Data Wrangling Exercise 2: Dealing With Missing Values

Springboard: Foundations of Data Science

Antoine Beauchamp

January 16th, 2017

Introduction

In this exercise, we handle some of the missing values in the Titanic data set. Let's begin by clearing the working environment, loading libraries, and setting the working directory.

```
#Clear working environment
rm(list=ls())

#Import library
library(readxl)
library(readr)
suppressMessages(library(dplyr))
library(tidyr)

#Set correct working directory
path_to_wd <- file.path("~","Documents","Work","DataScience","Springboard","FoundationsofDataScience","
setwd(path_to_wd)
rm(path_to_wd)
```

Section 0: Import Data

Let's start by converting the data from .xls to .csv, and import the data into a tibble format.

```
#Convert data from Excel format to CSV
read_excel("titanic3.xls") %>% write_csv("titanic_original.csv")
```

```
## Warning in xls_cols(path, sheet, col_names = col_names, col_types =
## col_types, : Expecting numeric in [1305, 13] got `328`
```

```
#Import data
titanicdata <- suppressMessages(read_csv("titanic_original.csv"))
```

Let's run some basic summary functions to get an idea of the data:

```
class(titanicdata)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
str(titanicdata)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1310 obs. of  14 variables:
##  $ pclass   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : num  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : chr  "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. 
##  $ sex      : chr  "female" "male" "female" "male" ...
##  $ age      : num  29 0.917 2 30 25 ...
##  $ sibsp    : num  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : num  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : chr  "24160.000000" "113781.000000" "113781.000000" "113781.000000" ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : chr  "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ embarked : chr  "S" "S" "S" "S" ...
##  $ boat     : chr  "2.000000" "11" NA NA ...
##  $ body     : num  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: chr  "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 14
##   .. ..$ pclass   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ survived : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ name     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ sex      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ age      : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ sibsp    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ parch    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ ticket   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ fare     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ cabin    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ embarked : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ boat     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ body     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ home.dest: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```r
glimpse(titanicdata)
```

```
## Observations: 1,310
## Variables: 14
## $ pclass    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ survived  <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1...
## $ name      <chr> "Allen, Miss. Elisabeth Walton", "Allison, Master. H...
## $ sex       <chr> "female", "male", "female", "male", "female", "male"...
## $ age       <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, ...
## $ sibsp     <dbl> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0...
## $ parch     <dbl> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1...
## $ ticket    <chr> "24160.000000", "113781.000000", "113781.000000", "1...
## $ fare      <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151.5500, 26...
## $ cabin     <chr> "B5", "C22 C26", "C22 C26", "C22 C26", "C22 C26", "E...
## $ embarked  <chr> "S", "S", "S", "S", "S", "S", "S", "S", "S", "C", "C...
## $ boat      <chr> "2.000000", "11", NA, NA, NA, "3", "10", NA, "D", NA...
## $ body      <dbl> NA, NA, NA, 135, NA, NA, NA, NA, NA, 22, 124, NA, NA...
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "...
```

```
titanicdata
```

```
## # A tibble: 1,310 × 14
##    pclass survived                                            name    sex
##     <dbl>    <dbl>                                           <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 1,300 more rows, and 10 more variables: age <dbl>, sibsp <dbl>,
## #   parch <dbl>, ticket <chr>, fare <dbl>, cabin <chr>, embarked <chr>,
## #   boat <chr>, body <dbl>, home.dest <chr>
```

Let's wrangle some data!

Section 1: Port of Embarkation

First we begin by replacing the missing values for the `port of embarkation` variable with a value of "S" for Southampton.

```
#Preserve original data
titanic_d1 = titanicdata

#Find NA values and replace with "S"
titanic_d1$embarked[titanicdata$embarked %>% is.na()] = "S"

#Let's just make sure that there aren't any missing values left.
titanic_d1 %>% filter(is.na(embarked))
```

```
## # A tibble: 0 × 14
## # ... with 14 variables: pclass <dbl>, survived <dbl>, name <chr>,
## #   sex <chr>, age <dbl>, sibsp <dbl>, parch <dbl>, ticket <chr>,
## #   fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <dbl>,
## #   home.dest <chr>
```

```
titanic_d1 %>% filter(embarked == '')
```

```
## # A tibble: 0 × 14
## # ... with 14 variables: pclass <dbl>, survived <dbl>, name <chr>,
## #   sex <chr>, age <dbl>, sibsp <dbl>, parch <dbl>, ticket <chr>,
## #   fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <dbl>,
## #   home.dest <chr>
```

