IMT 573: Lab 2 - Exploring Data

Alexander Davis

Thursday, October 08, 2020

Collaborators:

1.
2. …

Objectives

In this lab we will dive right in and explore a found dataset. Our aim is to practice getting to know our data. We will follow the steps of exploratory data analysis in this endeavor. This demo will give you an introduction to the very popular data visualization package `ggplot`. We will start with the basics today, and see more of this particular tool later on in the course.

```
# Load some helpful libraries for this course
library(tidyverse)
library(ggplot2)
```

Data Background

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.

Formulating 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.

```
library(titanic)
titanic_data <- data.frame(titanic_train)
```

Next, we want to inspect our data. We don't want to assume that our data is 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.

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?

```
head(titanic_data)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                   Name    Sex Age SibSp Parch
## 1                              Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                               Heikkinen, Miss. Laina female  26     0     0
## 4         Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                             Allen, Mr. William Henry   male  35     0     0
## 6                                     Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
tail(titanic_data)
```

```
##     PassengerId Survived Pclass                                     Name    Sex
## 886         886        0      3     Rice, Mrs. William (Margaret Norton) female
## 887         887        0      2                    Montvila, Rev. Juozas   male
## 888         888        1      1             Graham, Miss. Margaret Edith female
## 889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female
## 890         890        1      1                    Behr, Mr. Karl Howell   male
## 891         891        0      3                      Dooley, Mr. Patrick   male
##     Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886  39     0     5     382652 29.125              Q
## 887  27     0     0     211536 13.000              S
## 888  19     0     0     112053 30.000   B42        S
## 889  NA     1     2 W./C. 6607 23.450              S
## 890  26     0     0     111369 30.000  C148        C
## 891  32     0     0     370376  7.750              Q
```

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: Categorical variables are better suited as factors. You might want to check* `as.factor`

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

```
titanic_data$survived <- as.factor(titanic_data$Survived)
```

Trying the Easy Solution First

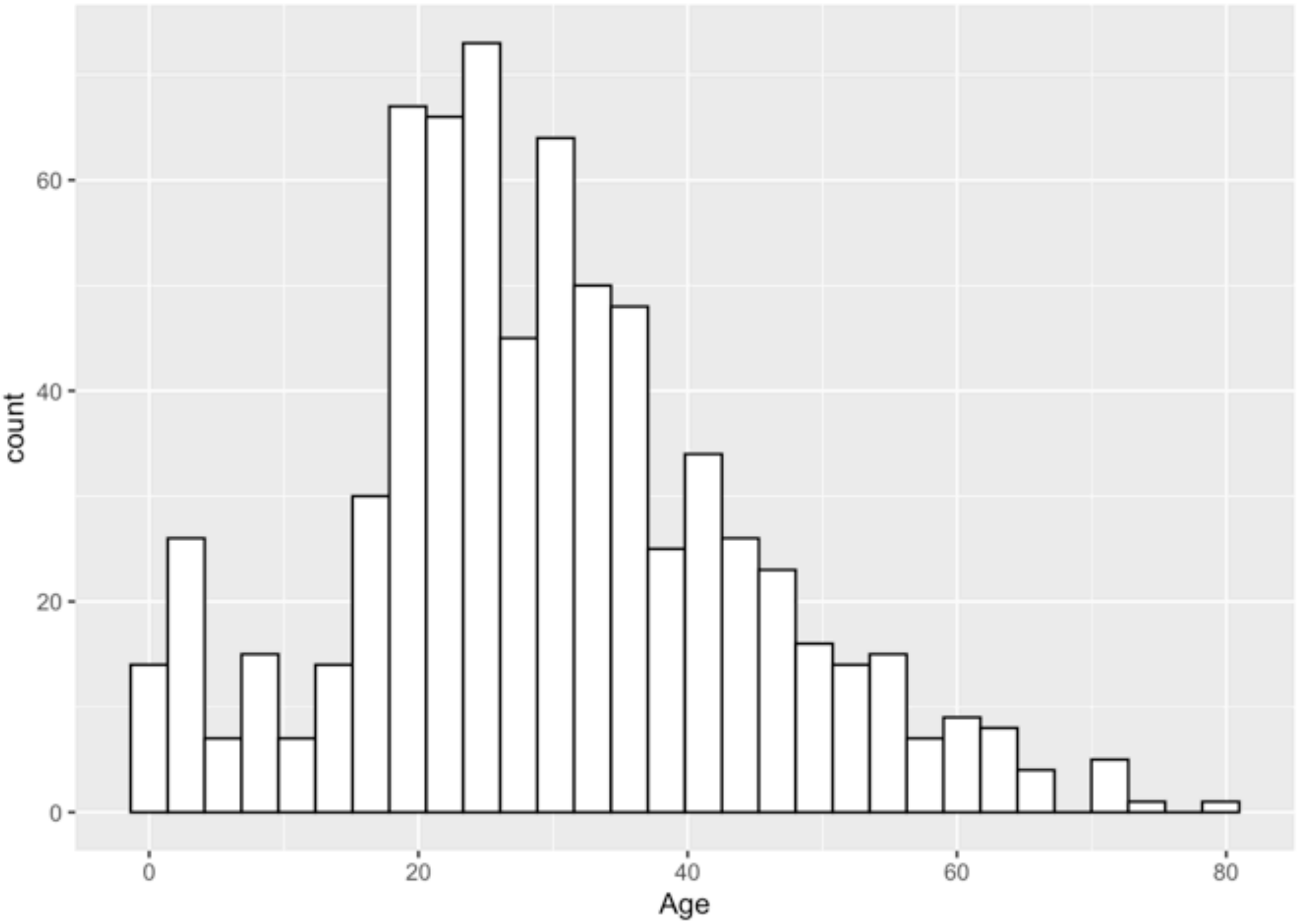
First, we want to explore who the passengers aboard the Titanic were. There are many ways we might go about this. Consider for example trying to understand the ages of passengers. We can create a basic visualization to help us understand the distributions of age for Titanic passengers.

