Data Wrangling 2

Georgie Knight

8 August 2016

The titanic data set.

Dplyr and tidyr

Load the *dplyr* and *tidyr* packages which will help us wrangle the data:

```
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")
```

Loading up the data

Load the file 'refine_original.csv' into R:

```
refine_original <- read.csv("titanic_original.csv")
dt_fr <- data.frame(refine_original)
```

Convert it to a table called 'titanic_table' within the dplyr package…

```
titanic_table<-dplyr::tbl_df(dt_fr)
```

Take a quick look at it:

```
## # A tibble: 1,309 x 14
##    pclass survived                                            name    sex
##     <int>    <int>                                          <fctr> <fctr>
## 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 <fctr>, fare <dbl>, cabin <fctr>, embarked <fctr>,
## #   boat <fctr>, body <int>, home.dest <fctr>
```

```
##  [1] "pclass"    "survived"  "name"      "sex"       "age"
##  [6] "sibsp"     "parch"     "ticket"    "fare"      "cabin"
## [11] "embarked"  "boat"      "body"      "home.dest"
```

```
## Observations: 1,309
## Variables: 14
## $ pclass    <int> 1, 1, 1, 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, 0, 0, 1...
## $ name      <fctr> Allen, Miss. Elisabeth Walton, Allison, Master. Hud...
## $ sex       <fctr> female, male, female, male, female, male, female, m...
## $ age       <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, ...
## $ sibsp     <int> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0...
## $ parch     <int> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1...
## $ ticket    <fctr> 24160, 113781, 113781, 113781, 113781, 19952, 13502...
## $ fare      <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151.5500, 26...
## $ cabin     <fctr> B5, C22 C26, C22 C26, C22 C26, C22 C26, E12, D7, A3...
## $ embarked  <fctr> S, S, S, S, S, S, S, S, S, C, C, C, C, S, S, S, C, ...
## $ boat      <fctr> 2, 11, , , , 3, 10, , D, , , 4, 9, 6, B, , , 6, 8, ...
## $ body      <int> NA, NA, NA, 135, NA, NA, NA, NA, NA, 22, 124, NA, NA...
## $ home.dest <fctr> St Louis, MO, Montreal, PQ / Chesterville, ON, Mont...
```

