INFX 573 Lab: Exploring Data

Charudatta Deshpande

October 10th, 2017

Collaborators: N/A

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("tufte")
```

1. Download the `lab03a_titanic.rmd` file from Canvas. Open `lab03a_titanic.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)
```

```
## Warning: Installed Rcpp (0.12.10) different from Rcpp used to build dplyr (0.12.11).
## Please reinstall dplyr to avoid random crashes or undefined behavior.
```

Exploring Data:

The sinking of the RMS Titanic[1] is a notable historical event. The

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

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 drowned, 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.

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

The data 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. That there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost. You can find the variable definitions at kaggle.com.

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.

```
titanic <- read.csv("titanic.csv")
# Note: you can read compressed files
# directly, no need to manually uncompress
```

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?

```
# Convert the file into data.table format for
# ease of analysis and exploration. I
# personally use this frequently.
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
as.data.table(titanic)
```

```
##       pclass survived
##    1:      1        1
##    2:      1        1
##    3:      1        0
##    4:      1        0
##    5:      1        0
##   ---
## 1305:      3        0
## 1306:      3        0
## 1307:      3        0
## 1308:      3        0
## 1309:      3        0
##                                                name
##    1:                    Allen, Miss. Elisabeth Walton
##    2:                   Allison, Master. Hudson Trevor
##    3:                     Allison, Miss. Helen Loraine
##    4:             Allison, Mr. Hudson Joshua Creighton
##    5: Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
##   ---
## 1305:                             Zabour, Miss. Hileni
## 1306:                            Zabour, Miss. Thamine
## 1307:                        Zakarian, Mr. Mapriededer
## 1308:                              Zakarian, Mr. Ortin
## 1309:                               Zimmerman, Mr. Leo
##          sex     age sibsp parch ticket
##    1: female 29.0000     0     0  24160
##    2:   male  0.9167     1     2 113781
```

```
##    3: female  2.0000     1     2 113781
##    4:   male 30.0000     1     2 113781
##    5: female 25.0000     1     2 113781
##   ---
## 1305: female 14.5000     1     0   2665
## 1306: female      NA     1     0   2665
## 1307:   male 26.5000     0     0   2656
## 1308:   male 27.0000     0     0   2670
## 1309:   male 29.0000     0     0 315082
##           fare   cabin embarked boat body
##    1: 211.3375      B5        S    2   NA
##    2: 151.5500 C22 C26        S   11   NA
##    3: 151.5500 C22 C26        S        NA
##    4: 151.5500 C22 C26        S       135
##    5: 151.5500 C22 C26        S        NA
##   ---
## 1305:  14.4542               C       328
## 1306:  14.4542               C        NA
## 1307:   7.2250               C       304
## 1308:   7.2250               C        NA
## 1309:   7.8750               S        NA
##                            home.dest
##    1:                    St Louis, MO
##    2: Montreal, PQ / Chesterville, ON
##    3: Montreal, PQ / Chesterville, ON
##    4: Montreal, PQ / Chesterville, ON
##    5: Montreal, PQ / Chesterville, ON
##   ---
## 1305:
## 1306:
## 1307:
## 1308:
## 1309:
```

```
# Below code with display first few rows of
# the titanic CSV file.
head(titanic)
```

```
##   pclass survived
## 1      1        1
## 2      1        1
## 3      1        0
## 4      1        0
## 5      1        0
```

```
## 6      1        1
##                                             name
## 1                    Allen, Miss. Elisabeth Walton
## 2                   Allison, Master. Hudson Trevor
## 3                     Allison, Miss. Helen Loraine
## 4             Allison, Mr. Hudson Joshua Creighton
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
## 6                               Anderson, Mr. Harry
##      sex     age sibsp parch ticket     fare
## 1 female 29.0000     0     0  24160 211.3375
## 2   male  0.9167     1     2 113781 151.5500
## 3 female  2.0000     1     2 113781 151.5500
## 4   male 30.0000     1     2 113781 151.5500
## 5 female 25.0000     1     2 113781 151.5500
## 6   male 48.0000     0     0  19952  26.5500
##     cabin embarked boat body
## 1      B5        S    2   NA
## 2 C22 C26        S   11   NA
## 3 C22 C26        S        NA
## 4 C22 C26        S       135
## 5 C22 C26        S        NA
## 6     E12        S    3   NA
##                          home.dest
## 1                    St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                    New York, NY
```

```
# Below code with display last few rows of the
# titanic CSV file.
tail(titanic)
```

```
##      pclass survived
## 1304      3        0
## 1305      3        0
## 1306      3        0
## 1307      3        0
## 1308      3        0
## 1309      3        0
##                            name    sex  age
## 1304     Yousseff, Mr. Gerious   male   NA
## 1305      Zabour, Miss. Hileni female 14.5
```

```
## 1306     Zabour, Miss. Thamine female   NA
## 1307 Zakarian, Mr. Mapriededer   male 26.5
## 1308       Zakarian, Mr. Ortin   male 27.0
## 1309        Zimmerman, Mr. Leo   male 29.0
##      sibsp parch ticket    fare cabin
## 1304     0     0   2627 14.4583
## 1305     1     0   2665 14.4542
## 1306     1     0   2665 14.4542
## 1307     0     0   2656  7.2250
## 1308     0     0   2670  7.2250
## 1309     0     0 315082  7.8750
##      embarked boat body home.dest
## 1304        C        NA
## 1305        C       328
## 1306        C        NA
## 1307        C       304
## 1308        C        NA
## 1309        S        NA
```

