Data visualization, joining data tables and reshaping data

Overview

Review of topics related to the homework

Data visualization continued

Additional practice joining data tables: dealing with duplicate keys

Bonus ggplot features

Reshaping data (if there is time)

Questions?

Review of concepts from the homework

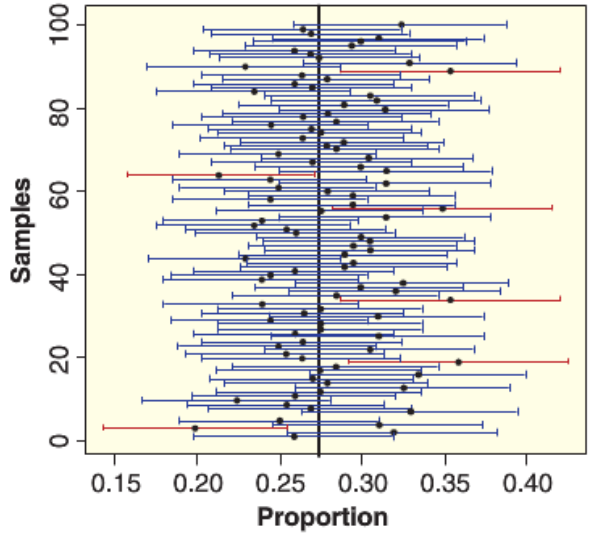
1. Getting confidence intervals for different confidence levels
   - E.g., confidence levels beyond 95%

2. Significance level α vs. p-value

3. Robustness

1. Confidence intervals for any confidence level

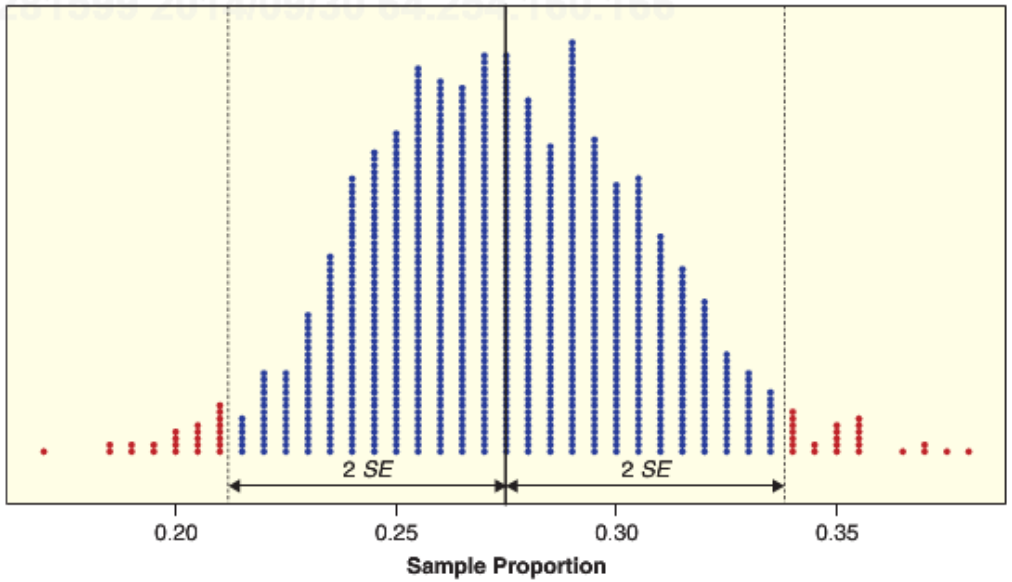
A **confidence interval** is an interval computed by a method that will contain the ***parameter*** a specified percent of times

