INFX 573 Lab: Exploring Data

Shuyang Wu

October 4th, 2016

Collaborators: Chaofan Han

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio. You will also need to install two R packages that we will be using throughout the course. You can install these packages in R using the following commands:

Hint: If you encounter any errors, you might need to install other dependences, including 'Rcpp' and 'tibble'.

```
# Install packages if you don't have them
install.packages("tidyverse")
install.packages("rticles")
install.packages("tufte")
```

1. Download the `week2a_lab.Rmd` file from Canvas. Open `week2a_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week2a_lab.Rmd`. You will also want to download the `titanic.txt` data file, containing a data about passengers aboard the Titanic.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, rename the R Markdown file to `YourLastName_YourFirstName_lab2a.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

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

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

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

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.

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

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?

Read and Inspect Data:

To begin, we need to load the Titanic dataset into R. You can do so by executing the following code.

Note: We will learn more about data frame tbl next week. For now, consider it a data frame with tidy printing.

```
titanic <- read.csv("titanic.csv")
titanic <- tbl_df(titanic)  # transform the data into a data frame tbl
```

Next, we want to inspect our data. We don't want to assume that are data in exactly as we expect it to be after reading it into R. It is helpful to inspect the data object, confirming to looks as expected.

Hint: Some helpful functions for inspecting data are: `head()`, `tail()`, `str()`, `nrow()`, `ncol()`, `table()`

Try editing to following code chunk to look at the top and bottom of your data frame. Perform any other inspection operations you deem necessary. Do you observe anything concerning?

```
# Edit me to add R code!
head(titanic)
```

```
## # A tibble: 6 x 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)
```

```
## # A tibble: 6 x 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>
```

```
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 ...
```

```
nrow(titanic)  #get the number of passenger
```

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

```
summary(titanic)  #basic summary
```

```
##      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
```

```
table(titanic$survived)  #find survival rate
```

```
##
##   0   1
## 809 500
```

Note: Remember to describe your results! You should write a reponse to accompany your analysis that comments on what you find.

The dataset contains the passengers class, name, survival status, sex, age, number of siblings or spouses aboard, number of parents or children aboard, ticket number, passenger fare, cabin that they stayed in, and the port of embarkation. There are 1309 passangers in total, 466 are female and 843 are male, 809 death and 500 survival.

Think about the variables in this data as they are defined. Which variables might you want to re-cast to be the appropriate data type in R?

Hint: Consider how variables are measured and how that matches available data types in R.

Transform the data type of varibles you identify as improperly cast.

```
# Edit me to add R code!
titanic$pclass <- as.factor(titanic$pclass)
titanic$survived <- as.logical(titanic$survived)  #turn into TRUE or FALSE
titanic$age <- as.integer(titanic$age)  #age should be integers
```

Trying the Easy Solution First:

First, we want to explore who the passengers aboard the Titanic were. There are many ways we might go about this. Consider for example trying to understand the ages of passengers. We can create a basic visualization to help us understand the distributions of age for Titanic passengers.

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

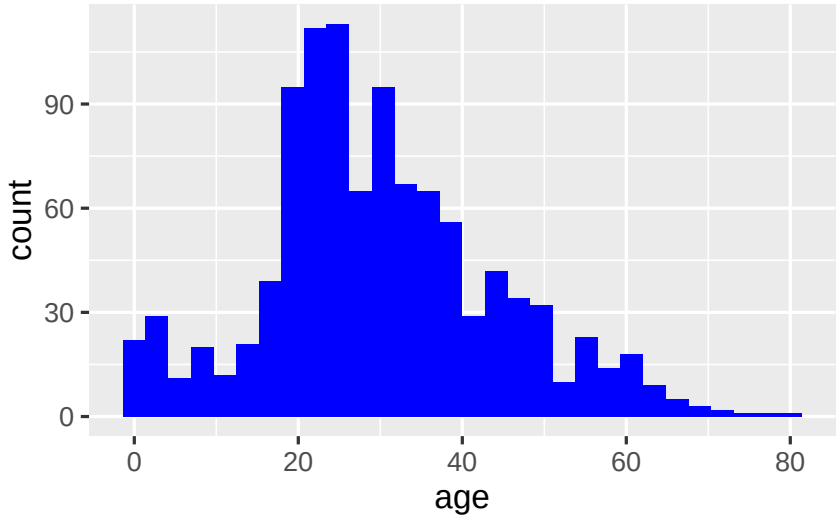

Figure 1: Age of Passenders Aboard the Titanic

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

```
ggplot(data = titanic, aes(age, survived)) + geom_point(size = 2,
    alpha = 0.5, color = "red")
```

Do you like the above figure? Why or why not? Produce a new figure that you think does a better job of helping you explore the association between passenger age and survival.

Note: You need to add a written response here!

Identify one additional data feature you want to explore. Produce one visualization that explore this feature. Describe why you think this is interesting and what you find.

Note: Don't forget to describe what you find!

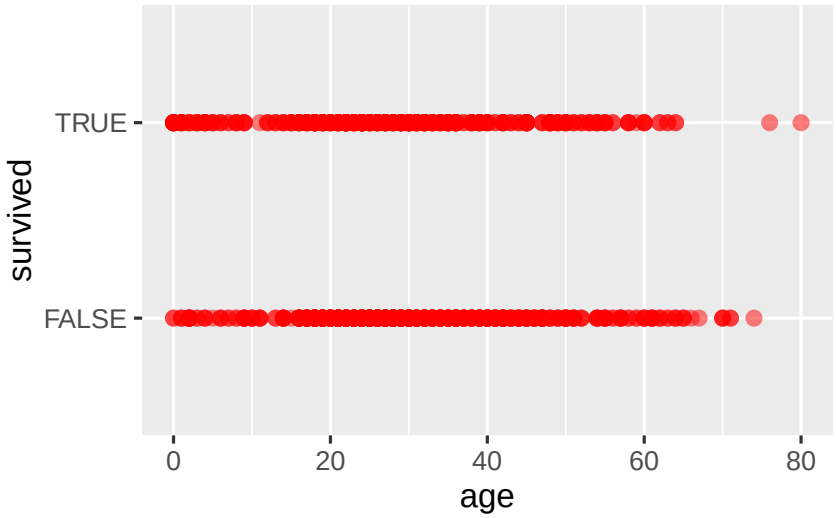

Figure 2: Survival and Passenger Age

What Next?

Consider the exploratory analysis you completed in the lab exercise. What would you do next?

Note: You need to add a written response here!