06 Introduction to data visualization

Soyoung Park

Pusan National University
Department of Statistics

The US murders data

Looking at the numbers and character strings that define a dataset is rarely useful. To convince yourself, print and stare at the data table:

```
library(dslabs)
data(murders)
head(murders)
```

```
##       state abb region population total
## 1   Alabama  AL  South    4779736   135
## 2    Alaska  AK   West     710231    19
## 3   Arizona  AZ   West    6392017   232
## 4  Arkansas  AR  South    2915918    93
## 5 California  CA   West   37253956  1257
## 6  Colorado  CO   West    5029196    65
```

The US murders data

What do you learn from staring at this table?

1) How quickly can you determine which states have the largest populations?

2) Which states have the smallest?

3) How large is a typical state?

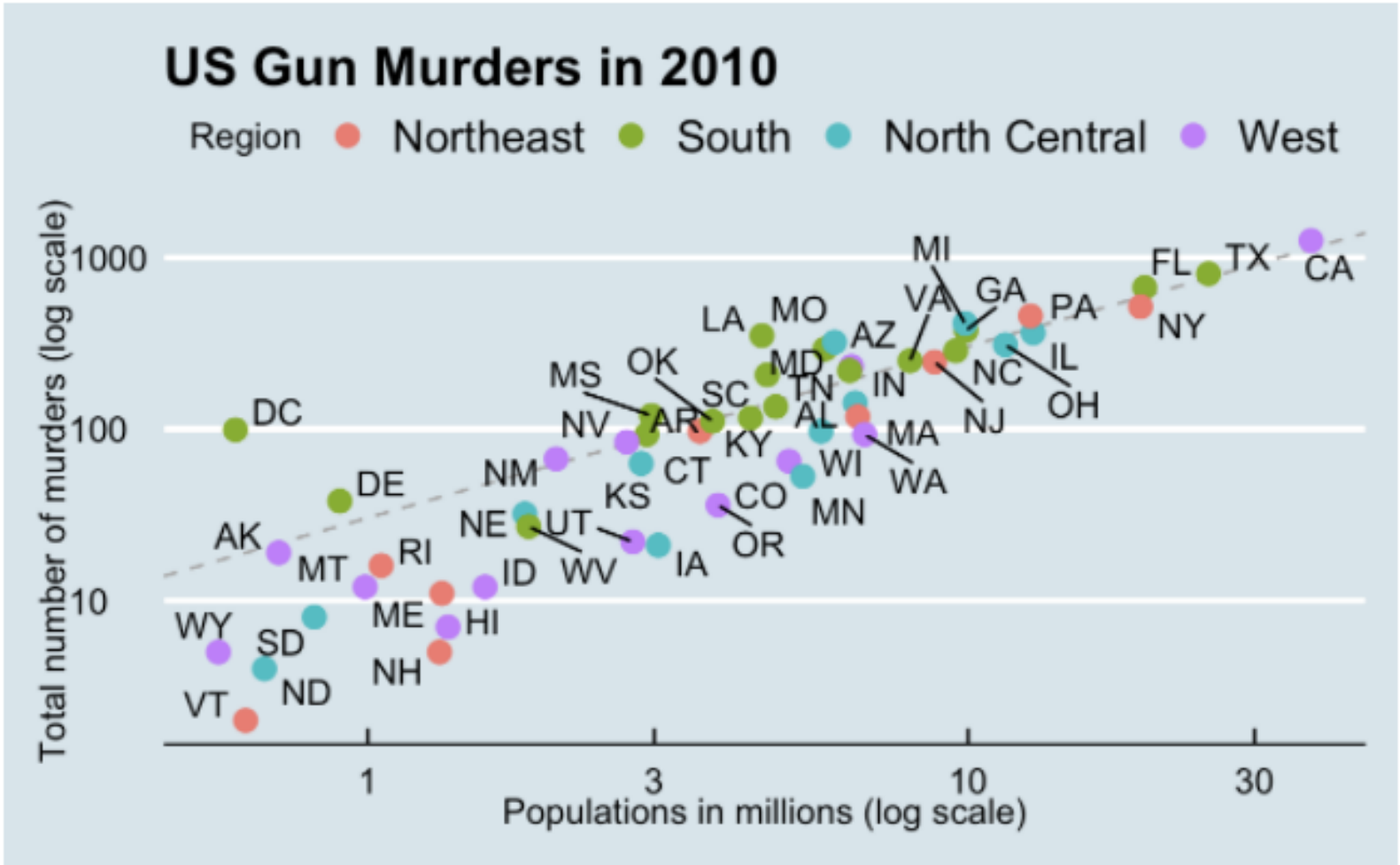
4) Is there a relationship between population size and total murders?

