Exploring Data : Titanic Analysis

Huong Thai

October 4th, 2016

Collaborators: Emma Sprio, Tim

```r
# Load some helpful libraries
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R
## version 3.3.2
```

Exploring Data:

The sinking of the RMS Titanic[1] is a notable historical event. The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning of 15 April 1912, after colliding with an iceberg during her maiden voyage from Southampton to New York City. Of the 2,224 passengers and crew aboard, more than 1,500 died in the sinking, making it one of the deadliest commercial peacetime maritime disasters in modern history.

The disaster was greeted with worldwide shock and outrage at the huge loss of life and the regulatory and operational failures that had led to it. Public inquiries in Britain and the United States led to major improvements in maritime safety. One of their most important legacies was the establishment in 1914 of the International Convention for the Safety of Life at Sea (SOLAS)[2], which still governs maritime safety today. Additionally, several new wireless regulations were passed around the world in an effort to learn from the many missteps in wireless communications which could have saved many more passengers.

The data we will explore in this lab were originally collected by the British Board of Trade in their investigation of the sinking. You can download these data in CSV format from Canvas. Researchers should note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost.

Formulate a Question:

Today, we will consider two questions in our exploration:

- Who were the Titanic passengers? What characteristics did they have?
- What passenger characteristics or other factors are associated with survival?

[1] https://en.wikipedia.org/wiki/RMS_Titanic

[2] https://en.wikipedia.org/wiki/International_Convention_for_the_Safety_of_Life_at_Sea

Read and Inspect Data:

Load the Titanic dataset into R. You can do so by executing the following code.

```
setwd("D:/Huong_Projects Archive/2016-Fall/INFX 573/Lab exercise/Lab2")
titanic <- read.csv("titanic.csv")
titanic <- tbl_df(titanic)  # transform the data into a data frame tbl
```