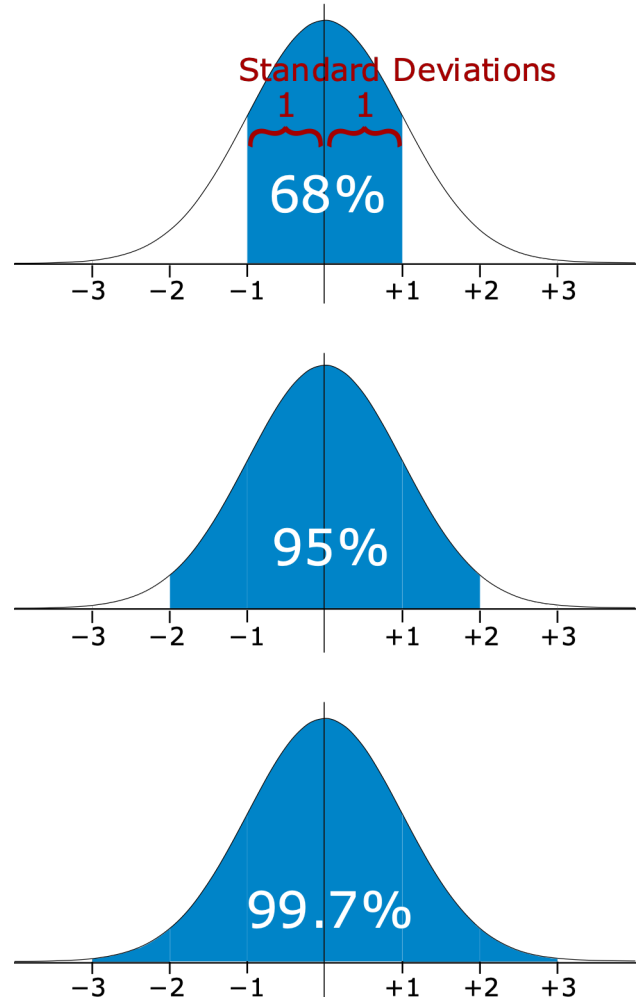


The **confidence level** is the percent of all intervals that contain the parameter

1. Confidence intervals for any confidence level

Recall we can use the bootstrap to get an estimate of the standard error SE*





We can then get a 95% confidence interval using:

$$\bar{x} \pm 2 \cdot SE^*$$

1. Confidence intervals for any confidence level

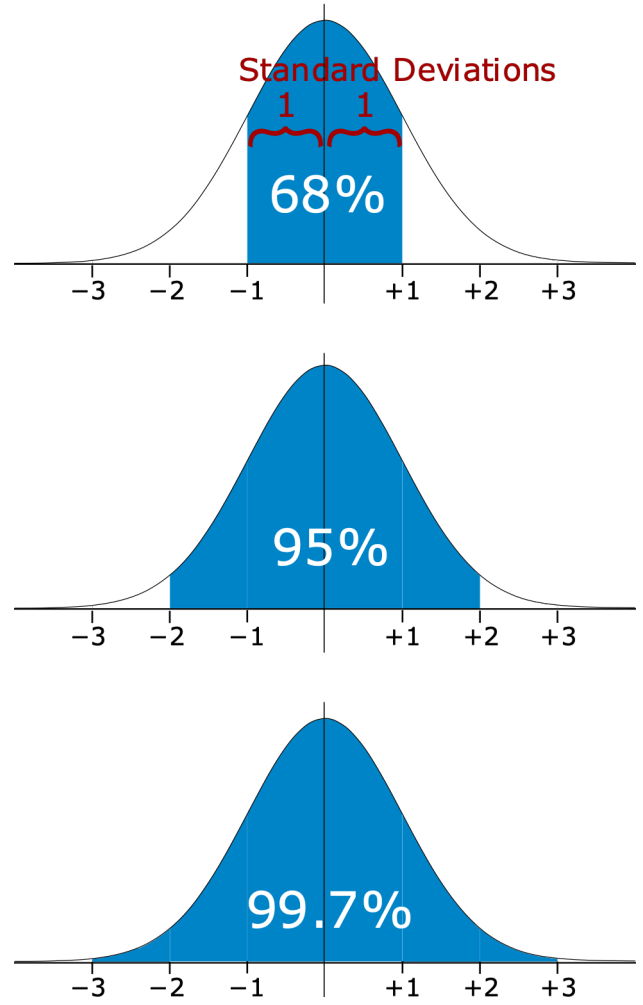
Q: What if we want a confidence interval for a different confidence level?

A: To do this we need different quantiles from the standard normal distribution.

