IMT 573: Module 2 Lab

Data Visualization

Jenny Skytta

Due: April 10, 2022

Collaborators: ***Independent work*** List collaborators here.

Objectives

In this module, we have focused on exploring data. Visualization is a great way to do this. Let's play around with visualization in this lab. You objective in this assignment is to create and reflect on different ways to visualize data. Think about what visuals you like and which enable you to tell compelling stories with data. And think about which charts you create might be misleading!

Instructions

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Open the `02_lab_viz.Rmd` and save a copy to your local directory. Supply your solutions to the assignment by editing `02_lab_viz.Rmd`.
2. First, replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and I encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`. When the PDF report is generated rename the knitted PDF file to `lab2_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup

In this lab you will need, at minimum, the following R packages.

```r
# Load standard libraries
library(tidyverse)
library(knitr) #loading this package to enable "kabble" function
```

In the demonstration for Module 2, we encountered data from the sinking of the RMS Titanic in the North Atlantic Ocean in the early morning of 15 April 1912. We will revisit this data.

```r
# Load titanic data
titanic_data <- read_csv("data/titanic.csv") # reading in the titanic.csv file and
# creating a variable called "titanic_data"
```

Recall, our two questions for exploration in the demonstration were:

- Who were the Titanic passengers and what characteristics did they have?
- What passenger characteristics or other factors seem to be associated with survival?

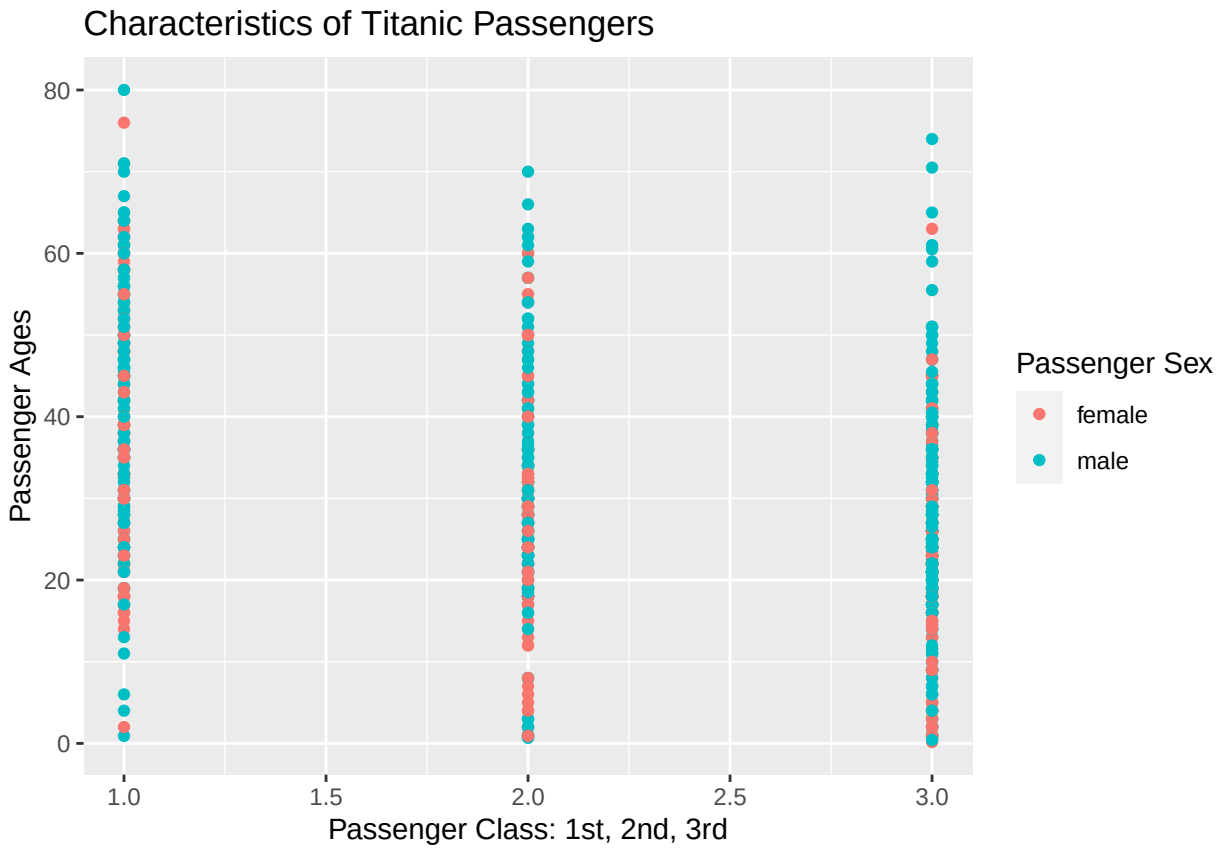
You job is to create a new visualization for each of these questions and comment on their ability to speak to these questions. Have fun and be creative!

