Data Wrangling - Missing Values

Tom Thorpe

March 27, 2018

Springboard Data Science Foundation Class

Exercise 01 - Data Wrangling - Refine Data

Objective

Practice using *tidy* and *dplyr* packages to clean up data in a practice dataset.

Exercise Results

Load the tidy and dplyr libraries for exercise.

```
#library(devtools)
library(tidyr)
library(dplyr)
# install.packages("gdata") needed for read.xls function
# library(gdata) - instructions were to convert to csv before loading, so not needed.
```

Identify input and output files

The practice data set was downloaded into a file called "titanic3.xls", resaved as "titanic_original.csv" and the cleaned data will be stored in a file named, "titanic_clean.csv".

```
infile = "C:/Users/Tom/git/datasciencefoundation/DataWrangleExer02/titanic_original.csv"
outfile = "C:/Users/Tom/git/datasciencefoundation/DataWrangleExer02/titanic_clean.csv"
```

Read CSV file into a local dataframe

```
# titanic <- read.xls(infile) %>% tbl_df # instructions were to convert to csv before loading
#titanic <- read.csv(file=infile, header=TRUE, sep=",") %>% tbl_df()
titanic1 <- read.csv(file=infile, header=TRUE, sep=",")
titanic <- tbl_df(titanic1)
```

Lets see what the data looks like:

```
str(titanic)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1309 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 22 24 25 26 27 31 46 47 51 55 ...
##  $ sex      : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age      : num  29 0.917 2 30 25 ...
##  $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
```

```
##  $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : Factor w/ 929 levels "110152","110413",..: 188 50 50 50 50 125 93 16 77 826 ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...
##  $ embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...
##  $ boat     : Factor w/ 28 levels "","1","10","11",..: 13 4 1 1 1 14 3 1 28 1 ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...
```

```
titanic
```

```
## # A tibble: 1,309 x 14
##    pclass survived name        sex      age sibsp parch ticket  fare cabin
##     <int>    <int> <fct>       <fct>  <dbl> <int> <int> <fct>  <dbl> <fct>
##  1      1        1 Allen, Mis~ fema~ 29.0       0     0 24160  211.  B5
##  2      1        1 Allison, M~ male   0.917     1     2 113781 152.  C22 ~
##  3      1        0 Allison, M~ fema~  2.00      1     2 113781 152.  C22 ~
##  4      1        0 Allison, M~ male  30.0       1     2 113781 152.  C22 ~
##  5      1        0 Allison, M~ fema~ 25.0       1     2 113781 152.  C22 ~
##  6      1        1 Anderson, ~ male  48.0       0     0 19952   26.6 E12
##  7      1        1 Andrews, M~ fema~ 63.0       1     0 13502   78.0 D7
##  8      1        0 Andrews, M~ male  39.0       0     0 112050   0.  A36
##  9      1        1 Appleton, ~ fema~ 53.0       2     0 11769   51.5 C101
## 10      1        0 Artagaveyt~ male  71.0       0     0 PC 17~  49.5 ""
## # ... with 1,299 more rows, and 4 more variables: embarked <fct>,
## #   boat <fct>, body <int>, home.dest <fct>
```

I noticed that there are three additional variables in the data set than in the data description pointed to by the exercise URL, *boat*, *body* and *home.dest*. I found the full description of the titanic3.xls data set here http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3info.txt. The *boat* is the lifeboat, *body* is the body identification number, and *home.dest* is the home destination of the passenger.

Default Missing *embarked* data

First fill in any missing data in the *embarked* column with 'S' to represent Southampton. First, Check to see how many are blank.

```
table(titanic$embarked)
```

```
##
##       C   Q   S
##   2 270 123 914
```

There are two empty *embarked* values. Make the updates.

```
titanic$embarked[titanic$embarked==""|titanic$embarked==" "] <- "S"
#titanic$embarked[grepl("^ |! {0}",titanic$embarked)] <- "S" # couldn't figure out a regular expression
```

Check the results of the change.

```
table(titanic$embarked)
```

```
##
##       C   Q   S
##   0 270 123 916
```