Task 1:Replace missing values

The embarked column has some missing values, which are known to correspond to passengers who actually embarked at Southampton. Find the missing values and replace them with S

```
titanic_table$embarked
```

```
##    [1] S S S S S S S S S C C C C S S S C C C C S S C C S C C C S S S C S S
##   [35] S C S S S C C C S C C S S S C C C S S S S S S S S C S S S S S S C S
##   [69] C S S C C S S S C C C S S S S S C C S S S S S S S S S S C C C C C C
##  [103] C C C C S S C S C S S S S S S C S S C S C C C S S S S C C C C C C C
##  [137] C S S S C C C C C C C S S S S C S C S S S S S C C S S S C S C S   C
##  [171] S S S C C S S S S S S S C C C S C S S S S S C S S S C S S C C S S C
##  [205] C S Q Q Q C S S C C C C S S C C C C C S S S S S C C S S S C S S C C
##  [239] S S S S C C C S C C S C C C C C S C S C C S S S S S C S S C S S S S
##  [273] S C C C C C C S C C C C   S S S S S S S S S C C C C C C S S C C C S
##  [307] S S C S S S C C C S C C S C S S C C C S S S S S S S S S S S S S S S
##  [341] S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S
##  [375] S S S S S S S S S S S S S S S S S C C S S S S S S S S C C S S S S S
##  [409] S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S
##  [443] S S S S S S S S S S S S S S S S S S S S S S S C S S S Q Q S Q S S S
##  [477] S Q C C C C C S S C S Q S S S C C C C C S S S S S S S S S S S S S S
##  [511] S Q C C S S S S S S C S S S C S C S S S S C S S S S S S C C S S S S
##  [545] S S S C S S S S S S S S S S S S S S Q S S S C S S S S S S S S S S S
##  [579] S S S S S S S S S S S S S S S S S S S S S S S S S S S S S C S S S S
##  [613] S C S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S S
##  [647] S C C S C C S C C S S C C C C S C S C C C Q S S S S C C S S S S C C
##  [681] C C Q Q Q S Q S S S S S Q Q Q Q S S S S S S Q Q S C C S S Q Q S S Q
##  [715] S S C C S S Q S S Q Q Q Q Q S S S S S S S S S S S S S C S S S Q Q S
##  [749] S S S S S S S S S S S S S S S S S S S S S S S Q S S S C Q S S Q S C
##  [783] Q S S S S S S C C C C S S C S Q S S Q Q Q Q Q S S S S S S S Q S Q S
##  [817] C S Q Q S S S S S S S S S S S S S S S S S S S S S Q S S S S S C S S
##  [851] S S S S Q C Q S S Q S S S S S Q S S S S S Q S S S C S S S S C S S S
##  [885] S S Q S S S S S S S S S S S S S S S S S S S S S S S S S S S S S C C
##  [919] C C Q S Q Q Q S Q C C Q Q Q S S S S S S S S C C C S S S S S Q S S S
##  [953] S C S S S S S S C Q Q S S S S S S S S Q S S S S S S S S S S S S Q S
##  [987] S Q Q S S C Q Q C C S C C Q Q Q Q Q Q Q Q Q Q Q S S Q S S S Q S S S
## [1021] S S S Q S S S S Q Q Q S Q S C C C C S Q Q S Q Q Q S S C C C C S S C
## [1055] Q S C C S S S S S S S S C S S Q Q Q Q Q Q S Q Q Q S Q Q S S S S S S
## [1089] S S S S S S S Q S S S S S S S S S S S S S C S S S S S S S S S S S C
## [1123] C C Q S S S S S S S S S S S S C S S S Q Q Q Q Q Q S S Q S S S S S S
## [1157] S S S S S S Q Q C C C Q S S S S S S S S S S S S S S S S C C C S S S
## [1191] S S S Q S Q S S Q S S C S S S S S S S S S S S S S Q Q S S S S S C S
## [1225] S S S S S S S S S S S S S S S S C C C C C S S S S Q S S S C S C C C
## [1259] C S S S S S S S S S S S S S S S S S S C S S S S S S S S C S S S S S
## [1293] S S S S S S S C C C C C C C C C S
## Levels:  C Q S
```

```
levels(titanic_table$embarked)
```

```
## [1] ""  "C" "Q" "S"
```

```
summary(titanic_table$embarked)
```

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

We see this is a factor data type it should an empty factor along with C, Q or S. We'll have to change the empty factor to S. We'll do this first by duplicating it, This forces R to put all labels into the lesser value, then relabeling it which will remove the empty factor.

```
titanic_table <- titanic_table %>%
  mutate(embarked = factor(embarked, labels=c('S','C','Q','S'))) %>%
  mutate(embarked = factor(embarked, labels=c('S','C','Q')))
```

```
## Warning in `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels)
## else paste0(labels, : duplicated levels in factors are deprecated

## Warning in `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels)
## else paste0(labels, : duplicated levels in factors are deprecated
```

```
summary(titanic_table$embarked)
```

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

Task 2: Repopulate the age column with the mean age

You'll notice that a lot of the values in the Age column are missing. While there are many ways to fill these missing values, using the mean or median of the rest of the values is quite common in such cases.

We'll first get the average age.

```
summarise(titanic_table,avg_age =mean(age, na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##    avg_age
##      <dbl>
## 1 29.88113
```

then use the *replace_na* function from *tidyr* to fill in the NAs

```
titanic_table <- titanic_table %>% replace_na(list(age = 29.88113))
```

we could have perhaps used the mean age rounded to the nearest half year to fit in with the rest of the data.