Inspect our data

```
summary(titanic)
```

```
##      pclass         survived    
##  Min.   :1.000   Min.   :0.000  
##  1st Qu.:2.000   1st Qu.:0.000  
##  Median :3.000   Median :0.000  
##  Mean   :2.295   Mean   :0.382  
##  3rd Qu.:3.000   3rd Qu.:1.000  
##  Max.   :3.000   Max.   :1.000  
##                                 
##                                 name     
##  Connolly, Miss. Kate            :   2  
##  Kelly, Mr. James                :   2  
##  Abbing, Mr. Anthony             :   1  
##  Abbott, Master. Eugene Joseph   :   1  
##  Abbott, Mr. Rossmore Edward     :   1  
##  Abbott, Mrs. Stanton (Rosa Hunt):   1  
##  (Other)                         :1301  
##      sex           age         
##  female:466   Min.   : 0.1667  
##  male  :843   1st Qu.:21.0000  
##               Median :28.0000  
##               Mean   :29.8811  
##               3rd Qu.:39.0000  
##               Max.   :80.0000  
##               NA’s   :263      
##      sibsp            parch      
##  Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.:0.000  
##  Median :0.0000   Median :0.000  
##  Mean   :0.4989   Mean   :0.385  
##  3rd Qu.:1.0000   3rd Qu.:0.000  
##  Max.   :8.0000   Max.   :9.000  
##                                  
##       ticket          fare       
##  CA. 2343:  11   Min.   :  0.000
##  1601    :   8   1st Qu.:  7.896
##  CA 2144 :   8   Median : 14.454
##  3101295 :   7   Mean   : 33.295
##  347077  :   7   3rd Qu.: 31.275
##  347082  :   7   Max.   :512.329
##  (Other) :1261   NA’s   :1
##              cabin      embarked
##                 :1014    :  2
##  C23 C25 C27    :   6   C:270
##  B57 B59 B63 B66:   5   Q:123
##  G6             :   5   S:914
##  B96 B98        :   4
##  C22 C26        :   4
##  (Other)        : 271
##       boat          body
##         :823   Min.   :  1.0
##  13     : 39   1st Qu.: 72.0
##  C      : 38   Median :155.0
##  15     : 37   Mean   :160.8
##  14     : 33   3rd Qu.:256.0
##  4      : 31   Max.   :328.0
##  (Other):308   NA’s   :1188
##                 home.dest
##                      :564
##  New York, NY        : 64
##  London              : 14
##  Montreal, PQ        : 10
##  Cornwall / Akron, OH:  9
##  Paris, France       :  9
##  (Other)             :639
```

```
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 ...
```

```
head(titanic)  # Look at the first few rows of the data frame
```

```
## # A tibble: 6 × 14
##   pclass survived
##    <int>    <int>
## 1      1        1
## 2      1        1
## 3      1        0
## 4      1        0
## 5      1        0
## 6      1        1
## # ... with 12 more variables: name <fctr>,
## #   sex <fctr>, age <dbl>, sibsp <int>,
## #   parch <int>, ticket <fctr>, fare <dbl>,
## #   cabin <fctr>, embarked <fctr>,
## #   boat <fctr>, body <int>,
## #   home.dest <fctr>
```

```
tail(titanic)  # Look at the last few rows of the data frame
```

```
## # A tibble: 6 × 14
##   pclass survived                      name
##    <int>    <int>                    <fctr>
## 1      3        0     Yousseff, Mr. Gerious
## 2      3        0      Zabour, Miss. Hileni
## 3      3        0     Zabour, Miss. Thamine
## 4      3        0 Zakarian, Mr. Mapriededer
## 5      3        0       Zakarian, Mr. Ortin
## 6      3        0        Zimmerman, Mr. Leo
## # ... with 11 more variables: sex <fctr>,
## #   age <dbl>, sibsp <int>, parch <int>,
## #   ticket <fctr>, fare <dbl>, cabin <fctr>,
## #   embarked <fctr>, boat <fctr>,
## #   body <int>, home.dest <fctr>
```

```
nrow(titanic)
```

```
## [1] 1309
```

```
ncol(titanic)
```

```
## [1] 14
```

```
nchar(titanic)

##    pclass  survived      name       sex
##      3935      3935      6759      3935
##       age     sibsp     parch    ticket
##      5285      3935      3935      6391
##      fare     cabin  embarked      boat
##      8901      4369      3935      4269
##      body home.dest
##      5325      5248

# Use the summary function to inspect
# variables
sum(is.na(titanic$pclass))

## [1] 0

sum(is.na(titanic$survived))

## [1] 0

sum(is.na(titanic$name))

## [1] 0

sum(is.na(titanic$sex))

## [1] 0

sum(is.na(titanic$age))

## [1] 263

sum(is.na(titanic$sibsp))

## [1] 0

sum(is.na(titanic$parch))

## [1] 0

sum(is.na(titanic$ticket))

## [1] 0

sum(is.na(titanic$fare))

## [1] 1

sum(is.na(titanic$cabin))
```

```
## [1] 0

sum(is.na(titanic$embarked))

## [1] 0

sum(is.na(titanic$boat))

## [1] 0

sum(is.na(titanic$body))

## [1] 1188

sum(is.na(titanic$home.dest))

## [1] 0

survivedpeople <- subset(titanic, survived ==
    1)
dim(survivedpeople)

## [1] 500  14

str(survivedpeople)

## Classes ’tbl_df’, ’tbl’ and ’data.frame’:    500 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ name     : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 22 24 31 46 51 70 73 93 94 100 ...
##  $ sex      : Factor w/ 2 levels "female","male": 1 2 2 1 1 1 1 1 2 1 ...
##  $ age      : num  29 0.917 48 63 53 ...
##  $ sibsp    : int  0 1 0 1 2 1 0 0 0 0 ...
##  $ parch    : int  0 2 0 0 0 0 0 0 0 1 ...
##  $ ticket   : Factor w/ 929 levels "110152","110413",..: 188 50 125 93 77 834 796 119 297 801 ...
##  $ fare     : num  211.3 151.6 26.6 78 51.5 ...
##  $ cabin    : Factor w/ 187 levels "","A10","A11",..: 45 81 151 147 63 99 35 1 10 50 ...
##  $ embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 2 2 4 4 2 ...
##  $ boat     : Factor w/ 28 levels "","1","10","11",..: 13 4 14 3 28 15 23 19 25 19 ...
##  $ body     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ home.dest: Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 238 163 23 238 259 1 159 231 ...

prop.table(table(titanic$survived))

## 
##        0        1 
## 0.618029 0.381971

table(titanic$sex, titanic$survived)
```

```
##
##            0   1
##   female 127 339
##   male   682 161
```

```
prop.table(table(titanic$sex, titanic$survived),
    margin = 1)
```

```
##
##                  0         1
##   female 0.2725322 0.7274678
##   male   0.8090154 0.1909846
```

Initial observations

First, data set lets us know we have 1309 observations, or passengers, to analyze here: 1309 entries, 0 to 1309.

Next it shows us all of the columns and each of them holds information of the observations, such as the name, sex or age, etc. These colunms let us know how many values each of them contains. Details: 14 colums show 14 variables ($ pclass, $ survived, $ name, $ sex , $ age , $ sibsp , $ parch, $ ticket, $ fare , $ cabin, $ embarked, $ boat, $ body, $ home.dest)

In general, it can be seen that the dataset contains various characteristics about individual passengers inluding their name, sex, and age. It can also noticed that there is a variable called 'survived' which is likely to contain data about whether that person survived after the crash. We also notice that some variables have missing data, represented by NAs. In details: - There are some variables appearing to have a lot of missing values: Age (263 counts), Fare (1 count), Body (1188 count) - Within 1309 observations (the survived people is set value ==1 in the dataset) there were 500 survived people. It means there was more than one-third of passengers survived. We also can see that the proporation of the survived female is much higher than it of the male.

