Homework 2: Married children on the Titanic

Tony Fraser

2023-07-17

Overview

For exploring, I decided to take the Titanic manifest down to just those under 18. In doing so, I noticed several females with "Mrs." in the Name column, and realized at that time, that was probably plenty old enough to be married. Back then, it was likely that children were either referred to by their name, or perhaps Miss or Master. As well, the age of consent was very differnet in the 1910's.

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
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)

#7.  BONUS - place the original .csv in a github file and read from a link.
loc<-"https://raw.githubusercontent.com/tonythor/pyspark-env/main/data/titanic.csv"
titanic_dataset <- loc
titanic <- read.csv(titanic_dataset)

#3. Create new column names for the new data frame.
titanic <- dplyr::rename(titanic, c("Class" = "Pclass"), c("Gender" = "Sex"))
titanic[c('LastName', 'FirstName')]<- str_split_fixed(titanic$Name, ',', 2)

#2. Create a new data frame with a subset of the columns and rows. Make
#   sure to rename it.
t_children = sqldf("select  Survived, Class, LastName, FirstName,
                        Gender, Age from titanic where age < 18
                        order by Class, Age desc")

#1 Use the summary function to gain an overview of the data set. Then
#  display the mean and median for at least two attributes.
```

```r
summary(t_children)
```

```
##     Survived          Class         LastName          FirstName
##  Min.   :0.0000   Min.   :1.000   Length:113         Length:113
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character
##  Median :1.0000   Median :3.000   Mode  :character   Mode  :character
##  Mean   :0.5398   Mean   :2.584
##  3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :3.000
##     Gender               Age
##  Length:113         Min.   : 0.420
##  Class :character   1st Qu.: 3.000
##  Mode  :character   Median : 9.000
##                     Mean   : 9.041
##                     3rd Qu.:16.000
##                     Max.   :17.000
```

```r
sprintf("P Class : mean:%s median:%s",
        round(mean(t_children$Class), digits=4),
        round(median(t_children$Class),digits=4))
```

```
## [1] "P Class : mean:2.5841 median:3"
```

```r
sprintf("Survived: mean:%s median:%s",
        round(mean(t_children$Survived), digits=4),
        round(mean(t_children$Survived), digits=4))
```

```
## [1] "Survived: mean:0.5398 median:0.5398"
```

```r
# Add that column in there based on salutations
t_children = t_children %>%
  mutate(PossiblyMarried = case_when(
    grepl("Mrs.", FirstName) ~ "True",
    grepl("Mister.", FirstName ) ~ "True",
    grepl("Mr.", FirstName ) ~ "True",
    .default = ""
  ))

#5 For at least 3 values in a column please rename so that every value in
#  that column is renamed.
replacements  = c('Miss.'='','Master.'='', 'Mrs.'='', 'Mister.'='', 'Mr.'='')
t_children$FirstName<- str_replace_all(t_children$FirstName, replacements )


# The first name column is pretty long. Let's truncate it so it'll fit on one page.
t_children = t_children %>% mutate(FirstName = str_trunc(FirstName, width = 10))

married = nrow(sqldf("select PossiblyMarried from t_children where length(PossiblyMarried) > 1"))
sprintf('Number of children in this dataset: %s', nrow(t_children))
```

```
## [1] "Number of children in this dataset: 113"
```

```r
sprintf('Number of children who might already be married: %s', married)
```

```
## [1] "Number of children who might already be married: 26"
```

```r
#6. Display enough rows to see examples of all of steps 1-5 above.
head(t_children, 50)
```

```
##    Survived Class             LastName  FirstName Gender   Age PossiblyMarried
## 1         1     1 Penasco y Castellana   Victo... female 17.00            True
## 2         1     1               Thayer   John ...   male 17.00            True
## 3         1     1                 Dick   Alber... female 17.00            True
## 4         1     1               Hippach   Jean ... female 16.00
## 5         1     1               Maioni    Roberta female 16.00
## 6         1     1                Lines   Mary ... female 16.00
## 7         1     1               Madill   Georg... female 15.00
## 8         1     1               Carter   Lucil... female 14.00
## 9         1     1               Carter   Willi...   male 11.00
## 10        1     1                Dodge   Washi...   male  4.00
## 11        0     1              Allison   Helen... female  2.00
## 12        1     1              Allison   Hudso...   male  0.92
## 13        1     2                Ilett     Bertha female 17.00
## 14        1     2              Lehmann     Bertha female 17.00
## 15        0     2              Gaskell     Alfred   male 16.00            True
## 16        0     2                 Mudd   Thoma...   male 16.00            True
## 17        1     2               Nasser   Nicho... female 14.00            True
## 18        1     2            Mellinger   Madel... female 13.00
## 19        1     2              Collyer   Marjo... female  8.00
## 20        1     2               Davies   John ...   male  8.00
## 21        1     2                 Hart   Eva M... female  7.00
## 22        1     2               Harper   Annie... female  6.00
## 23        1     2                 West   Const... female  5.00
## 24        1     2               Becker   Mario... female  4.00
## 25        1     2                Wells       Joan female  4.00
## 26        1     2              Laroche   Simon... female  3.00
## 27        1     2             Navratil   Michel M   male  3.00
## 28        1     2             Richards   Willi...   male  3.00
## 29        1     2             Navratil   Edmon...   male  2.00
## 30        1     2                Quick   Phyll... female  2.00
## 31        1     2               Becker   Richa...   male  1.00
## 32        1     2               Mallet      Andre   male  1.00
## 33        1     2             Caldwell   Alden...   male  0.83
## 34        1     2             Richards   Georg...   male  0.83
## 35        1     2           Hamalainen      Viljo   male  0.67
## 36        1     3            Andersson   Erna ... female 17.00
## 37        0     3              Attalah     Malake female 17.00
## 38        0     3                Calic       Jovo   male 17.00            True
## 39        0     3               Kallio   Nikol...   male 17.00            True
## 40        0     3                Calic      Petar   male 17.00            True
## 41        0     3                Elias   Josep...   male 17.00            True
## 42        0     3               Jensen   Svend...   male 17.00            True
## 43        0     3            Culumovic       Jeso   male 17.00            True
## 44        0     3              Goodwin   Lilli... female 16.00
## 45        0     3                 Ford   Willi...   male 16.00            True
## 46        0     3                 Osen   Olaf ...   male 16.00            True
## 47        1     3              Gilnagh   Kathe... female 16.00
## 48        1     3                 Carr   Helen... female 16.00
## 49        1     3           Sunderland   Victo...   male 16.00            True
## 50        0     3               Panula   Ernes...   male 16.00            True
```