- Say we want a 90% confidence interval we can get the z* value such that 90% of the mass of a normal is within ± z*
- In R we can do this using the qnorm(a) function
  - *a* the amount of area under a normal curve: $Pr(Z < z^*) = a$

We can then get a confidence interval using:

$$\bar{x} \pm z^* \cdot SE^*$$

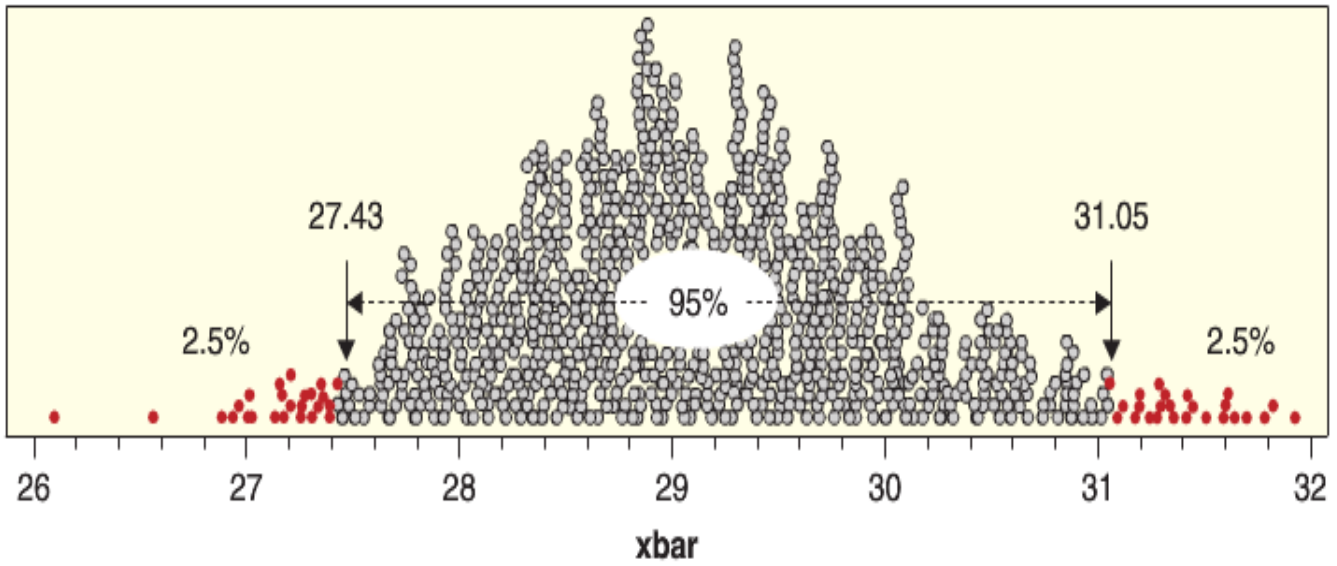


The bootstrap percentile method

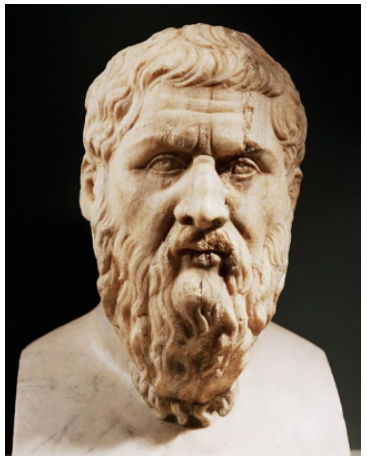
We can also get confidence intervals for any confidence levels using the percentiles of the bootstrap distribution

In R we can do this using the quantile() function:

```
> quantile(bootstrap_dist, c(.005, .995)))      # 99% CI
```