Review the type of variables, I have some points to improve the correctness of the dataset. 1. Variable "survived": we can use logical type so we can make sure that there are only two value True or False, instead of "int" type which provides more than two options. 2. Variable "age": it should be set to be "int" instead of "dbl" or "num". "int" type only allow integer, however in this dataset, the age value can be any type of numeric such as 22.5. The age should not be fomatted in so-called decimal 3. Variable "body": There are many missing values so it should be removed from the dataset. 4. Varible "boat": There are many missing values and unconsistent code like number or character in this column. So I recommend to clear it out from the dataset to make the dataset non-distracted.

The summary function also reveals that the 'survived' variable is treated as a numeric variable in R. This characteristic is more appropriately a categorical variable and thereforeit will be re-casted it as a factor. The same goes for 'pclass'.

```
# Re-cast categorical variables to be factor
# data types
titanic$pclass <- as.factor(titanic$pclass)
titanic$survived <- as.factor(titanic$survived)
```

Easy Solution:

I will create a basic visualization to help us understand the distributions of age for Titanic passengers.

```
ggplot(data = titanic, aes(age)) + geom_histogram(fill = "blue")
```

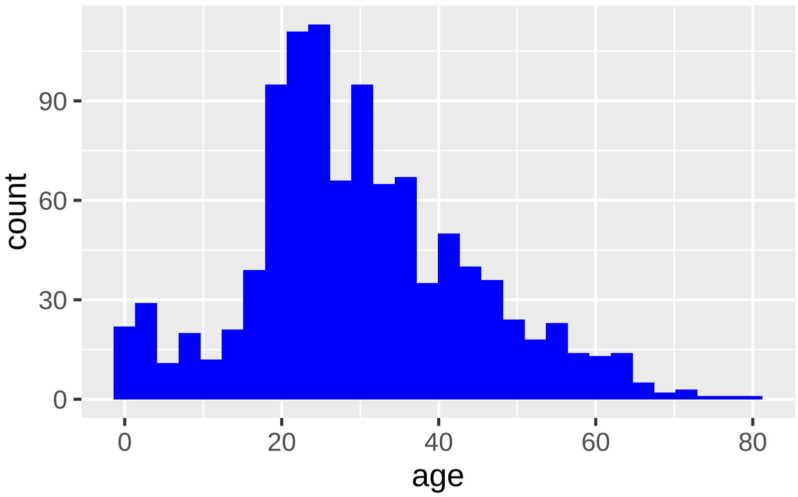


Figure 1: Age of Passenders Aboard the Titanic

We might go further to look at how passenger age might be related to survival.

Figure to show age distribution by survival

```
ggplot(data = titanic, aes(age, survived)) + geom_point(size = 2,
    alpha = 1, color = "red")
```

This simple visualization is good at showing the number of survival and non-survival. I can interpret the difference in age distribution between 2 variables. However, I can't easily see the correlation

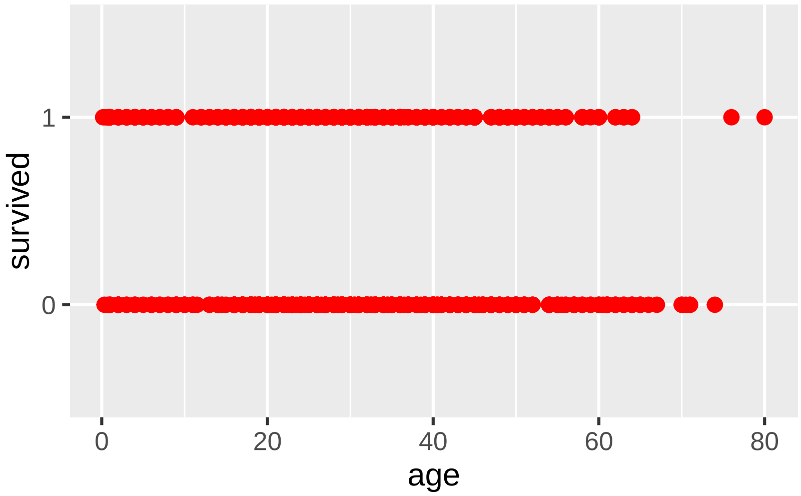


Figure 2: Survival and Passenger Age

between Age and Survival. In other words, I don't see which age had the highest survival count. In addition, the axis X distracted my attention to the variable survived.