Charu's observations -

1. Not all data fields are available for all passangers, especially for the ones that did not survive.

2. Ticket numbers are in inconsistent format. Some of the ticket numbers are same for multiple rows, which indicates one ticket was issued per family.

3. Fare is probably not supposed to be four decimal points. It is most likely that passangers in 1912 paid only up to cents (2 decimal places).

4. Age is not properly formatted. A three byte integer should be enough for this analysis.

5. One row can contain more than one cabin number.

6. Most passangers who do not have a 'boat' assigned did not survive. 'Boat' most likely refers to a rescue boat.

Continued - Data types

Following data types are observed in this dataset -

pclass - Integer

survived - Integer (this will be later converted to character for ease of exploration)

name - character

sex - character

age - Numeric (to be converted to Integer later)

sibsp - Integer

parch - Integer

ticket - Character

fare - Numeric
cabin - character
embarked - character
boat - character
body - Integer
home.dest - character

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?

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

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.

Charu's response -

Age needs to be converted into integer format.

Also, the 'survived' variable needs to be converted to character for the purpose of filling.

Some other changes can be made, but they are not useful for the purpose of exploratory analysis.

```
# We will use a data.table function to change
# type of column age from numeric to integer.
titanic$age = as.integer(titanic$age)
titanic$survived <- as.character(titanic$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.

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

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")
```

Note: You need to add a written response here!

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.

Charu's Response -

The above figure (Survival and Passenger Age) does not do a good job of explaining the relationship between survival and pas-

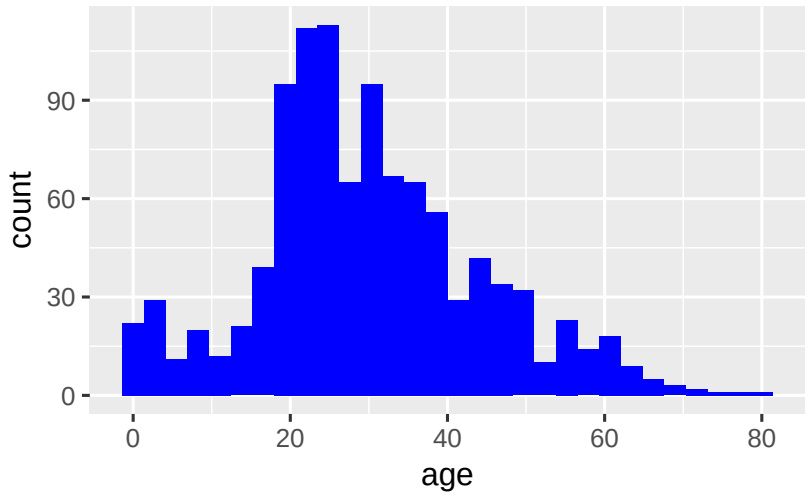

Figure 1: Age of Passenders Aboard the Titanic

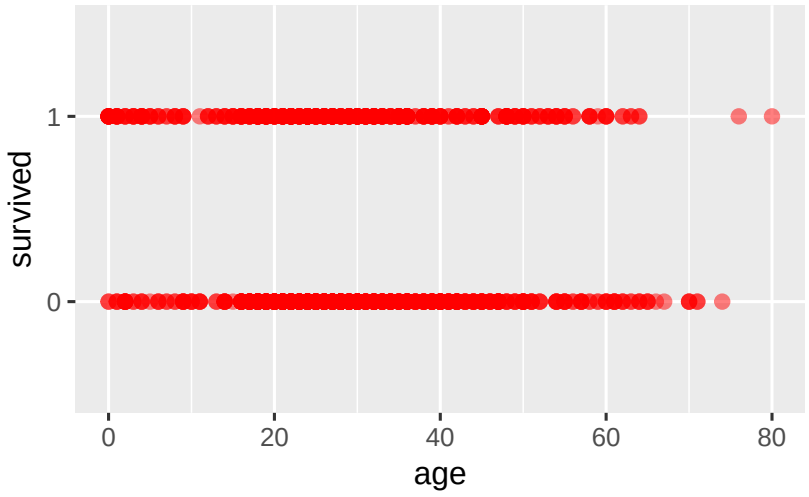

Figure 2: Survival and Passenger Age

sanger age. It is hard to understand and no conclusions can be drawn from the figure.