2. Significance level α vs. p-value

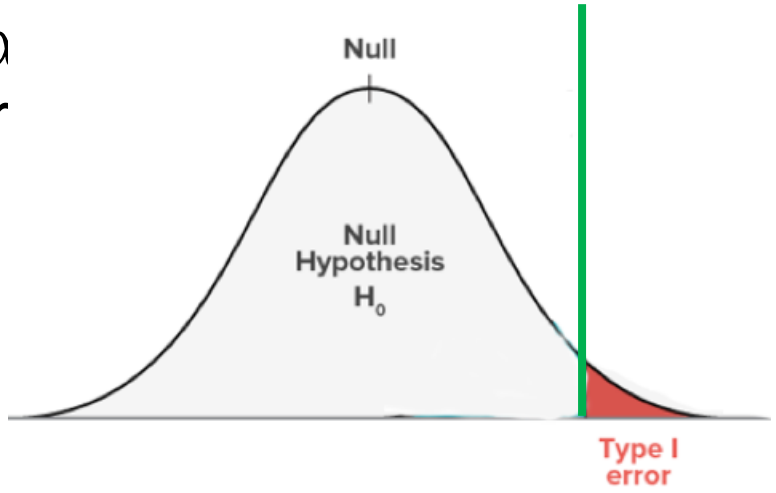


The significance level α is a value we set prior to running a hypothesis tests

- Life wisdom: If you are going to make a bet with a nihilist, you'd better agree to the rules first!
- $\alpha = 0.05$

In the Neyman-Pearson accept/reject H_0 paradigm, α is the proportion of times we will make a type I error

- i.e, proportion of time we incorrectly reject the null hypothesis
- i.e. the proportion of times Gorgias loses the bet even though he was correct

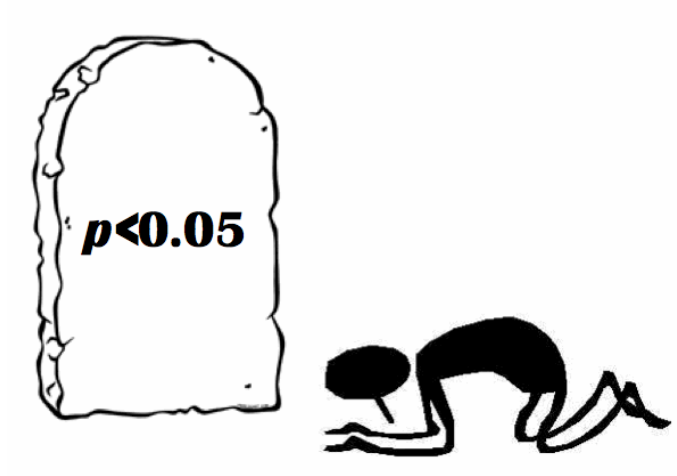


2. Significance level α vs. p-value

A **p-value** is the probability, of obtaining a statistic as (or more) extreme than the observed sample *if the null hypothesis was true*

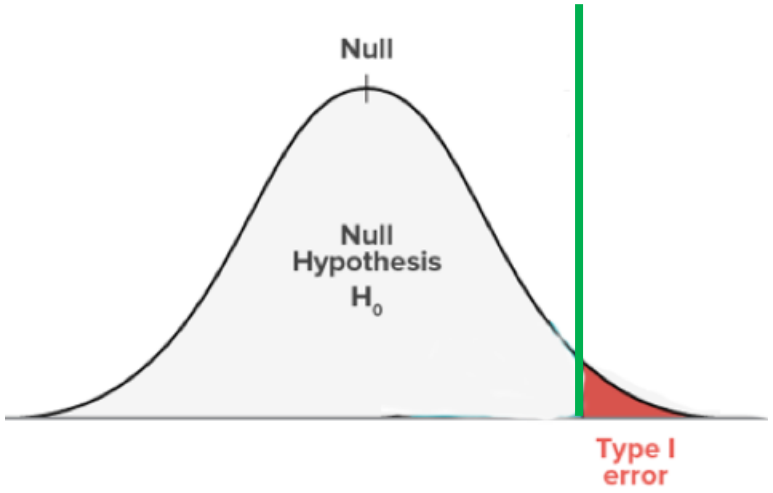
- i.e., the probability that we would get a statistic as extreme as our observed statistic from the null distribution

$$\Pr(\text{STAT} \geq \text{observed statistic} \mid H_0 = \text{True})$$



In the Neyman-Pearson accept/reject H_0 paradigm, if our p-value is less than the α level we will reject the null hypothesis.

- This will ensure we will only make α% type I errors
  - E.g., if α = 0.05 we will only have type I errors 5% of the time



3. Robustness

Statistical procedures are **robust** if they perform well under a wide range of conditions (i.e., a range of underlying probability distributions).

Example: mathematical derivation of the t-test assumes that the underlying data is normally distributed, but the t-test still works well when n > 30 for many other distributions.

- i.e., the t-test will still give the correct type I error rate even if the data is not coming from a normal distribution.

In homework 4 I was using the term to mean do we get similar results under slightly different conditions

- paperback books instead of hard cover books, etc.

