0 Andrews, Mr. Thomas Jr male  
## 9 1 1 Appleton, Mrs. Edward Dale (Charlotte Lamson) female  
## 10 1 0 Artagaveytia, Mr. Ramon male  
## # ... with 1,299 more rows, and 11 more variables: age <dbl>, sibsp <int>,  
## # parch <int>, ticket <chr>, fare <dbl>, cabin <chr>, embarked <chr>,  
## # boat <chr>, body <int>, home.dest <chr>, has\_cabin\_number <dbl>

# Step 5: Save as a new dataframe and submit on Github.

#### Creating the cleaned file and submitting it into Github

write\_csv(titanic\_original,"titanic\_clean.csv")