Data Story Clinic

November 26, 2017

The Great Influenza of 1918

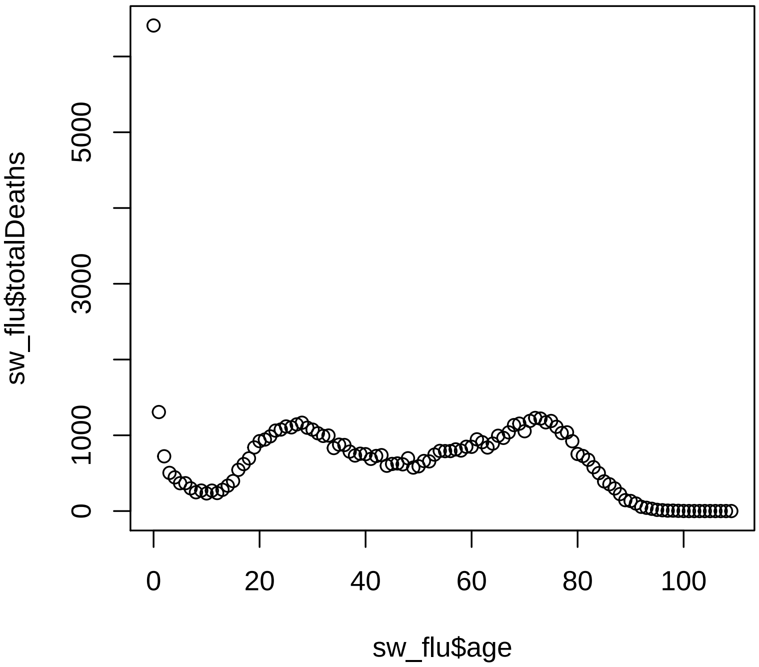
Exploratory data analysis is a highly creative activity. To give you a sense of how this is done, we will explore a couple of influenza mortality data sets. The first contains Swiss deaths from the 1918 flu pandemic. This outbreak was one of the deadliest in modern history and a fair amount of research has been done to understand why this was so.

Question 1 – Search for information on the 1918 pandemic. What books or articles did you find? Are there any data sets available? Also check the CDC (Center for Disease Control) website. Now create a list of your sources and why you chose them.

With a list of resources now compiled, let's load our Swiss data set and create a simple plot to visualize the distribution.

```
sw_flu <- read.csv("swissFlu.csv", header = TRUE,
    stringsAsFactors = FALSE)

plot(sw_flu$age, sw_flu$totalDeaths)
```

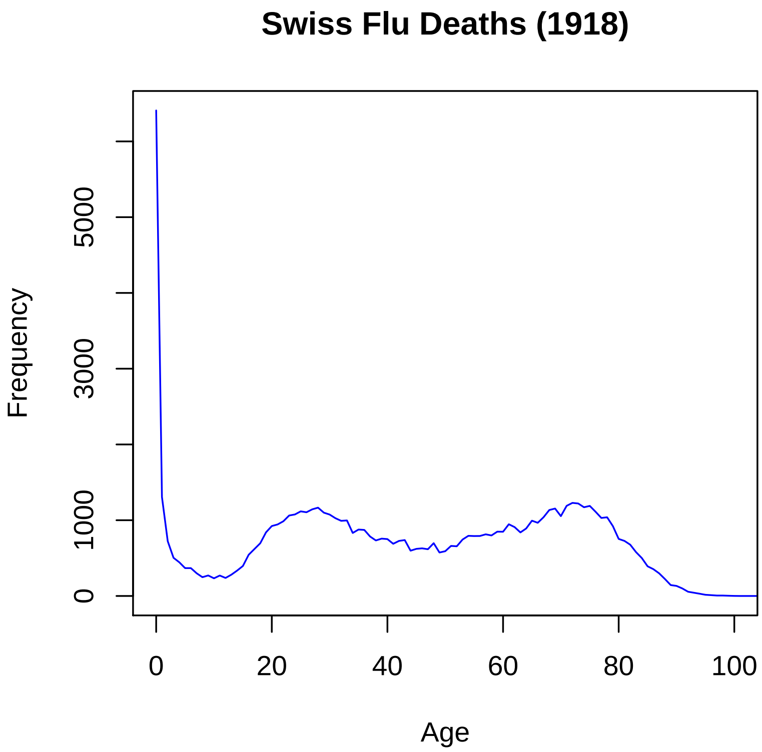


Although this initial plot provides some interesting information, let's spiff it up just a bit. Here are some of the arguments that can be added to the plot command.

- Type - the type of plot to be drawn (Points(default) or Lines)
- Main - the title of the plot
- Xlab, Ylab - the axis labels
- Xlim, Ylim - the range of values on each axis
- Col - color of symbols

```
# Create a blue line plot.

plot(sw_flu$age, sw_flu$totalDeaths, type = "l",
    ylab = "Frequency", xlab = "Age", main = "Swiss Flu Deaths (1918)",
    col = "blue", xlim = range(0:100))
```



Historians write that the 1918 pandemic was devastating. Millions died around the world. But how devastating was it? Sometimes, it's helpful to compare data sets to get a sense of scale. In this case, we'll compare the Swiss mortality data with influenza data found at flhealthcharts.com. In the past 20 years, the 1998 flu season stands out, having the highest total number of deaths.

