Titanic Wikipedia Data Grab

Gina Reynolds, Claus O. Wilke

Dec. 2017

This file is written to collect the information about those on board Titanic from the Wikipedia pages on passengers and crew.

```r
library(htmltab)
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
```

```r
library(stringr)
library(here)
```

```
## here() starts at /Users/wilke/github/titanic.complete
```

```r
sessionInfo()
```

```
## R version 3.4.2 (2017-09-28)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] here_0.1      stringr_1.2.0 dplyr_0.7.4   htmltab_0.7.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14     assertthat_0.2.0 digest_0.6.12    rprojroot_1.2
##  [5] R6_2.2.2         backports_1.1.1  magrittr_1.5     evaluate_0.10.1
##  [9] rlang_0.1.4      stringi_1.1.5    bindrcpp_0.2     rmarkdown_1.7
## [13] tools_3.4.2      glue_1.2.0       yaml_2.1.14      compiler_3.4.2
## [17] pkgconfig_2.0.1  htmltools_0.3.6  bindr_0.1        knitr_1.17
## [21] tibble_1.3.4
```

Download raw data from Wikipedia

We grab the passenger and crew tables as raw html files from Wikipedia and store them in directory `./data-raw/RawData` for further processing.

```
if (!dir.exists(here("data-raw", "RawData"))) {
  dir.create(here("data-raw", "RawData"))
}

if (!file.exists(here("data-raw", "RawData", "Passengers2017-12-17.html")))  {
  download.file("https://en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic",
                destfile=here("data-raw", "RawData",
                              paste0("Passengers", Sys.Date(), ".html")))
}

if (!file.exists(here("data-raw", "RawData", "Crew2017-12-17.html"))) {
  download.file("https://en.wikipedia.org/wiki/Crew_of_the_RMS_Titanic",
                destfile=here("data-raw", "RawData",
                              paste0("Crew", Sys.Date(), ".html")))
}
```

Passengers

```
url <- here("data-raw", "RawData", "Passengers2017-12-17.html")
Table1 <- htmltab(url, 1, rm_nodata_cols = F)
Table2 <- htmltab(url, 2, rm_nodata_cols = F)
Table3 <- htmltab(url, 3, rm_nodata_cols = F)
Table1$Class <- "First"
Table2$Class <- "Second"
Table3$Class <- "Third"
Passengers <- bind_rows(Table1, Table2, Table3); dim(Passengers)
```

```
## [1] 1319    9
```

Passengers data cleanup

Note wikipedia mistake for passengers for Everett, Washington, USA.

```
Passengers[str_detect(Passengers$Boarded, "Everett"),]
```

```
##                                        Name Age          Hometown
## 1025 Jeanie, Mrs. Beanie The (née Meanie)    6 London, England, UK
## 1026  Meanie, Miss Maliza Mae (née Jones)   24 London, England, UK
##                    Boarded Destination Lifeboat Body Class Home country
## 1025 Everett, Washington, USA        14            <NA> <NA> Third  Southampton
## 1026 Everett, Washington, USA        14            <NA> <NA> Third  Southampton
```

Several entries are shifted one column to the left.

```
Passengers[str_detect(Passengers$Boarded, "Everett"),"Lifeboat"] <- 14
Passengers[str_detect(Passengers$Boarded, "Everett"),"Destination"] <- "Everett, Washington, USA"
```

```r
Passengers[str_detect(Passengers$Boarded, "Everett"),"Boarded"] <- NA
Passengers[c(1025,1026),]
```

```
##                               Name Age            Hometown Boarded
## 1025 Jeanie, Mrs. Beanie The (née Meanie)   6 London, England, UK   <NA>
## 1026  Meanie, Miss Maliza Mae (née Jones)  24 London, England, UK   <NA>
##               Destination Lifeboat Body Class Home country
## 1025 Everett, Washington, USA        14 <NA> Third  Southampton
## 1026 Everett, Washington, USA        14 <NA> Third  Southampton
```