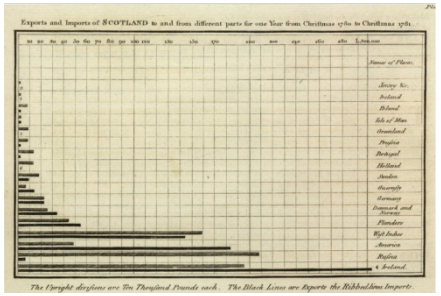
Data visualization

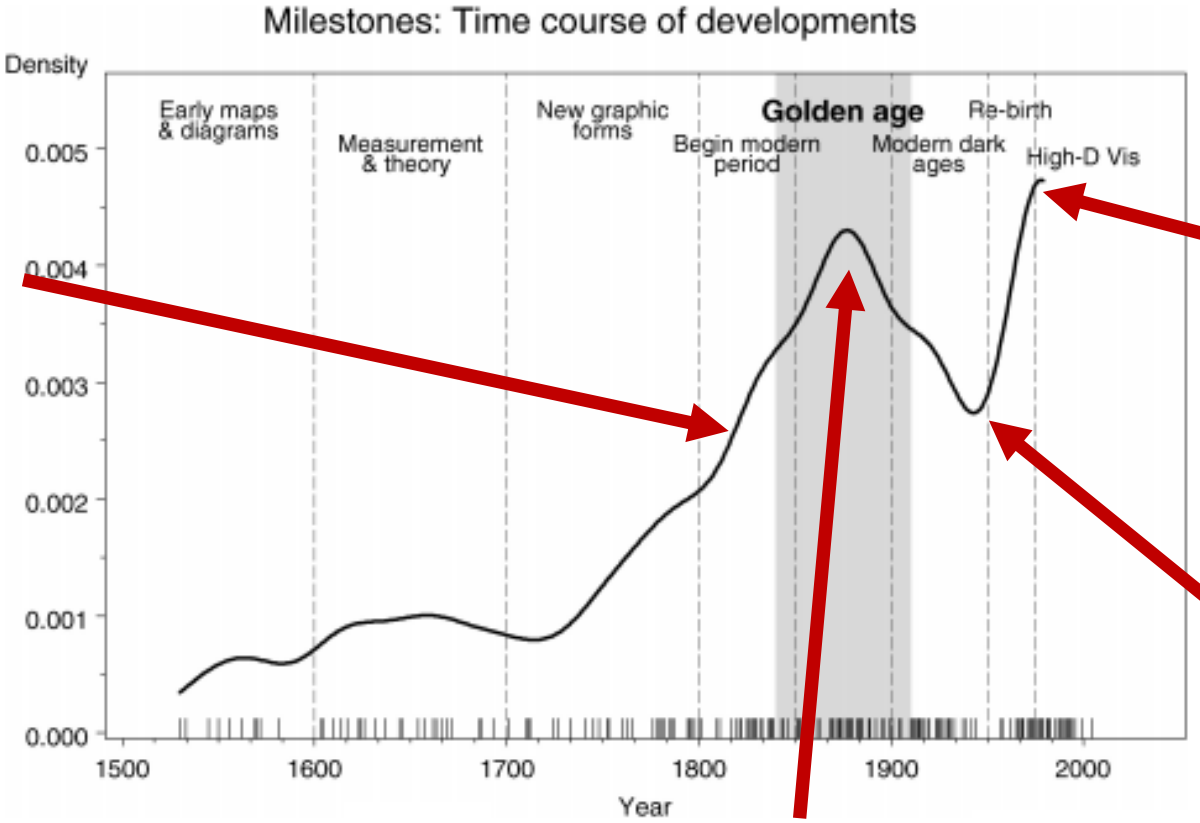
*Statistical projections which **speak to the senses without fatiguing the mind**, possess the advantage of fixing the attention on a great number of important facts.*

—Alexander von Humboldt, 1811

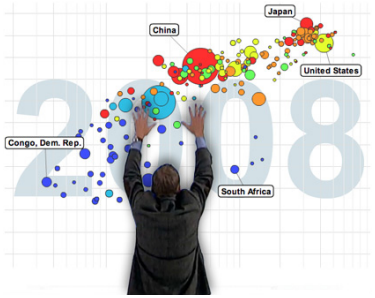
Review: A very brief history of data visualization

