

Springboard Data Wrangling Exercise 2 - Titanic

Chinpei Tang

May 19, 2016

Set working directory

First, save the titanic3.xls file to titanic_original.csv file, and set to the correct working directory.

```
setwd("C:/Users/Chinpei/Documents/GitHub/Springboard_FDS/DW_Ex2")
```

Load original data

Since “Titanic” is one of the preloaded dataset in RStudio, and to avoid overriding the dataset, the dataset is imported as “titanic_ex” (_ex means exercise). Also assign blank and space data to NA.

```
titanic_ex = read.csv("titanic_original.csv", header = T, na.strings = c("", " "))
```

Examine the data.

```
dim(titanic_ex)
```

```
## [1] 1310 14
```

```
summary(titanic_ex)
```

```
##      pclass      survived      name
##  Min.   :1.000   Min.   :0.000   Connolly, Miss. Kate      : 2
## 1st Qu.:2.000   1st Qu.:0.000   Kelly, Mr. James         : 2
##  Median:3.000   Median:0.000   Abbing, Mr. Anthony      : 1
##  Mean   :2.295   Mean   :0.382   Abbott, Master. Eugene Joseph: 1
## 3rd Qu.:3.000   3rd Qu.:1.000   Abbott, Mr. Rossmore Edward : 1
##  Max.   :3.000   Max.   :1.000   (Other)                  :1302
## NA's    :1      NA's    :1      NA's                      : 1
##      sex      age      sibsp      parch
## female:466   Min.   : 0.1667   Min.   :0.0000   Min.   :0.000
## male :843    1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000
## NA's : 1     Median :28.0000   Median :0.0000   Median :0.000
##                               Mean   :29.8811   Mean   :0.4989   Mean   :0.385
##                               3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##                               Max.   :80.0000   Max.   :8.0000   Max.   :9.000
##                               NA's    :264      NA's    :1      NA's    :1
##      ticket      fare      cabin      embarked
## CA. 2343: 11   Min.   : 0.000   C23 C25 C27      : 6   C :270
## 1601 : 8      1st Qu.: 7.896   B57 B59 B63 B66: 5   Q :123
## CA 2144 : 8      Median :14.454   G6                : 5   S :914
## 3101295 : 7      Mean   :33.295   B96 B98           : 4   NA's: 3
## 347077 : 7      3rd Qu.:31.275   C22 C26           : 4
## (Other) :1268   Max.   :512.329   (Other)           :271
```

```
## NA's      : 1  NA's      :2      NA's      :1015
##      boat      body      home.dest
## 13      : 39  Min.      : 1.0  New York, NY      : 64
## C      : 38  1st Qu.: 72.0  London          : 14
## 15      : 37  Median :155.0  Montreal, PQ     : 10
## 14      : 33  Mean   :160.8  Cornwall / Akron, OH: 9
## 4       : 31  3rd Qu.:256.0  Paris, France    : 9
## (Other):308  Max.   :328.0  (Other)          :639
## NA's     :824  NA's     :1189  NA's             :565
```

There are 1310 observations, and 14 columns.

Port of embarkation

Examine the NA's in embarked column.

```
summary(titanic_ex$embarked)
```

```
##      C      Q      S NA's
## 270 123 914      3
```

In fact, there are actually 3 missing values instead of 1. Substitute the missing port of embarkation to “S”.

```
titanic_ex$embarked[is.na(titanic_ex$embarked)] = "S"
summary(titanic_ex$embarked)
```

```
##      C      Q      S
## 270 123 917
```

Now there is no more NA.

Age

Examine the NA's in age column.

```
summary(titanic_ex$age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
## 0.1667 21.0000 28.0000 29.8800 39.0000 80.0000    264
```

There are 264 NA entries. Calculate the mean of the age ignoring the NA's.

```
mean(titanic_ex$age, na.rm = T)
```

```
## [1] 29.88113
```

Substitute the NA's with the mean values.

```
titanic_ex_agemean = titanic_ex
titanic_ex_agemean$age[is.na(titanic_ex$age)] = mean(titanic_ex$age, na.rm = T)
summary(titanic_ex_agemean$age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1667 22.0000 29.8800 29.8800 35.0000 80.0000
```

Some other ways to populate the missing values are taking the median value.

```
titanic_ex_agemed = titanic_ex
titanic_ex_agemed$age[is.na(titanic_ex$age)] = median(titanic_ex$age, na.rm = T)
summary(titanic_ex_agemed$age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1667 22.0000 28.0000 29.5000 35.0000 80.0000
```

However, there is not much statistical difference between taking median or mean values. So use the mean age to substitute the average data.

```
titanic_ex = titanic_ex_agemean
```

Lifeboat

NA has already been assigned to the blank or space data in the boat column.

```
summary(titanic_ex$boat)
```

```
##      1      10      11      12      13      13 15 13 15 B      14      15
##      5      29      25      19      39      2      1      33      37
##    15 16      16      2      3      4      5      5 7      5 9      6
##      1      23      13      26      31      27      2      1      20
##      7      8      8 10      9      A      B      C      C D      D
##     23      23      1      25      11      9      38      2      20
##    NA's
##     824
```

```
titanic_ex$boat[is.na(titanic_ex$boat)]
```

```
##      [1] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [15] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [29] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [43] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [57] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [71] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [85] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##     [99] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##    [113] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##    [127] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
##    [141] <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
```

```
summary(titanic_ex$boat)
```

4

```
##      15 16      16      2      3      4      5      5 7      5 9      6
##          1      23      13      26      31      27      2      1      20
##          7      8      8 10      9      A      B      C      C D      D
##          23      23      1      25      11      9      38      2      20
##      NA's
##      824
```

Cabin

Finally, create the “has_cabin_number” column for the passenger with cabin numbers.

```
summary(titanic_ex$cabin)
```

```
##      C23 C25 C27 B57 B59 B63 B66      G6      B96 B98
##          6      5
##      C22 C26      C78      D      F2
##          4      4      4
##          F33      F4      A34      B51 B53 B55
##          4      4      3
##      B58 B60      C101      E101      E34
##          3      3      3
##          B18      B20      B22      B28
##          2      2      2
##          B35      B41      B45      B49
##          2      2      2
##          B5      B69      B71      B77
##          2      2      2
##          B78      C106      C116      C123
##          2      2      2
##      C124      C125      C126      C2
##          2      2      2
##          C31      C32      C46      C52
##          2      2      2
##          C54      C55 C57      C6      C62 C64
##          2      2      2
##          C65      C68      C7      C80
##          2      2      2
##          C83      C85      C86      C89
##          2      2      2
##          C92      C93      D10 D12      D15
##          2      2      2
##          D17      D19      D20      D21
##          2      2      2
##          D26      D28      D30      D33
##          2      2      2
##          D35      D36      D37      E121
##          2      2      2
##          E24      E25      E31      E33
##          2      2      2
##          E44      E46      E50      E67
##          2      2      2
##          E8      F G63      F G73      A10
##          2      2      2      1
```

##	A11	A14	A16	A18
##	1	1	1	1
##	A19	A20	A21	A23
##	1	1	1	1
##	A24	A26	A29	A31
##	1	1	1	1
##	A32	A36	A5	A6
##	1	1	1	1
##	A7	A9	(Other)	NA's
##	1	1	88	1015

```
titanic_ex$has_cabin_number = as.integer(!is.na(titanic_ex$cabin))
```

Write to clean file

Write the new dataset to the clean csv file.

```
write.csv(titanic_ex, file = "titanic_clean.csv")
```