Passenger survival

Survival is ID'd with Color. html is style in <tr field.

```r
Lines <- readLines(url)
BeforeTablesLine <- which(str_detect(Lines, '<th>Lifeboat'))
GrabWhich <- which(c(rep(T, nrow(Table1)), F, rep(T, nrow(Table2)), F, rep(T, nrow(Table3))))
TempLines <- Lines[BeforeTablesLine[1]:length(Lines)]
Passengers$survival_outcome <- str_detect(TempLines[str_detect(TempLines, "<tr")], "style")[GrabWhich]
```

Crew

```r
url <- here("data-raw", "RawData", "Crew2017-12-17.html")
Lines <- readLines(url)
BeforeTablesLine <- which(str_detect(Lines,'<th>Hometown'))
Crew <- data_frame()
for (i in 1:8){
  temp <- htmltab(url, i,rm_nodata_cols = F)
  TempLines <- Lines[BeforeTablesLine[i]:length(Lines)]
  temp$survival_outcome <- str_detect(TempLines[str_detect(TempLines, "<tr")], "style")[1:nrow(temp)]
  Crew <- bind_rows(Crew, temp)
}

Crew$Crew <- "Crew"
Table <- bind_rows(Passengers,Crew); dim(Table)
```

```
## [1] 2186    12
```

Join passenger and crew tables

Preparation for full join. Some people are classified as crew and passengers!

```r
Passengers$Hometown[Passengers$Hometown=="Belfast, Ireland, UK"] <- "Belfast, Ireland"
Passengers$Name[Passengers$Name=="Frost, Mr. Anthony Wood \"Archie\""] <- "Frost, Mr. Anthony Wood"
Passengers$Name[Passengers$Name=="Frost, Mr. Anthony Wood \"Artie\""] <- "Frost, Mr. Anthony Wood"

Table <- full_join(Passengers,Crew); dim(Table) #
```

```
## Joining, by = c("Name", "Age", "Hometown", "Boarded", "Lifeboat", "Body", "Class", "survival_outcome"
```

```
## [1] 2179    12
```

```r
Table$Crew[is.na(Table$Crew)] <- "Not Crew"
Table$survival_outcome <- ifelse(Table$survival_outcome, "Survived", "Perished")
```

Sex and age

I inspected titles to see if first names were all male. There is a Dr. Alice. I overwrite the case below, designating this individual as female. Also, any last names like John, Wallace and the like will be overwriten if there is a woman's title.

```r
# Sex
Table$sex <- NA
Table$sex[str_detect(Table$Name, "Master |Mr. |Mr |Father |Dr. |Sir |Don |Commander |Captain |Major |Col
Table$sex[str_detect(Table$Name, "Miss |Mrs.|Doña |Countess |Lady |Alice")] <- "Female"
table(Table$sex, as.numeric(Table$Age) >= 18, useNA = "ifany")
```

```
## Warning in table(Table$sex, as.numeric(Table$Age) >= 18, useNA = "ifany"):
## NAs introduced by coercion
```

```
##
##          FALSE TRUE <NA>
##   Female    81  406    3
##   Male     110 1565   12
##   <NA>       0    2    0
```

```r
table(Table$survival_outcome, Table$Lifeboat)
```

```
##
##             ?  1 10 11 12 13 14 14? 15 15? 16  2  3  4  5  6  7  8  9  A
##   Perished  0  0  0  0  0  0  1   0  1   0  1  0  0  1  0  0  1  0  0  4
##   Survived 18 12 33 48 20 66 45   1 58   1 33 18 38 41 36 25 25 27 41 12
##
##           A/14  B  C  D
##   Perished    0  0  0  0
##   Survived    1 29 48 21
```

```r
Table[is.na(Table$sex),] # These are probably men too - Position Trimmer and Fireman/Stoker
```

```
##                Name Age                         Hometown     Boarded
## 1500  Gosling, S.   26 Southampton, Hampshire, England Southampton
## 1529 Instance, T.   33 Southampton, Hampshire, England Southampton
##      Destination Lifeboat Body Class Home country survival_outcome
## 1500        <NA>     <NA> <NA>  <NA>          <NA>         Perished
## 1529        <NA>     <NA> <NA>  <NA>          <NA>         Perished
##           Position Crew  sex
## 1500       Trimmer Crew <NA>
## 1529 Fireman/Stoker Crew <NA>
```