Task 3: Lifeboat

Fill these empty slots with a dummy value e.g. the string 'None' or 'NA'

```
titanic_table$boat
```

```
##    [1] 2       11                              3       10
##    [9] D                       4       9       6       B
##   [17]         6       8       A       5       5       5       4
##   [25] 8               7       7       8       D               7
##   [33] 8       8               4       6       9
##   [41]         6       D       8       3                       5
##   [49] 6       3       3                               4       4
##   [57] C       4               5               6               4
##   [65] 5       5       4       8       7       6
##   [73] 4       11                      14              14      2
##   [81]                 7       7               4       5 9     3
##   [89] 8               3       3       3       13      5       5
##   [97]         6       2       1       1               7       4
##  [105] 4                       7       4       5               10
##  [113] 10      10                      10      1               5
##  [121] 5       5       5       5       5                       D
##  [129]         4       7       7               5       5
##  [137] B       3               3       7       7               5
##  [145] 5       3       3                       D               3
##  [153] 3       7               3                               4
##  [161] 4       10              8       15      D       14      D
##  [169] 6               C
##  [177] 8       5       5               2       8       2       3
##  [185]                 6       9       9               10
##  [193]         2               8       7               10      6
##  [201]                 7               6                       14
##  [209] 14      11                                      6       6
##  [217] 5               8       7       5
##  [225]                         8               8       3       6
##  [233]         7               A                       2
##  [241] 9               11                      8               6
##  [249]         4       4       4               4       3       9
##  [257] 1       11      4       7       3       5               11
##  [265] 3       7                                       6       7
##  [273] 7       3       3       3               6       3
##  [281] 1       5       4               6
##  [289] 8       8               8       5 7     5 7             B
##  [297] 4       D       7                               3
##  [305] 5                               8       8               8
##  [313]                 4       8 10            A               3
##  [321] D               8               10
##  [329]         11                              10
##  [337] 13      13              11      11      13      11      13
##  [345] 12                                      11      14
##  [353] 14      12              9                               13
##  [361] 13      13      14
##  [369]                 12      12              14
##  [377] 9       14              14      14
##  [385]         14              14      13
##  [393] 12                                              10
##  [401] 10      12      12                                      16
##  [409]                                                         12
##  [417]
##  [425]                         4       4               11
##  [433] 15              14              14      9       9
##  [441] 9       13                                      4
##  [449]         4                       10              10
##  [457]                                 12
##  [465]         11              12              10              9
##  [473]                                                 14      14
##  [481]         14      12      11      14
##  [489]                                 10              10
##  [497]                                         14      14      B
##  [505]
##  [513]                 D       D
##  [521] 7       11              13      9               9
##  [529]         12                              12      11
##  [537] 9               14              11      11      11
##  [545] 12              9               4       4       4       13
##  [553]         12                              12      16      11
##  [561] 10              13              9
##  [569]                 9                       16
##  [577]         14                      10              9       9
##  [585] 12              10      14      14      14      10      10
##  [593]         10                      9       14      9
##  [601]                         A       16      A       15      C
##  [609]                         11      13      15
##  [617]                                         16
##  [625]         D                                       A
##  [633]
##  [641]         15              15              13      15      C
##  [649]                                         C               D
##  [657]         C       C       C       C       C
##  [665] 15
##  [673]                 C
##  [681]                                                 13
##  [689]                                         13
##  [697]
##  [705]                                         16
##  [713]                 C                                       12
##  [721]                                         13
##  [729]                                         2       2       2
##  [737]                 12                              15
##  [745]         15      13 15 B
##  [753]                                         16      15      13
##  [761] 11              10      10              10
##  [769]                         C
##  [777]         B       13                                      D
##  [785]         16
##  [793]                 13                              15
##  [801]                                         13
##  [809]
##  [817]                 16      13      C D                     C D
##  [825]
##  [833]
##  [841]                                         15
##  [849]                         11              B
##  [857] 16      15      C                               C
##  [865]                 15      15
##  [873] C               C
##  [881] 15      A                                       D       13
##  [889]         15                                      15      15
##  [897]                         15
##  [905]         15                              15
##  [913]         13                      15      15
##  [921]         A       16      D
##  [929]                                                 2       2
##  [937] 2                                       10
##  [945]         C               13              14
##  [953]         6
##  [961]
##  [969] A       A       15
##  [977]                 15              10      15
##  [985] 15      13                                              15
##  [993]         16                              C               15 16
## [1001]         16      16      16      13              13
## [1009]
## [1017]         15                                              16
## [1025]         14      14              16
## [1033]         B       C       C       C                       16
## [1041] 15              16      16      16                      C
## [1049] C       C       C
## [1057] C       C                       D       13
## [1065] 9                       13
## [1073]                                         D               C
## [1081] B       13      13
## [1089] A
## [1097]
## [1105]
## [1113]
## [1121] 15      C       D       D
## [1129]                         9
## [1137]
## [1145]                                         13
## [1153]                                                         C
## [1161]                 14
## [1169]
## [1177]                                                 C
## [1185]                         13      13      13      11
## [1193]                                 11
## [1201]                                         16
## [1209]
## [1217] 13                                                      C
## [1225]                                 9
## [1233] B       15                      13              13 15
## [1241] 16                              14                      10
## [1249]                                                 15
## [1257] C       C       C               15      15
## [1265]
## [1273]                                         13 15
## [1281]                                                 C
## [1289]
## [1297]                         C
## [1305]
## 28 Levels:  1 10 11 12 13 13 15 13 15 B 14 15 15 16 16 2 3 4 5 5 7 ... D
```

```
summary(titanic_table$boat)
```

```
##                1      10      11      12      13   13 15 13 15 B      14
##      823       5      29      25      19      39       2       1      33
##       15   15 16      16       2       3       4       5     5 7     5 9
##       37       1      23      13      26      31      27       2       1
##        6       7       8    8 10       9       A       B       C     C D
##       20      23      23       1      25      11       9      38       2
##        D
##       20
```