I come up to produce a new figure that you think does a better job of helping you explore the association between passenger age and survival.

```
ggplot(data = titanic, aes(age, survived)) + geom_bar(stat = "identity")
```

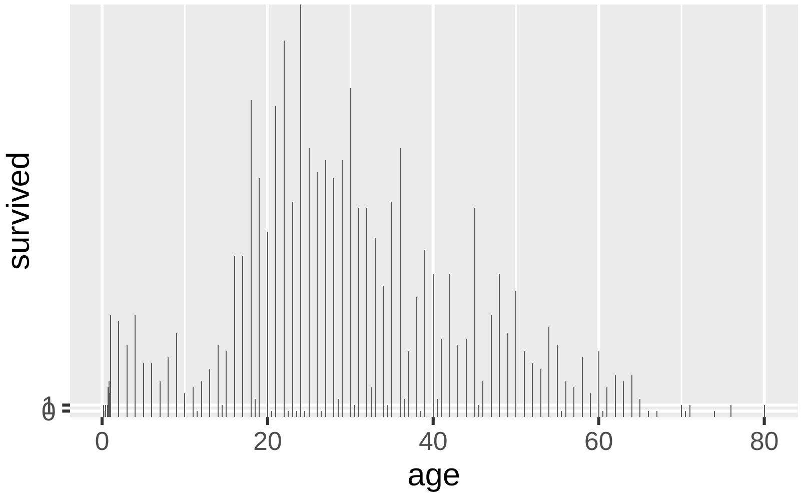


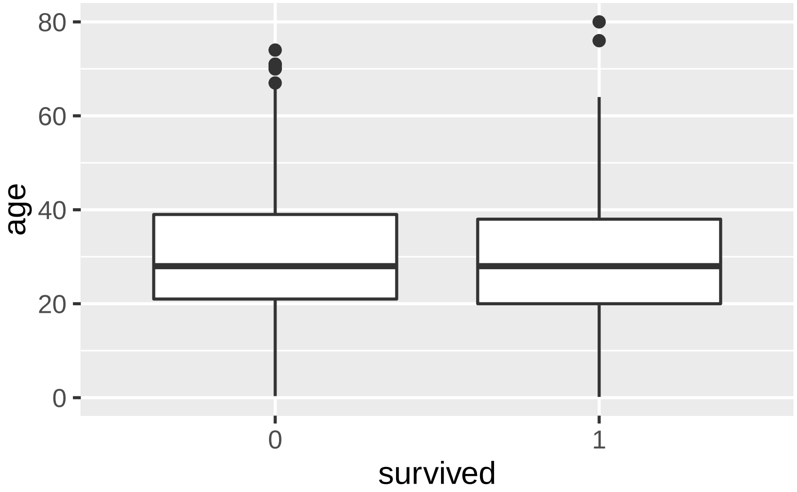
Figure 3: Survival and Passenger Age

I decide to use bar chart to compare the number of survival in age categories. Because I can see which age group had the best chance to survive.

However, these figures does not do a good job of displaying the data. In particular, it does not help us understand the distribution of ages by survival. A better plot would be a boxplot to show the age distribution for each value of the survival variable.

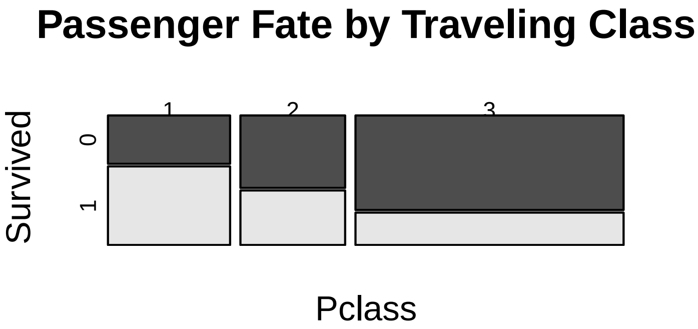
```
ggplot(data = titanic, aes(survived, age)) + geom_boxplot()
```

```
## Warning: Removed 263 rows containing non-finite
## values (stat_boxplot).
```



We want to look at the relationship between survival and passenger class to determine if evidence suggests high surivial rates for upper class passengers. In the following figure we see not only how many passengers fall into each class, relatively, but also what proportion survived. Data suggests that passengers in 1st and 2nd class cabins had higher rates of surivival, compared the 3rd class passengers.

```
mosaicplot(titanic$pclass ~ titanic$survived,
    main = "Passenger Fate by Traveling Class",
    shade = FALSE, color = TRUE, xlab = "Pclass",
    ylab = "Survived")
```



From this visualization, we might want to build a statistical model to identify the relative influence of each factor on survival or predict the survival of a passenger.