Section 2: Age

To deal with missing `age` values, we will can use the **mean** or **median** of the rest of the values to estimate the data. Here I used the mean.

```
titanic_d2 = titanic_d1

#Calculate the mean of the age variable
age_mean <- titanic_d1 %>% summarise(mean(age, na.rm=TRUE))
#Calculate the median of the age variable (why not?)
age_median <- titanic_d1 %>% summarise(median(age,na.rm=TRUE))

age_mean[[1]]
```

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

```
#Replace the missing age values with the mean of the variable.
titanic_d2$age[titanic_d1$age %>% is.na()] = round(age_mean[[1]])

#Making sure we haven't missed anything.
titanic_d2 %>% filter(is.na(age))
```

```
## # A tibble: 0 × 14
## # ... with 14 variables: pclass <dbl>, survived <dbl>, name <chr>,
## #   sex <chr>, age <dbl>, sibsp <dbl>, parch <dbl>, ticket <chr>,
## #   fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <dbl>,
## #   home.dest <chr>
```

```
titanic_d2 %>% filter(age=='')
```

```
## # A tibble: 0 × 14
## # ... with 14 variables: pclass <dbl>, survived <dbl>, name <chr>,
## #   sex <chr>, age <dbl>, sibsp <dbl>, parch <dbl>, ticket <chr>,
## #   fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <dbl>,
## #   home.dest <chr>
```

```
length(subset(titanic_d2$age, titanic_d2$age %>% is.na()))
```

```
## [1] 0
```

In addition to using the mean or the median of the age variable to estimate the missing values, we also could have used the **mode** of the variable. This would be a reasonable estimate, as the mode represents the most common value of the variable. This means that, assuming age to be a random variable, the mode describes the value with the highest probability of occurring. A passenger is then more likely to have this value as their age, making it a good estimate of missing values.

Section 3: Lifeboat

Let's find the missing values for the `boat` variable and replace them with "None"

```
titanic_d3 = titanic_d2

#Replace missing values with "None"
titanic_d3$boat[titanic_d2$boat %>% is.na()] = "None"

#Make sure we haven't missed anything
titanic_d3 %>% filter(is.na(boat))
```

```
## # A tibble: 0 × 14
## # ... with 14 variables: pclass <dbl>, survived <dbl>, name <chr>,
## #   sex <chr>, age <dbl>, sibsp <dbl>, parch <dbl>, ticket <chr>,
## #   fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <dbl>,
## #   home.dest <chr>
```

```
titanic_d3 %>% filter(boat=='')
```

```
## # A tibble: 0 × 14
## # ... with 14 variables: pclass <dbl>, survived <dbl>, name <chr>,
## #   sex <chr>, age <dbl>, sibsp <dbl>, parch <dbl>, ticket <chr>,
## #   fare <dbl>, cabin <chr>, embarked <chr>, boat <chr>, body <dbl>,
## #   home.dest <chr>
```

```
length(subset(titanic_d3$boat, titanic_d3$boat %>% is.na()))
```

```
## [1] 0
```

Section 4: Cabin

Presumably, the passengers that don't have a cabin number associated with them were not staying in a cabin. More likely they were staying in some common bunker area of sorts. We could replace the missing cabin values with something like "None" or "Bunker", whatever is most appropriate. This might not be telling the whole story however, since some passengers in class 1, who have the means to afford a cabin, also appear not to have cabins. It might not be reasonable to assume that all missing values for this variable correspond to passengers without cabins.

In any case, we will create a new variable, `has_cabin_number`, that represents whether the passenger had a cabin or not.

```
#Create a new variable describing the existence of a cabin number
titanic_clean <- titanic_d3 %>% mutate(has_cabin_number = sapply(titanic_d3$cabin, function (x) {if (is

titanic_clean %>% select(cabin, has_cabin_number)
```

```
## # A tibble: 1,310 × 2
##       cabin has_cabin_number
##       <chr>            <dbl>
## 1        B5                1
## 2   C22 C26                1
## 3   C22 C26                1
## 4   C22 C26                1
## 5   C22 C26                1
## 6       E12                1
## 7        D7                1
## 8       A36                1
## 9      C101                1
## 10     <NA>                0
## # ... with 1,300 more rows
```

Section 5: Write to File

Finally, let's write the data to a .csv file.

```r
write_csv(titanic_clean, "titanic_clean.csv")
```