5) How do murder rates vary across regions of the country?

It is quite difficult to extract this information just by looking at the numbers.

The US murders data

In contrast, the answer to all the questions above are readily available from examining this plot:



Data visualization

Data visualization provides a powerful way to communicate a data-driven finding. In some cases, the visualization is so convincing that no follow-up analysis is required.

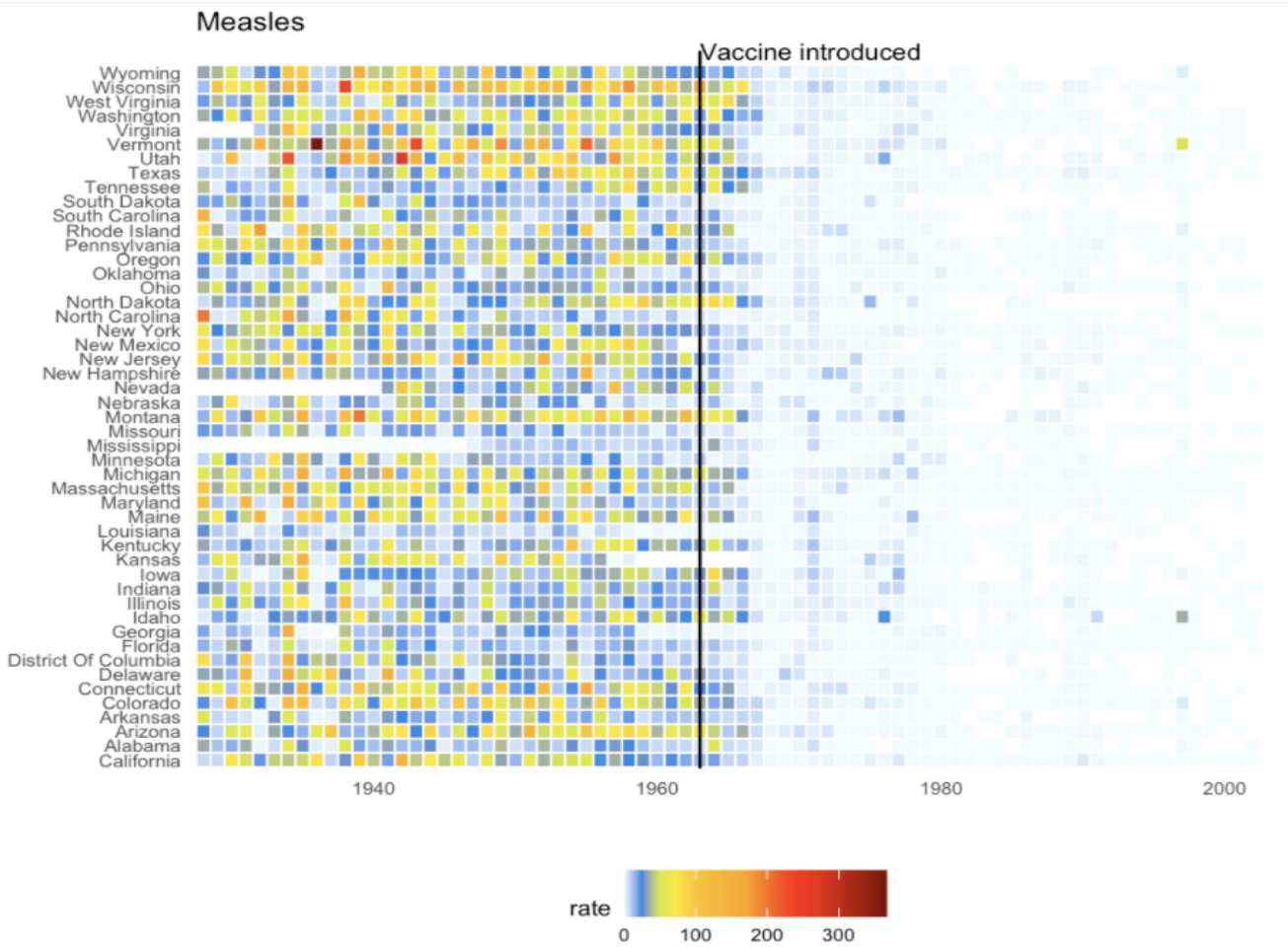
The growing availability of informative datasets and software tools has led to increased reliance on data visualizations across many industries, academia, and government.