We'll rewrite the levels for this column:

```
fct =levels(titanic_table$boat)
fct
```

```
##  [1] ""        "1"       "10"      "11"      "12"      "13"      "13 15"
##  [8] "13 15 B" "14"      "15"      "15 16"   "16"      "2"       "3"
## [15] "4"       "5"       "5 7"     "5 9"     "6"       "7"       "8"
## [22] "8 10"    "9"       "A"       "B"       "C"       "C D"     "D"
```

```
fct[1]='NA'
titanic_table <- titanic_table %>%
                 mutate(boat = factor(boat, labels=fct))
summary(titanic_table$boat)
```

```
##      NA       1      10      11      12      13   13 15 13 15 B      14
##     823       5      29      25      19      39       2       1      33
##      15   15 16      16       2       3       4       5     5 7     5 9
##      37       1      23      13      26      31      27       2       1
##       6       7       8    8 10       9       A       B       C     C D
##      20      23      23       1      25      11       9      38       2
##       D
##      20
```

Task 4: Cabin Numbers

You notice that many passengers don't have a cabin number associated with them. Create a new column has_cabin_numberwhich has 1 if there is a cabin number, and 0 otherwise.

```
summary(titanic_table$cabin)
```

```
##                     C23 C25 C27 B57 B59 B63 B66              G6
##            1014               6               5               5
##         B96 B98         C22 C26             C78               D
##               4               4               4               4
##              F2             F33              F4             A34
##               4               4               4               3
##     B51 B53 B55         B58 B60            C101            E101
##               3               3               3               3
##             E34             B18             B20             B22
##               3               2               2               2
##             B28             B35             B41             B45
##               2               2               2               2
##             B49              B5             B69             B71
##               2               2               2               2
##             B77             B78            C106            C116
##               2               2               2               2
##            C123            C124            C125            C126
##               2               2               2               2
##              C2             C31             C32             C46
##               2               2               2               2
##             C52             C54         C55 C57              C6
##               2               2               2               2
##         C62 C64             C65             C68              C7
##               2               2               2               2
##             C80             C83             C85             C86
##               2               2               2               2
##             C89             C92             C93         D10 D12
##               2               2               2               2
##             D15             D17             D19             D20
##               2               2               2               2
##             D21             D26             D28             D30
##               2               2               2               2
##             D33             D35             D36             D37
##               2               2               2               2
##            E121             E24             E25             E31
##               2               2               2               2
##             E33             E44             E46             E50
##               2               2               2               2
##             E67              E8           F G63           F G73
##               2               2               2               2
##             A10             A11             A14             A16
##               1               1               1               1
##             A18             A19             A20             A21
##               1               1               1               1
##             A23             A24             A26             A29
##               1               1               1               1
##             A31             A32             A36              A5
##               1               1               1               1
##              A6              A7              A9         (Other)
##               1               1               1              88
```

```
titanic_table <-mutate( titanic_table,has_cabin_number = as.integer(cabin != ''))
select(titanic_table,cabin, has_cabin_number)
```

```
## # A tibble: 1,309 x 2
##       cabin has_cabin_number
##      <fctr>            <int>
## 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                         0
## # ... with 1,299 more rows
```

Now save the cleaned table as a .csv.

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