Titanic Wikipedia Data Grab

Gina Reynolds

12/11/2017

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

```
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
```

```
library(stringr)
```

```
if(!dir.exists("RawData")){dir.create("RawData")}
download.file("https://en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic",
              destfile="RawData/Passengers.html")
download.file("https://en.wikipedia.org/wiki/Crew_of_the_RMS_Titanic",
              destfile="RawData/Crew.html")
```

Passengers

```
url="RawData/Passengers.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
```

note wikipedia mistake for passengers for Everett, Washington, USA

```
######## passengers #######
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
```

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

Survival is ID'd with Color... html is style in

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$Survived=str_detect(TempLines[str_detect(TempLines, "<tr")], "style")[GrabWhich] "'

Crew

```
##### crew ########
url="RawData/Crew.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$Survived=str_detect(TempLines[str_detect(TempLines, "<tr")], "style")[1:nrow(temp)]
Crew=bind_rows(Crew, temp)
}

Crew$Crew=1
Table=bind_rows(Passengers,Crew); dim(Table)
```

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

Join Passenger and Crew Tables

```
# Preparation for full join - some people classified as crew and passengers!
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", "Survived")
```

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

```
Table$Crew[is.na(Table$Crew)]=0
```

Sex and Age

```
# Sex
Table$sex=NA
# I inspected titles to see is first names were all male.  There is a Dr. Alice.
# I overwrite the case below, designating this individual as female.
# Also, any last names like John, Wallace and the like will be overwriten if there is a woman's title.
Table$sex[str_detect(Table$Name, "Master |Mr. |Mr |Father |Dr. |Sir |Don |Commander |Captain |Major |Co
Table$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
```

```
table(Table$Survived,Table$Lifeboat)
```

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

```
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 Survived       Position
## 1500        <NA>     <NA> <NA>  <NA>         <NA>    FALSE        Trimmer
## 1529        <NA>     <NA> <NA>  <NA>         <NA>    FALSE Fireman/Stoker
##      Crew  sex
## 1500    1 <NA>
## 1529    1 <NA>
```

Age

```r
# Age
Table$AgeCharacter=Table$Age
table(Table$AgeCharacter, 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
```

Save Data

```r
if(!dir.exists("DataProducts")){dir.create("DataProducts")}
save(Table, file = "DataProducts/PeopleOnTitantic.RData")
write.csv(Table, "DataProducts/PeopleOnTitantic.csv", row.names = F)
```