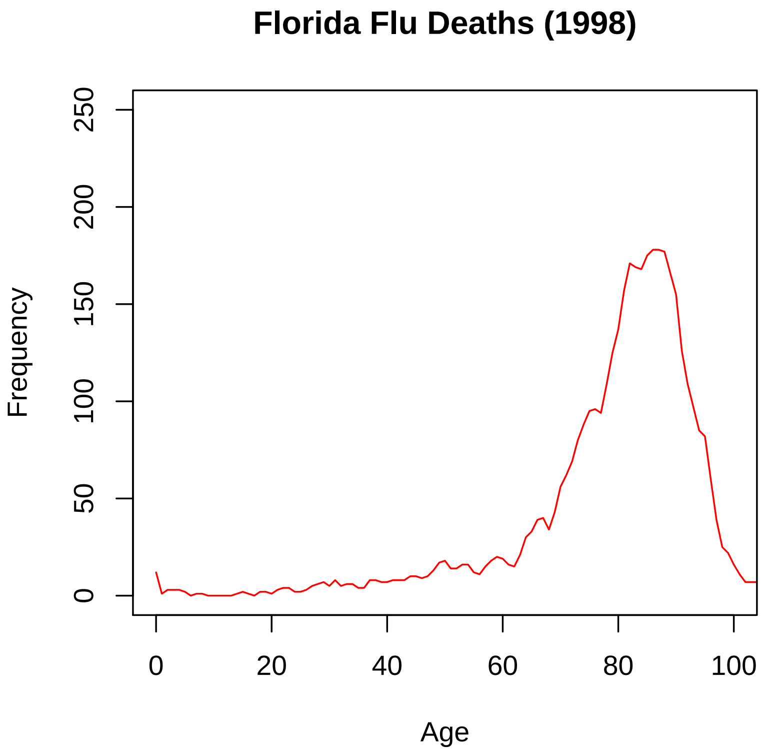
To save time, the 1998 Florida data set has already been created. So let's load that dataset now and plot it with the same arguments used for the Swiss plot. But instead of blue, we will create a red line.

Question 2 – When conducting exploratory data analysis (EDA), it is important that you pay close attention to the shape of the data. How does it present visually? Take some time to carefully examine this graph and then describe what you see in words. Is the data skewed to one side or the other? Does it have a mound in the center? Does anything stand out or appear odd, given what you know about influenza?

```
fl_flu <- read.csv("floridaFlu.csv", header = TRUE,
    stringsAsFactors = FALSE)

# Create a red line plot.

plot(fl_flu$age, fl_flu$totalDeaths, type = "l",
    ylab = "Frequency", xlab = "Age", main = "Florida Flu Deaths (1998)",
    col = "red", xlim = range(0:100), ylim = range(0:250))
```



To take a closer look, let's stack the two graphs.

Question 3 – the Florida influenza dataset comes from fl-healthcharts.com. Find the Influenza and Pneumonia Deaths page by Googling it. Given what you see on this page, define the steps a data scientist must take to get a count of deaths by age for the 1998 flu season. Data does not always come pre-formatted!

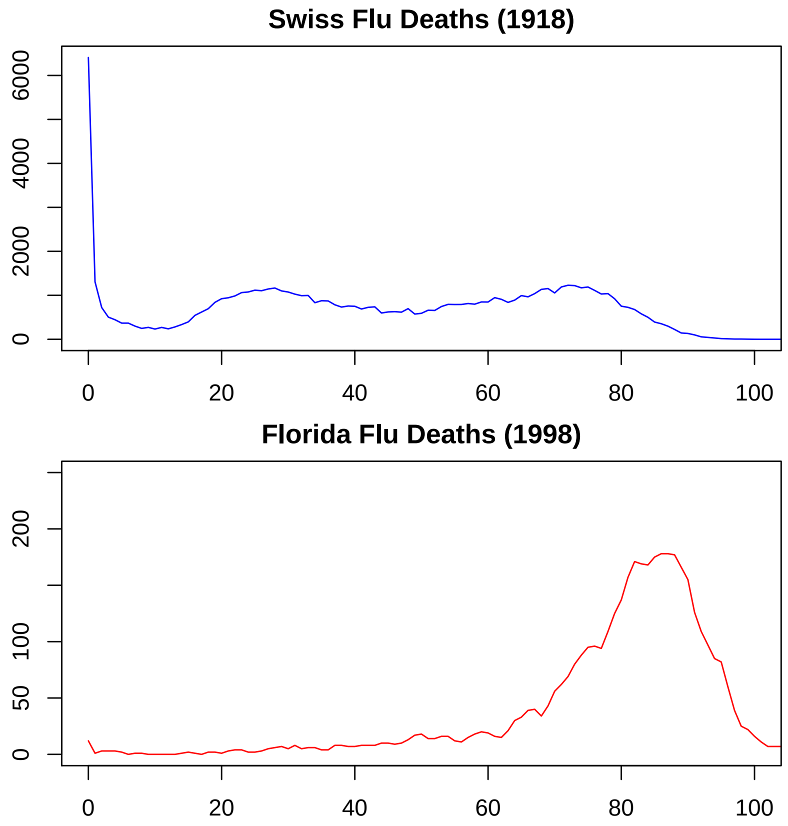
```
opar <- par(no.readonly = TRUE)

par(mar = c(2, 2, 2, 2))

# Two rows and one column of plots in the
# graph.
par(mfrow = c(2, 1))

plot(sw_flu$age, sw_flu$totalDeaths, type = "l",
    ylab = "Frequency", xlab = "Age", main = "Swiss Flu Deaths (1918)",
    col = "blue", xlim = range(0:100))
```

```r
plot(fl_flu$age, fl_flu$totalDeaths, type = "l",
    ylab = "Frequency", xlab = "Age", main = "Florida Flu Deaths (1998)",
    col = "red", xlim = range(0:100), ylim = range(0:250))
```



```r
par(opar)
```

