IMT 573 Lab: Exploring Data

Naga Soundari Balamurugan

October 2nd, 2018

Collaborators:

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 dependencies, including 'Rcpp' and 'tibble'.

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

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

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

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.

[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?

```r
#Top rows of the titanic table
head(titanic)
```

```
## # A tibble: 6 x 14
##   pclass survived name    sex      age sibsp
##    <int>    <int> <fct>   <fct>  <dbl> <int>
## 1      1        1 Allen,~ fema~ 29.0       0
## 2      1        1 Alliso~ male   0.917     1
## 3      1        0 Alliso~ fema~  2.00      1
## 4      1        0 Alliso~ male  30.0       1
## 5      1        0 Alliso~ fema~ 25.0       1
## 6      1        1 Anders~ male  48.0       0
## # ... with 8 more variables: parch <int>,
## #   ticket <fct>, fare <dbl>, cabin <fct>,
## #   embarked <fct>, boat <fct>, body <int>,
## #   home.dest <fct>
```

```r
#Bottom rows of the titanic table
tail(titanic)
```

```
## # A tibble: 6 x 14
##   pclass survived name     sex     age sibsp
##    <int>    <int> <fct>    <fct> <dbl> <int>
## 1      3        0 Youssef~ male   NA       0
## 2      3        0 Zabour,~ fema~  14.5     1
## 3      3        0 Zabour,~ fema~  NA       1
## 4      3        0 Zakaria~ male   26.5     0
## 5      3        0 Zakaria~ male   27.0     0
## 6      3        0 Zimmerm~ male   29.0     0
## # ... with 8 more variables: parch <int>,
## #   ticket <fct>, fare <dbl>, cabin <fct>,
## #   embarked <fct>, boat <fct>, body <int>,
## #   home.dest <fct>
```

```r
#Data type of all the columns
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 ...
```

```
#number of rows
nrow(titanic)
```

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

```
#number of columns
ncol(titanic)
```

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

```
#To find the number of unique entries in each column, unique could be used
unique(titanic$sex) #This lists male and female
```

```
## [1] female male
## Levels: female male
```

The head command provides us the top 6 rows of the titanic data. Similarly the tail command provides the bottom 6 rows of the titanic data. This helps us to confirm that the data is uniform and does not have any junk values at bottom rows.

The str command provides a list of all the column names with the corresponding data type. It also helps us to know the number of responses in each column. For example, the sex column has 2 levels i.e., Male and Female. This is really helpful to understand the number of levels in each column.

Finally the command nrow and ncol provides us with the number of rows and columns repectively. The titanic dataset has **1309** rows and **14** columns of different datatype.

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

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?

The column age is of type num(decimal) which can be denoted as integer. For the columnn survived, there are only two levels either yes or no and is denoted as 1(survived) and 0(not survived). It is of the type int which could be denoted as character as we are not going to make any arithmetic calculations on this column.

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

Hint: Consider how variables are measured and how that matches available data types in R. See: https://www.statmethods.net/input/datatypes.html

```
#Recasting the datatype of Age
titanic$age <- as.integer(titanic$age)
titanic$survived <- as.character(titanic$survived)
```

By using the str(titanic), we can reassure that the datatypes have been changed.

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")
```

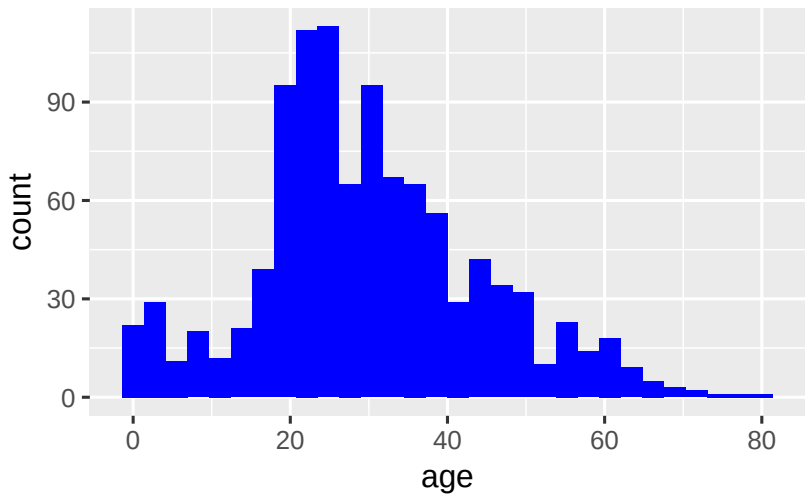

Figure 1: Age of Passenders Aboard the Titanic

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")
```

Note: You need to add a written response here!

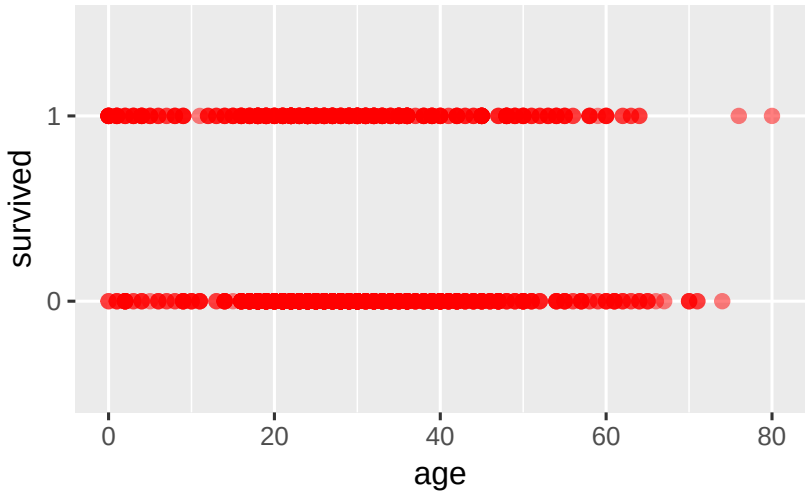

Figure 2: Survival and Passenger Age

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.

> This visualization does not clearly explain the purpose of number of people survived in each age group. Hence a barplot could be used to show the number of people survived in each age bracket.

```
#Filters only the data for the people who survived

survived <- titanic %>% filter(titanic$survived == '1')

#Histogram that displays the number of people survived in each age group
hist(survived$age, main = "No.of People survived vs Age")
```

> From the histogram, we can see that most of the people who survived are from the age group 20 to 30.

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!

> I would like to explore gender data and check if the female or male survived in more numbers.

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!

Figure 3: No of person survived vs Age