Problem 1: Who were the Titanic passengers and what characteristics did they have?

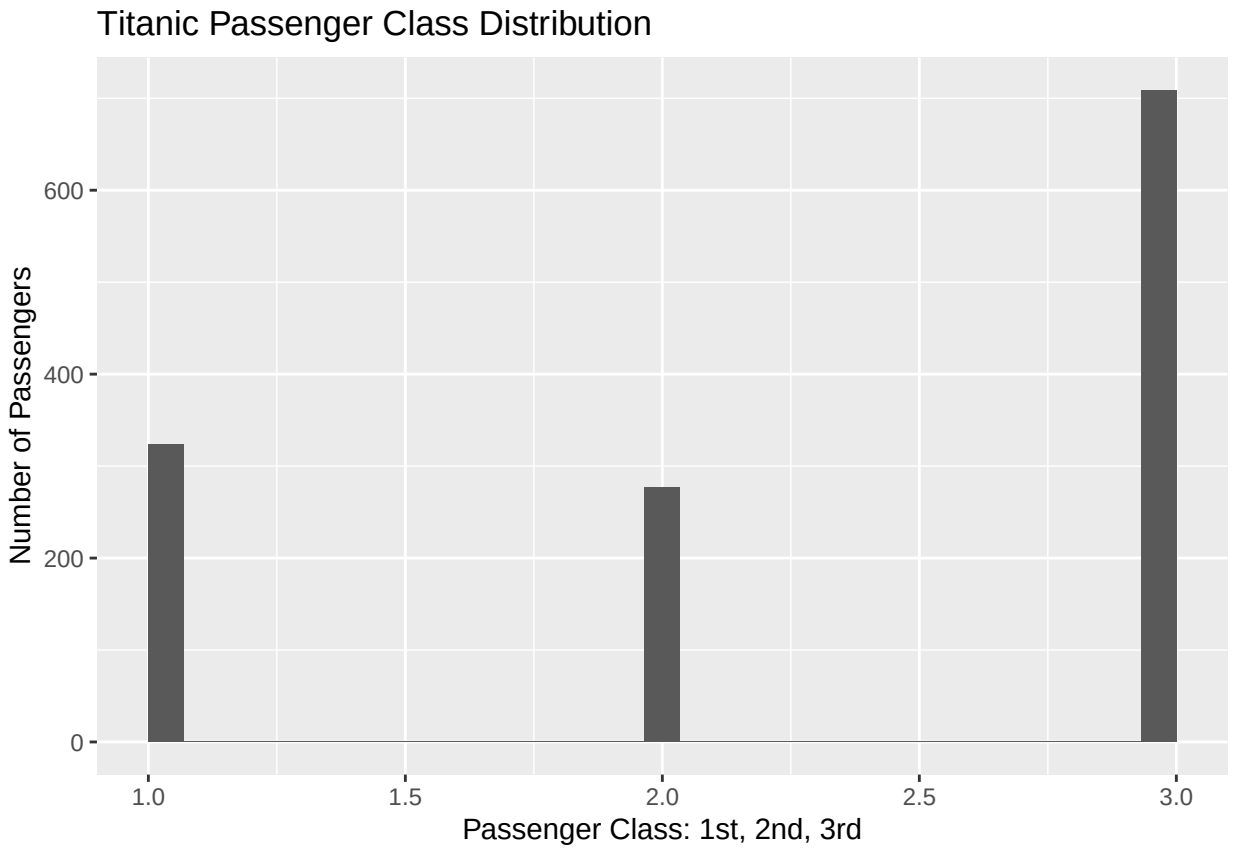
```r
summary_of_titanic <- titanic_data %>%
  select(pclass, survived, sex, age)

kable(head(summary_of_titanic))
```

pclass	survived	sex	age
1	1	female	29.0000
1	1	male	0.9167
1	0	female	2.0000
1	0	male	30.0000
1	0	female	25.0000
1	1	male	48.0000