At this point, you can begin to write the narrative to your data story. Pay close attention to the scale of the 1918 tragedy in comparison to the Florida mortality numbers for 1998. How many deaths would Florida have experienced if the mortality rate was equivalent to 1918? How might a data scientist calculate that?

Question 4 – What do you see when you compare these two graphs? Start to create a data story narrative by writing down everything that you find significant while comparing these two.

You will also want to consider the spread of the two distributions. Did the 1918 pandemic impact the same age groups as Florida's flu outbreak in 1998? What would calculating the mean or median age at death tell you? Would those numbers differ between the two distributions? And if so, how?

In order to calculate a central tendency (mean or median), we must create frequency tables for each dataframe. As is often the case, the data is not in a format that allows us to use the *mean()* or *median()* functions. Running either of these functions on the totalDeaths column is meaningless as we want to calulate the mean or median age at death. In the code that follows, we transform the data into a format (vector) that is usable by either of these functions as well as *hist()* and *boxplot()*.

```
for (i in 1:nrow(fl_flu)) {
    tmp <- rep(fl_flu$age[i], fl_flu$totalDeaths[i])

    if (i == 1) {
        fl_freq <- tmp
    } else {
        fl_freq <- c(fl_freq, tmp)
    }
}

for (i in 1:nrow(sw_flu)) {
    tmp <- rep(sw_flu$age[i], sw_flu$totalDeaths[i])

    if (i == 1) {
        sw_freq <- tmp
    } else {
        sw_freq <- c(sw_freq, tmp)
    }
}
```

Question 5 – Can you describe what this code is doing? Create a Help document that explains this code to a data scientist who needs to understand its function.

Many times as a data scientist you will need to understand and/or modify code written by other R programmers. In fact, a good way to master R is by analyzing code line-by-line.

With the two vectors now created, we will calculate the mean for each.

```
mean(sw_freq)
```

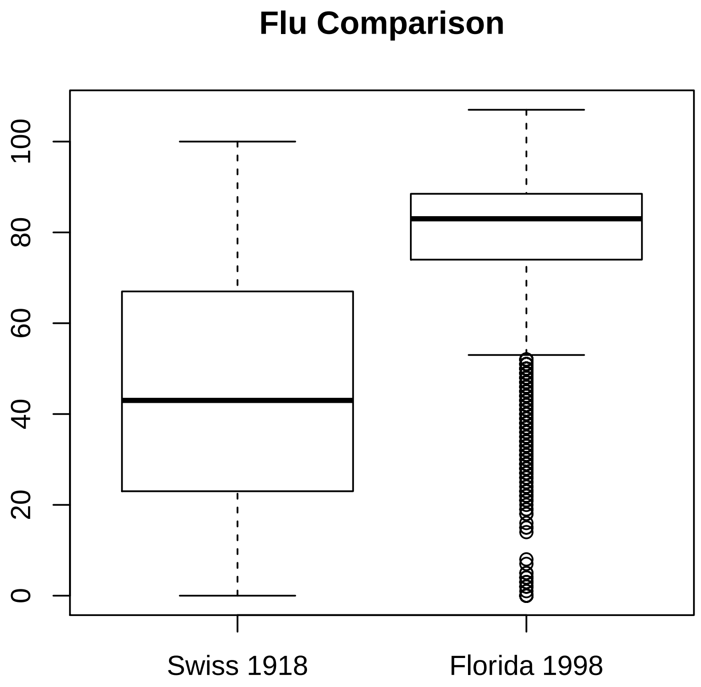
```
## [1] 43.21765
```

```
mean(fl_freq)
```

```
## [1] 79.06936
```

Additionally, let’s create a boxplot to *see* the difference. Remember: a boxplot shows the distribution of the data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.

```
boxplot(sw_freq, fl_freq, main = "Flu Comparison",
    names = c("Swiss 1918", "Florida 1998"))
```



With this information in-hand, you should now be able to produce a compelling data story of the 1918 pandemic. Or, if you wish, you might decide to investigate this tragedy from another angle. For example, what was its economic impact? Did it result in the loss of GDP (gross domestic product)? How would a data scientist calculate that?