Age

```r
# Age
Table$age_character <- Table$Age
```

```r
table(Table$age_character, useNA = "ifany")
```

```
##
##      --       1      10 10 mo.      11 11 mo.      12      13      14      15
##       2      11       6       3       4       1       6       6       8      11
##      16      17      18      19       2  2 mo.      20      21      22      23
##      28      38      57      62      13       1      80      81      98      68
##      24      25      26      27      28      29       3      30      31      32
##      93      85      74      78      91      76       7      97      74      91
##      33      34      35      36      37      38      39       4  4 mo.      40
##      51      51      60      69      42      43      51      15       1      43
##      41      42      43      44      45      46      47      48      49       5
##      29      39      24      27      32      17      19      24      13       5
##   5 mo.      50      51      52      53      54      55      56      57      58
##       1      16      10      12       3      10       9       5       7       7
##      59       6      60      61      62      63      64      65      66      67
##       9       6       8       7       8       6       5       2       3       1
##      69       7  7 mo.      70      71      74       8       9  9 mo.     n/a
##       1       9       1       1       3       1       9       9       2       2
##    <NA>
##       1
```

```r
Table$Age[str_detect(Table$Age,"m")] <- 0
Table$Age <- as.numeric(Table$Age)
```

```
## Warning: NAs introduced by coercion
```

```r
table(Table$Age, useNA = "ifany")
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14
##    10    11    13     7    15     5     6     9     9     9     6     4     6     6     8
##    15    16    17    18    19    20    21    22    23    24    25    26    27    28    29
##    11    28    38    57    62    80    81    98    68    93    85    74    78    91    76
##    30    31    32    33    34    35    36    37    38    39    40    41    42    43    44
##    97    74    91    51    51    60    69    42    43    51    43    29    39    24    27
##    45    46    47    48    49    50    51    52    53    54    55    56    57    58    59
##    32    17    19    24    13    16    10    12     3    10     9     5     7     7     9
##    60    61    62    63    64    65    66    67    69    70    71    74  <NA>
##     8     7     8     6     5     2     3     1     1     1     3     1     5
```

Convert variable names to snake case

```r
names(Table) <- str_replace(tolower(names(Table)), " ", "_")
```

Save data

```r
if(!dir.exists(here("data-raw", "DataProducts"))) {
  dir.create(here("data-raw", "DataProducts"))
}
str(Table)
```

```
## 'data.frame':    2179 obs. of  14 variables:
##  $ name            : chr  "Allen, Miss Elizabeth Walton" "Allison, Mr. Hudson Joshua Creighton" "and
##  $ age             : num  29 30 19 18 25 33 2 0 22 47 ...
##  $ hometown        : chr  "St. Louis, Missouri, US" "Montreal, Quebec, Canada" "Montreal, Quebec, Cal
##  $ boarded         : chr  "Southampton" "Southampton" "Southampton" "Southampton" ...
##  $ destination     : chr  "St. Louis, Missouri, US" "Montreal, Quebec, Canada" "Montreal, Quebec, Cal
##  $ lifeboat        : chr  "2" NA NA "11" ...
##  $ body            : chr  NA "135" "294" NA ...
##  $ class           : chr  "First" "First" "First" "First" ...
##  $ home_country    : chr  NA NA NA NA ...
##  $ survival_outcome: chr  "Survived" "Perished" "Perished" "Survived" ...
##  $ position        : chr  NA NA NA NA ...
##  $ crew            : chr  "Not Crew" "Not Crew" "Not Crew" "Not Crew" ...
##  $ sex             : chr  "Female" "Male" "Male" "Female" ...
##  $ age_character   : chr  "29" "30" "19" "18" ...
```

```r
write.csv(Table, here("data-raw", "DataProducts", "PeopleOnTitanic.csv"), row.names = F)

# rename to final data table name and save for package use
titanic_complete <- Table
devtools::use_data(titanic_complete, overwrite = TRUE)
```

```
## Saving titanic_complete as titanic_complete.rda to /Users/wilke/github/titanic.complete/data
```