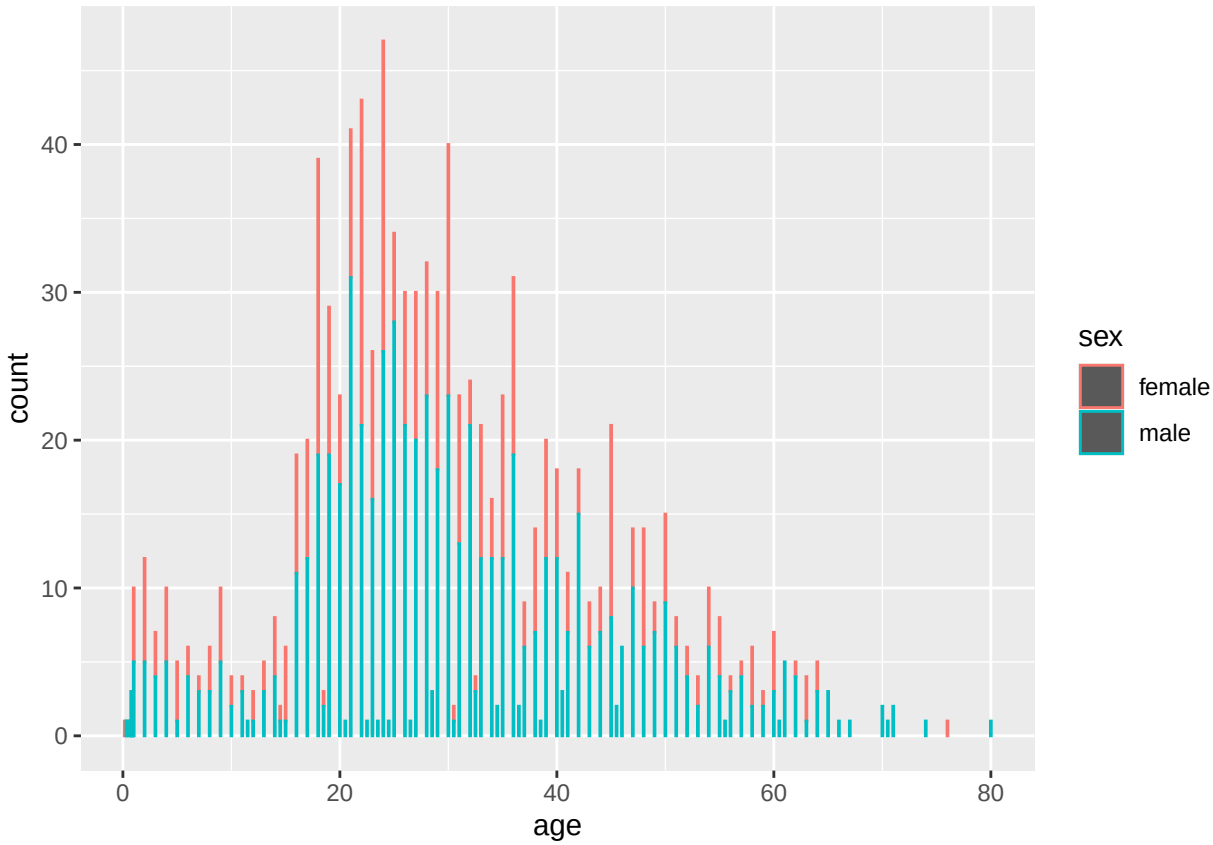
```r
ggplot(data = titanic_data, mapping = aes(x = pclass, y = age)) +
  geom_point(mapping = aes(color = sex)) +
  geom_smooth() +
  labs(
    title = "Characteristics of Titanic Passengers", # plot title
    x = "Passenger Class: 1st, 2nd, 3rd", # x-axis label
    y = "Passenger Ages", # y-axis label
    color = "Passenger Sex" # legend label for the "color" property
  )
```




```
ggplot(data = titanic_data, aes(pclass)) +
         geom_histogram() +
    labs(
    title = "Titanic Passenger Class Distribution", # plot title
    x = "Passenger Class: 1st, 2nd, 3rd", # x-axis label
    y = "Number of Passengers") # y-axis label
```




```
ggplot(data = titanic_data) +
  geom_bar(mapping = aes(x = age, colour = sex))
```



From our data, we see that the titanic passengers were described by their *class*, *survival status*, *sex*, and *age*. Data points pertaining to their accommodations and destination also are contained within the dataset. Most of the passengers were in 3rd class accommodations, accounting for more than double that of 1st class. In the IMT 573 *Module 2.9 EDA demonstration*, the histogram showed that the largest proportion of passengers were aged within their twenties.