Populate missing *age* data

Populate the missing data with the mean of the other rows with age data. Look at the age data before populating.

```
table(titanic$age)
```

```
##
## 0.1667 0.3333 0.4167 0.6667   0.75 0.8333 0.9167      1      2      3
##      1      1      1      1      3      3      2     10     12      7
##      4      5      6      7      8      9     10     11   11.5     12
##     10      5      6      4      6     10      4      4      1      3
##     13     14   14.5     15     16     17     18   18.5     19     20
##      5      8      2      6     19     20     39      3     29     23
##   20.5     21     22   22.5     23   23.5     24   24.5     25     26
##      1     41     43      1     26      1     47      1     34     30
##   26.5     27     28   28.5     29     30   30.5     31     32   32.5
##      1     30     32      3     30     40      2     23     24      4
##     33     34   34.5     35     36   36.5     37     38   38.5     39
##     21     16      2     23     31      2      9     14      1     20
##     40   40.5     41     42     43     44     45   45.5     46     47
##     18      3     11     18      9     10     21      2      6     14
##     48     49     50     51     52     53     54     55   55.5     56
##     14      9     15      8      6      4     10      8      1      4
##     57     58     59     60   60.5     61     62     63     64     65
##      5      6      3      7      1      5      5      4      5      3
##     66     67     70   70.5     71     74     76     80
##      1      1      2      1      2      1      1      1
```

```
summary(titanic$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1667 21.0000 28.0000 29.8811 39.0000 80.0000     263
```

```
filter(titanic,is.na(age)) %>% count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   263
```

```
select(titanic,name,age) %>% print(n=18)
```

```
## # A tibble: 1,309 x 2
##    name                                              age
##    <fct>                                           <dbl>
##  1 Allen, Miss. Elisabeth Walton                    29.0
##  2 Allison, Master. Hudson Trevor                    0.917
##  3 Allison, Miss. Helen Loraine                      2.00
##  4 Allison, Mr. Hudson Joshua Creighton             30.0
##  5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)  25.0
##  6 Anderson, Mr. Harry                              48.0
##  7 Andrews, Miss. Kornelia Theodosia                63.0
##  8 Andrews, Mr. Thomas Jr                           39.0
##  9 Appleton, Mrs. Edward Dale (Charlotte Lamson)    53.0
## 10 Artagaveytia, Mr. Ramon                          71.0
## 11 Astor, Col. John Jacob                           47.0
```

```
## 12 Astor, Mrs. John Jacob (Madeleine Talmadge Force) 18.0
## 13 Aubart, Mme. Leontine Pauline                      24.0
## 14 "Barber, Miss. Ellen \"Nellie\""                   26.0
## 15 Barkworth, Mr. Algernon Henry Wilson               80.0
## 16 Baumann, Mr. John D                                NA
## 17 Baxter, Mr. Quigg Edmond                           24.0
## 18 Baxter, Mrs. James (Helene DeLaudeniere Chaput)    50.0
## # ... with 1,291 more rows
```

Notice that the 16th entry has NA for age. Set the empty age data to the mean of the non-empty age.

```
meanAge= mean(titanic$age,na.rm=TRUE)
medianAge= median(titanic$age,na.rm=TRUE)
# meanAge= mean(titanic$age[titanic$age>=0],na.rm=TRUE)
meanAge
```

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

```
medianAge
```

```
## [1] 28
```

```
meanAge<-trunc(meanAge) # truncate the age since only age less than 1 have fractional values.
#titanic$age[titanic$age==NA] <- meanAge
titanic$age[is.na(titanic$age)] <- meanAge

#Check the results
select(titanic,name,age) %>% print(n=18)
```

```
## # A tibble: 1,309 x 2
##    name                                                age
##    <fct>                                              <dbl>
##  1 Allen, Miss. Elisabeth Walton                     29.0
##  2 Allison, Master. Hudson Trevor                     0.917
##  3 Allison, Miss. Helen Loraine                       2.00
##  4 Allison, Mr. Hudson Joshua Creighton              30.0
##  5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)   25.0
##  6 Anderson, Mr. Harry                               48.0
##  7 Andrews, Miss. Kornelia Theodosia                 63.0
##  8 Andrews, Mr. Thomas Jr                            39.0
##  9 Appleton, Mrs. Edward Dale (Charlotte Lamson)     53.0
## 10 Artagaveytia, Mr. Ramon                           71.0
## 11 Astor, Col. John Jacob                            47.0
## 12 Astor, Mrs. John Jacob (Madeleine Talmadge Force) 18.0
## 13 Aubart, Mme. Leontine Pauline                     24.0
## 14 "Barber, Miss. Ellen \"Nellie\""                  26.0
## 15 Barkworth, Mr. Algernon Henry Wilson              80.0
## 16 Baumann, Mr. John D                               29.0
## 17 Baxter, Mr. Quigg Edmond                          24.0
## 18 Baxter, Mrs. James (Helene DeLaudeniere Chaput)   50.0
## # ... with 1,291 more rows
```

```
filter(titanic,is.na(age)) %>% count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
```

```
## 1     0
```

How else could the missing age data be populated? The median is an alternative measure that may be closer to a representative age. The median in this case is 28. The mean is 29.88 rounded to 30. I believe that since the median is less than the mean, there are more younger people than old people on the ship. I would tend to use the median over the mean in this case. Somehow it just feels better. Maybe I will have a better answer after taking the statistics part of the course.

Populate missing Lifeboat data

There is missing data for the lifeboat column.

Does it make sense to fill in the cabin numbers with a value? To me, it does not make sense to try to come up with an estimate of the cabin number. How could any meaningful data be created? It makes sense to set the missing data to "NONE".