A particularly effective example is a Wall Street Journal article[1] showing data related to the impact of vaccines on battling infectious diseases.

[1] http://graphics.wsj.com/infectious-diseases-and-vaccines/?mc_cid=711ddeb86e

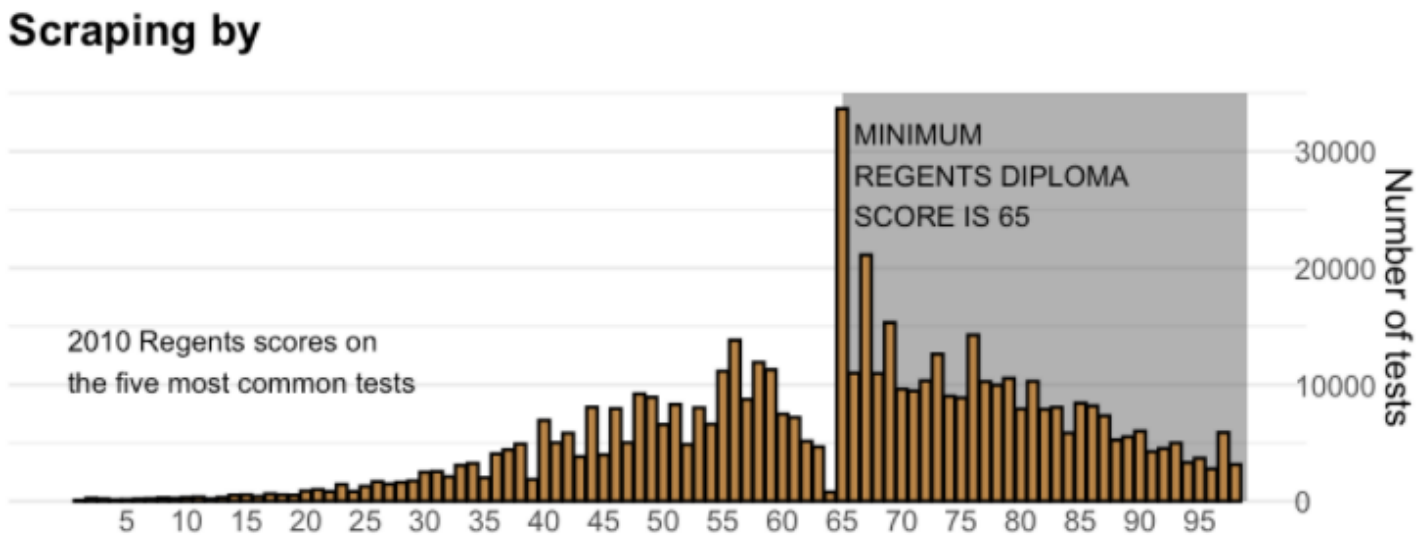
Data visualization

One of the graphs shows measles cases by US state through the years with a vertical line demonstrating when the vaccine was introduced.



Data visualization

Another striking example comes from a New York Times chart, which summarizes scores from the NYC Regents Exams. The distribution of the test scores forces us to notice something somewhat problematic:



The most common test score is the minimum passing grade, with very few scores just below the threshold. This unexpected result is consistent with students close to passing having their scores bumped up.

Data visualization

This is an example of how data visualization can lead to discoveries which would otherwise be missed if we simply subjected the data to a battery of data analysis tools or procedures.

Data visualization is the strongest tool of what we call *exploratory data analysis* (EDA).

Many widely used data analysis tools were initiated by discoveries made via EDA. EDA is perhaps the most important part of data analysis, yet it is one that is often overlooked.

We will use the **ggplot2** package to code. To learn the very basics, we will start with a somewhat artificial example