Playfair





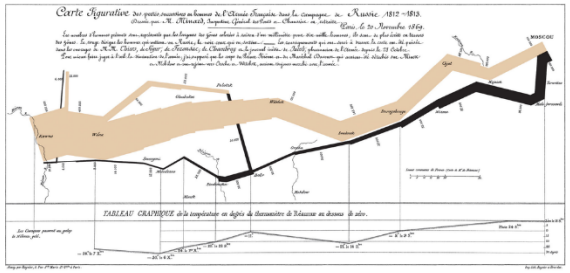
Info graphics



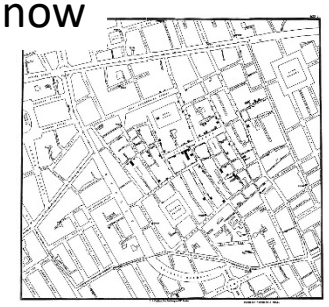
Gapminder

Math

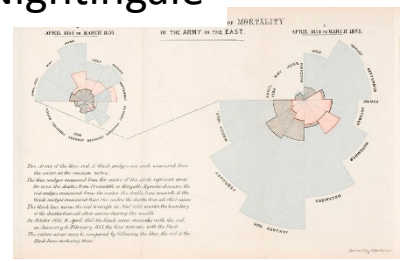
Minard



Snow



Nightingale

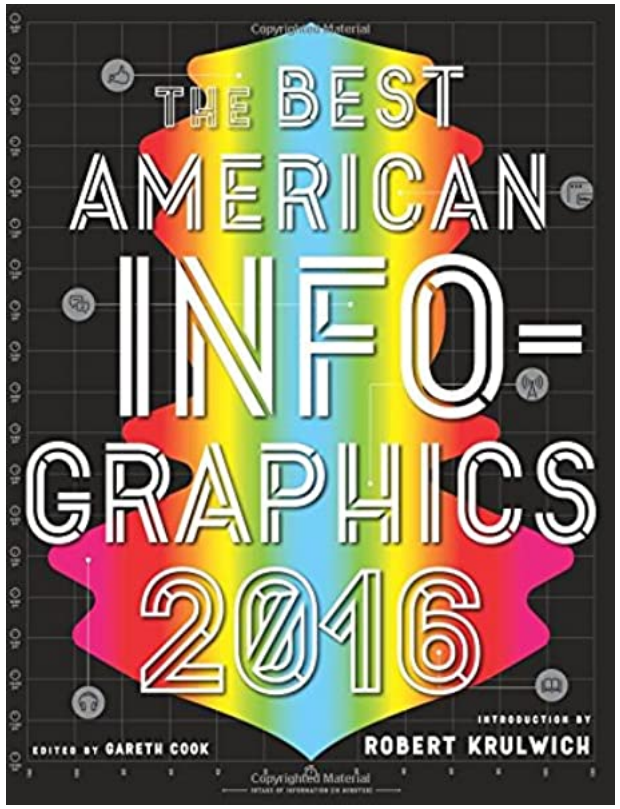


Survey question 1

Find an interesting data visualization on the web:

1. Write down the URL link to the image
2. Explain why you think it is interesting

Brief class share on Thursday



Let’s share with each other interesting visualizations you found

Go around alphabetically by first name

Share link to visualization in chat window

Describe why you found it interesting for 1 minute

Did anyone see anything particularly interesting?