What does a missing value mean here? Missing data could mean the records for the cabin assignment were lost or the people were assigned to a big dormitory area instead of a room. But there was probably more than one dormitory and the specific dormitory is unknown. Either way, the best that can be done is to do as the assignment suggests and assign missing data the value of "NONE".

Update the missing data in the lifeboat data with "NONE". I tried using these commands

titanic$boat[titanic$boat==""] <- as.factor("NONE")

titanic$boat[titanic$boat==""] <- "NONE"

But both resulted in the following warning:

Warning in `[<-.factor(*tmp*`, titanic$boat == "", value = structure(c(13L, : invalid factor level, NA generated

So I did some research and converted the *boat* column from a factor to character data. But I don't understand why the column was a factor and not character data to begin with. I need to research this.

```
filter(titanic, boat =="") %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   823
```

```
titanic$boat <- as.character.factor (titanic$boat)
titanic$boat[titanic$boat==""] <- "NONE"
#Check the results
select(titanic,name,boat) %>% print(n=15)
```

```
## # A tibble: 1,309 x 2
##    name                                            boat
##    <fct>                                           <chr>
##  1 Allen, Miss. Elisabeth Walton                   2
##  2 Allison, Master. Hudson Trevor                  11
##  3 Allison, Miss. Helen Loraine                    NONE
##  4 Allison, Mr. Hudson Joshua Creighton            NONE
##  5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) NONE
##  6 Anderson, Mr. Harry                             3
##  7 Andrews, Miss. Kornelia Theodosia               10
##  8 Andrews, Mr. Thomas Jr                          NONE
##  9 Appleton, Mrs. Edward Dale (Charlotte Lamson)   D
```

```
## 10 Artagaveytia, Mr. Ramon                              NONE
## 11 Astor, Col. John Jacob                               NONE
## 12 Astor, Mrs. John Jacob (Madeleine Talmadge Force) 4
## 13 Aubart, Mme. Leontine Pauline                        9
## 14 "Barber, Miss. Ellen \"Nellie\""                     6
## 15 Barkworth, Mr. Algernon Henry Wilson                 B
## # ... with 1,294 more rows
```

```
filter(titanic, boat =="") %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

Create has_cabin_number column

Create a *has_cabin_number* column with a value of 1 if the *cabin* column is non-blank.

```
titanic <- mutate(titanic,has_cabin_number=ifelse(cabin!="",1,0))
# Check results
select(titanic,name,cabin,has_cabin_number) %>% print(n=15)
```

```
## # A tibble: 1,309 x 3
##    name                                      cabin  has_cabin_number
##    <fct>                                     <fct>             <dbl>
##  1 Allen, Miss. Elisabeth Walton             B5                    1.
##  2 Allison, Master. Hudson Trevor            C22 C~                1.
##  3 Allison, Miss. Helen Loraine              C22 C~                1.
##  4 Allison, Mr. Hudson Joshua Creighton      C22 C~                1.
##  5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) C22 C~          1.
##  6 Anderson, Mr. Harry                       E12                   1.
##  7 Andrews, Miss. Kornelia Theodosia         D7                    1.
##  8 Andrews, Mr. Thomas Jr                    A36                   1.
##  9 Appleton, Mrs. Edward Dale (Charlotte Lamson)   C101            1.
## 10 Artagaveytia, Mr. Ramon                   ""                    0.
## 11 Astor, Col. John Jacob                    C62 C~                1.
## 12 Astor, Mrs. John Jacob (Madeleine Talmadge For~ C62 C~          1.
## 13 Aubart, Mme. Leontine Pauline             B35                   1.
## 14 "Barber, Miss. Ellen \"Nellie\""          ""                    0.
## 15 Barkworth, Mr. Algernon Henry Wilson      A23                   1.
## # ... with 1,294 more rows
```

```
glimpse(titanic)
```

```
## Observations: 1,309
## Variables: 15
## $ pclass          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ survived        <int> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, ...
## $ name            <fct> Allen, Miss. Elisabeth Walton, Allison, Maste...
## $ sex             <fct> female, male, female, male, female, male, fem...
## $ age             <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48...
## $ sibsp           <int> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, ...
## $ parch           <int> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ ticket          <fct> 24160, 113781, 113781, 113781, 113781, 19952,...
## $ fare            <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151.5...
```

```
## $ cabin            <fct> B5, C22 C26, C22 C26, C22 C26, C22 C26, E12, ...
## $ embarked         <fct> S, S, S, S, S, S, S, S, S, C, C, C, C, S, S, ...
## $ boat             <chr> "2", "11", "NONE", "NONE", "NONE", "3", "10",...
## $ body             <int> NA, NA, NA, 135, NA, NA, NA, NA, NA, 22, 124,...
## $ home.dest        <fct> St Louis, MO, Montreal, PQ / Chesterville, ON...
## $ has_cabin_number <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, ...
```

Save the cleaned data

```
write.csv(titanic, file=outfile,row.names=FALSE)
```

That concludes the exercise.