Also the Y axis has values between 0 and 1 which are never going to be valid.

Based on my understanding, I believe following figure will do a better job of explaining this relationship.

```
ggplot(titanic, aes(age)) + geom_bar(aes(fill = survived),
    position = "stack")
```

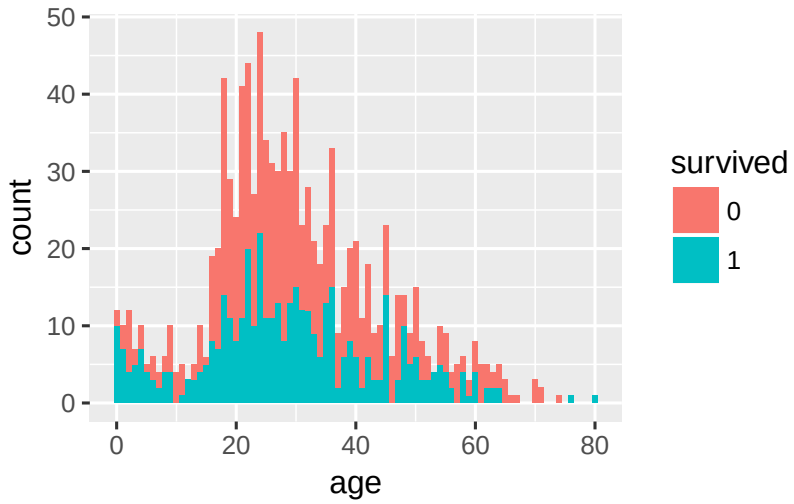

Figure 3: Survival and Passenger Age - Stacked Bar Plot

Charu's Response continued -

Based on the above figure, it is seen that generally for any age, except under age of 10, for most ages the pattern is the same. Number of people that did not survive is greater than those survived. However, under the age of 10, more children have survived. This may be due to the fact that children were rescued with priority, or as a courtesy. But for the most part, age of a passanger does not seem to have any effect on their survival

Do Additional Analysis

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.

Describe/explain what you find!

Charu's Response -

I am interested in finding out the relationship between the effect of sex and class of travel (pclass) on the survival of a passenger. It

was implied by some critics that the wealthy passengers travelling in first class had an unfair advantage of having better access to emergency supplies like life jackets, rescue boats etc. While those travelling in third class, mostly working class passengers, were claimed to have very limited access to these supplies, and it was implied that the working class passangers had a less survival rate compared to wealthy passangers. I would like to explore if this is true. Also, it is claimed that women were given priority for rescue, so their survival rate was better than men. We will also analyze if this is true.

We will plot the jitter plot. It is suitable since both pclass and sex are discrete variables.

```
ggplot(titanic, aes(sex, pclass, color = survived)) +
    geom_jitter(size = 0.6)
```

Figure 4: Effect of Passanger Class and Sex on Survival

Charu's response continued -

All of the claimed theories appear to be true.

1. In all passanger classes, female survival rate is better than males for a specific passenger class. This indicates that women were rescued with priority.

2. Chances of survival decrease as the passenger class decreases from first to third. This does indicate that first class wealthy passengers had better access to rescue services than working class passengers in third class.

3. Males in third class constitute the biggest share of the victims (not survived).

4. Almost all women in first class survived. The chances of survival are extremely high in this group.

What Next?

Note: You need to add a written response here!

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

Charu's Response -

Based on this exploratory analysis, further analysis can be done in following ways -

1. Is there a relationship between 'boats' and 'survived'? Boats variable is not explained on Kaggle. However, it may indicate the use of a rescue boat if the passenger used it. Did getting a boat help the passenger survive? Did anyone not survive after getting a boat?

2. Is there a relationship between getting a boat and class of travel? Did wealthier passengers have better access to boats?

3. Some outside data is needed for this specific analysis. How were the rooms structured for different classes? Did the location of first and second class allow easy access to decks and rescue boats? Did third class passengers have poor access to these because of their locations?

4. Did the first class and second class passengers actually pay more fare than third class?

5. For those travelling with families, did all family members survive,or did all family members die? Can a relationship be found where only children survived and parents died?

5. Use machine learning methods to calculate best possibilty of survival using age, sex, boat, class and travelling companions. Which group fares the best and why?