One of my favorites…

NYTimes: TheUpshot

What 2,000 Calories Looks Like

Chipotle

Shake shack

Ruth's Chris Steakhouse

Potbelly

Sonic

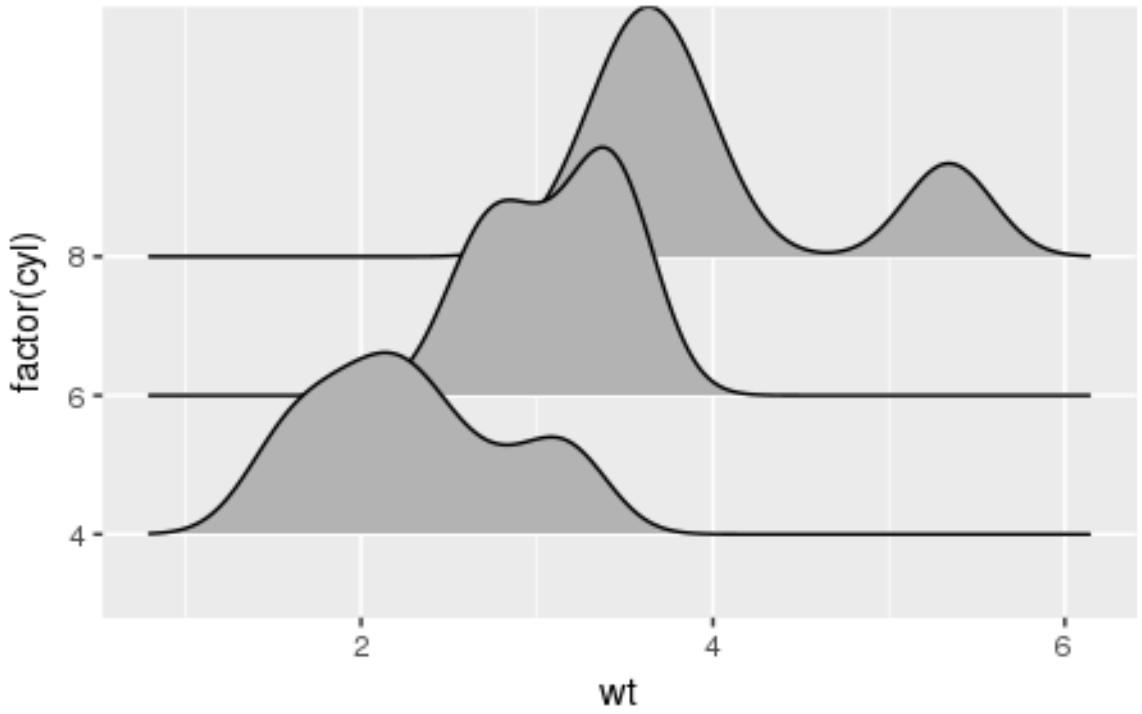
Subway

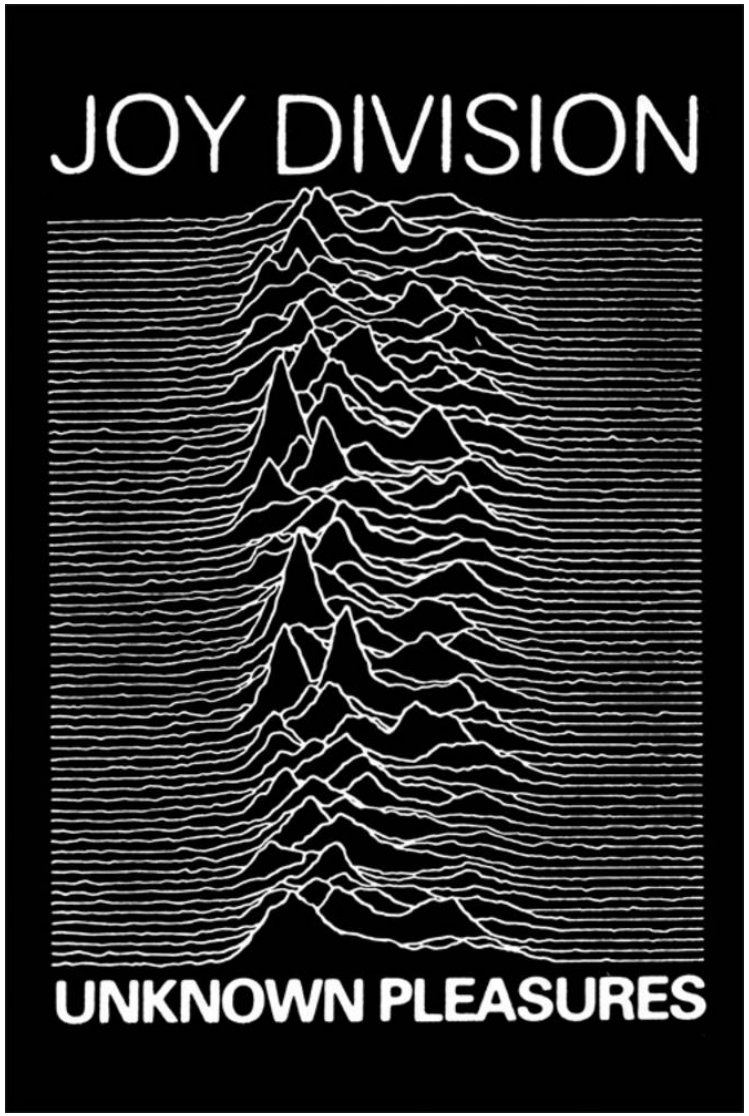
iHop

At home

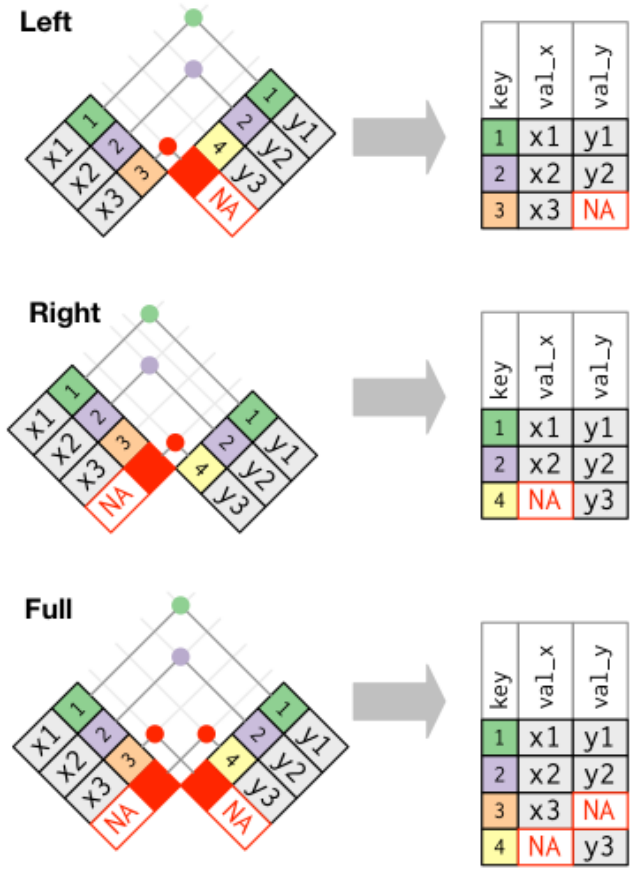
Violin and Joy plots

Any ideas why they are called joy plots?





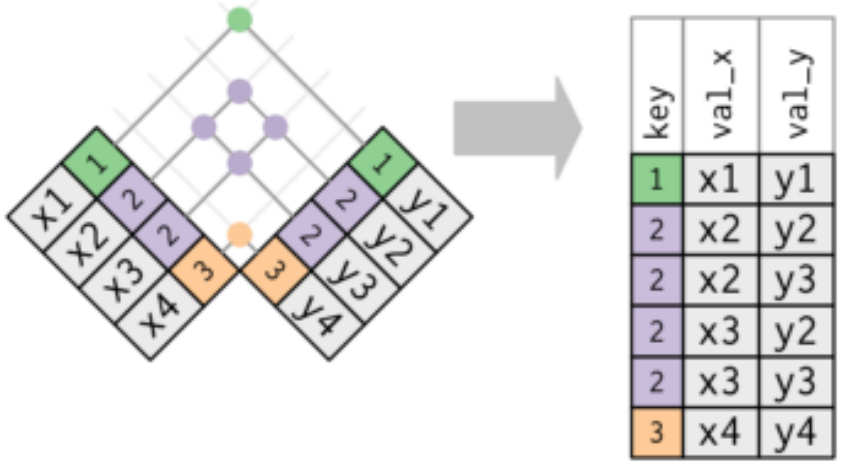
Joining data frames



Duplicate keys

If both tables have duplicate keys you get all possible combinations of joined values (Cartesian product).

- **This is usually an error!**



key	val_x	val_y
1	x1	y1
2	x2	y2
2	x2	y3
2	x3	y2
2	x3	y3
3	x4	y4

Always check the output after you join a table because even if there is not a syntax error you might not get the table you are expecting!

- You can check how many rows a data frame has using the nrow() function

Let’s try a real example in R…