You could also explore other characteristics, like passenger class, gender.

```
passenger_age <- data.frame(titanic_data[1],titanic_data[6])
##passenger_age
ggplot(passenger_age, aes(x=Age)) + geom_histogram(color="black", fill="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```

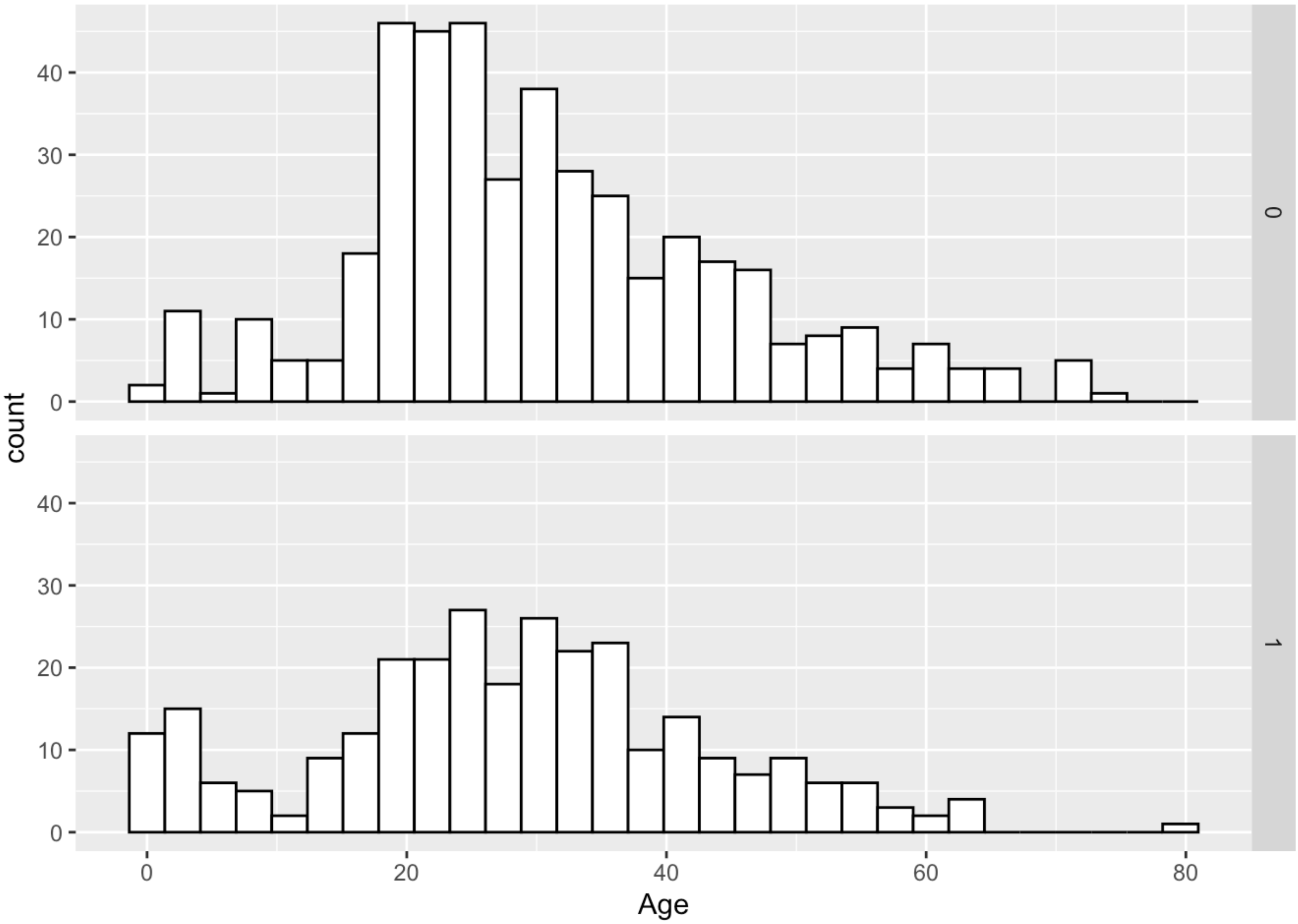


We might go further to look at how passenger age might be related to survival (or whichever other characteristic you are interested in).

```
passenger_age_survival <- data.frame(passenger_age, titanic_data[2])
#passenger_age_survival
ggplot(passenger_age_survival, aes(x=Age)) + geom_histogram(color="black", fill="white") + facet_grid(vars(Survived))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



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.

I think this does an okay job at exploring some association between age and survival. You can pick out that a significant portion of younger individuals survived in the below histogram versus the above histogram.

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.

I would take a look at class and how that could determine a possible survival rate.

What Next?

Consider the exploratory analysis you just completed. What would you do next? Can you challenge your first approach and come up with additional questions and solutions? See hint [3].

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

2. https://en.wikipedia.org/wiki/International_Convention_for_the_Safety_of_Life_at_Sea (https://en.wikipedia.org/wiki/International_Convention_for_the_Safety_of_Life_at_Sea)↩

3. https://bio304-class.github.io/bio304-fall2017/data-story-titanic.html (https://bio304-class.github.io/bio304-fall2017/data-story-titanic.html)↩