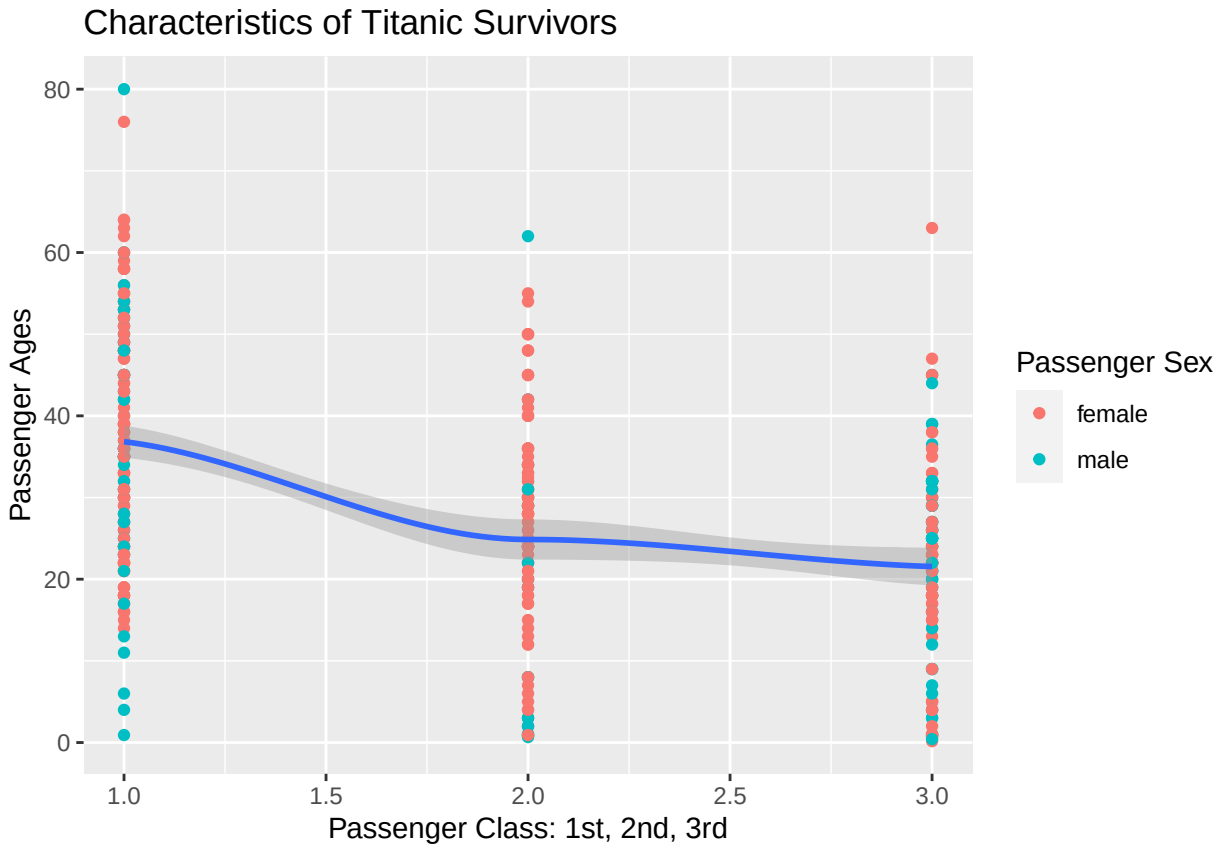
Problem 2: What passenger characteristics or other factors seem to be associated with survival?

```
Survivors_stats <- titanic_data %>% # Creating a tibble of data
  filter(survived == 1) %>%  # filtering for survival
   select(sex, pclass, age) # only selecting sex, class, and age for my data

kable(head(Survivors_stats)) #A small sampling table of the data informing the visualization
```

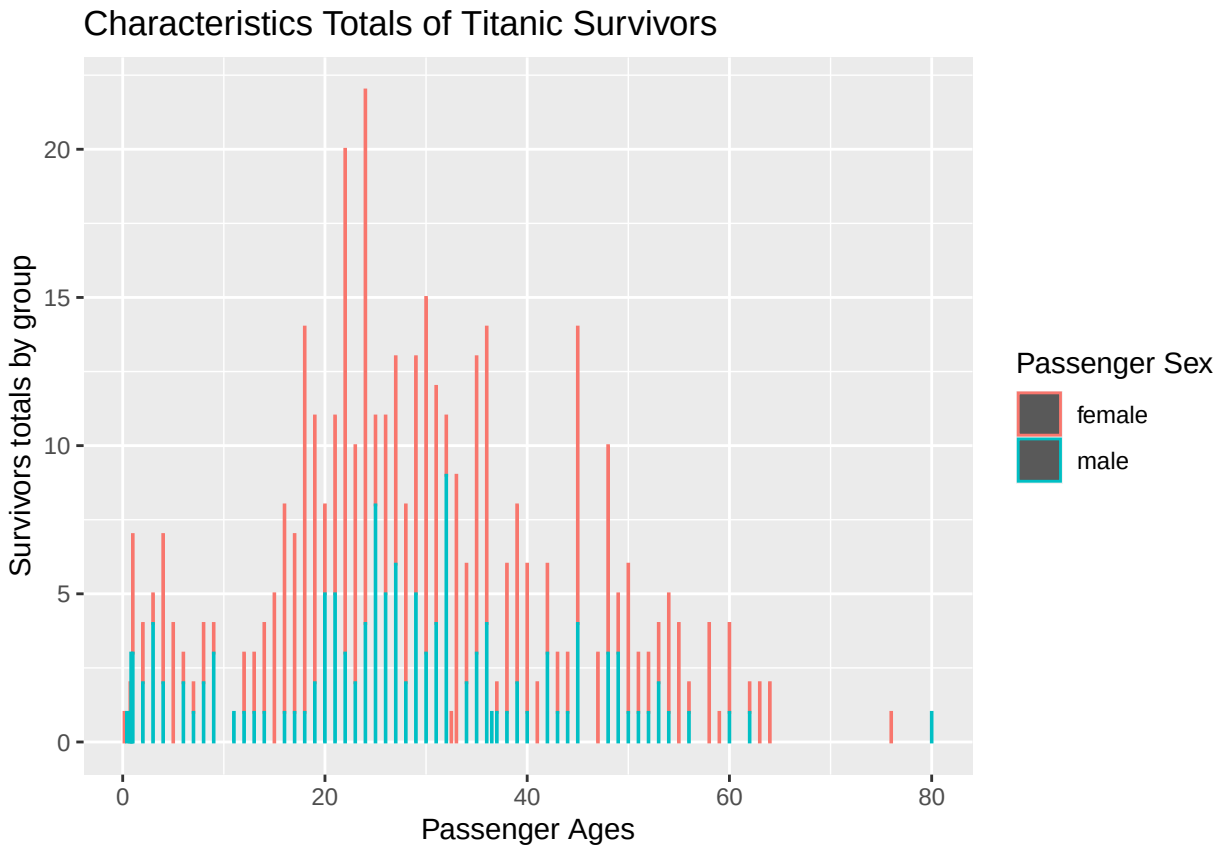
sex	pclass	age
female	1	29.0000
male	1	0.9167
male	1	48.0000
female	1	63.0000
female	1	53.0000
female	1	18.0000

```
ggplot(data = Survivors_stats, mapping = aes(x = pclass, y = age)) +
  geom_point(mapping = aes(color = sex)) +
  geom_smooth() +
    labs(
    title = "Characteristics of Titanic Survivors", # plot title
    x = "Passenger Class: 1st, 2nd, 3rd", # x-axis label
    y = "Passenger Ages", # y-axis label
    color = "Passenger Sex" # legend label for the "color" property
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#creating a small point plot that shows sex, age, and class intersections

ggplot(data = Survivors_stats) +
  geom_bar(mapping = aes(x = age, colour = sex)) +
    labs(
    title = "Characteristics Totals of Titanic Survivors", # plot title
    x = "Passenger Ages", # x-axis label
    y = "Survivors totals by group", # y-axis label
    color = "Passenger Sex" # legend label for the "color" property
  )
```



From the *visualization*, we see that survivors of the titanic were predominantly *first class*, *female*, and mostly below the age of *40*. Class and access to escape appear to have been correlated.

Citations:

Code written above is from the previous course IMT 511 which used the below text to support class scripts. https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover

[^titanic_ref] https://en.wikipedia.org/wiki/RMS_Titanic