Journey into the 2d world!
MVE080/MMG640 Lecture 2

Sebastian Persson
sebpe@chalmers.se

November 2, 2022

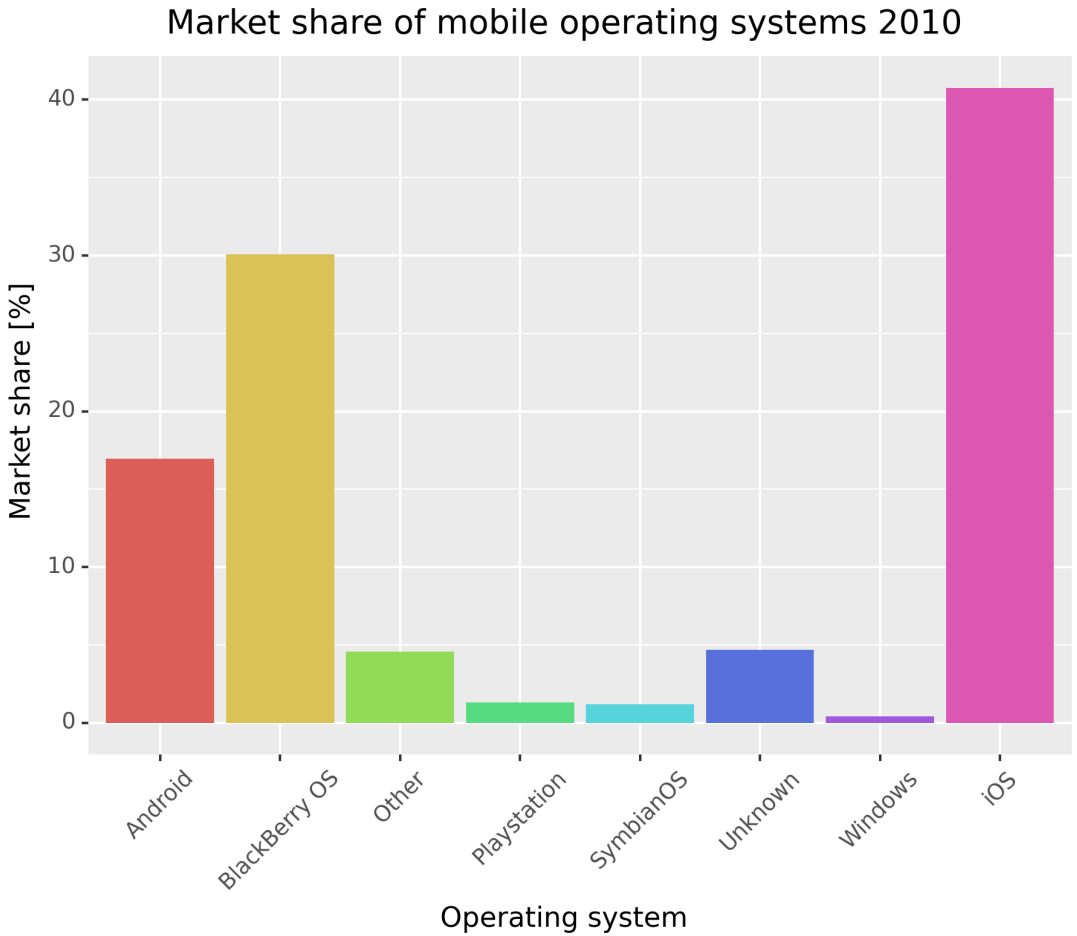
Discussion forum - Campuswire

- Campuswire - *https://campuswire.com/p/G327E9A1D*
  - Password 4516
- Forum to ask question which can be answered by other students, or myself

How do we visualise amounts?

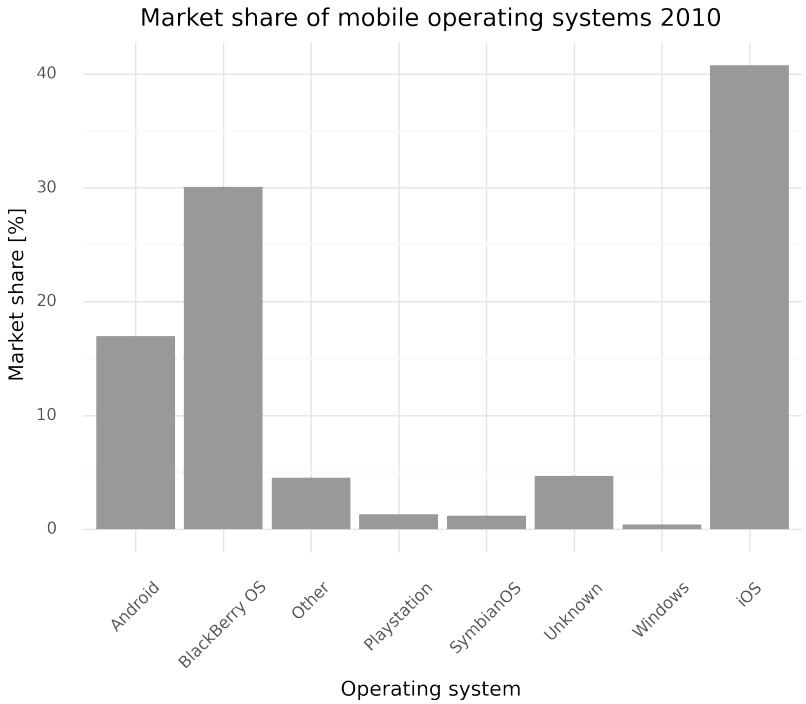
From last lecture...

- Why is this plot unreadable?
  - For the next two minutes write down your response



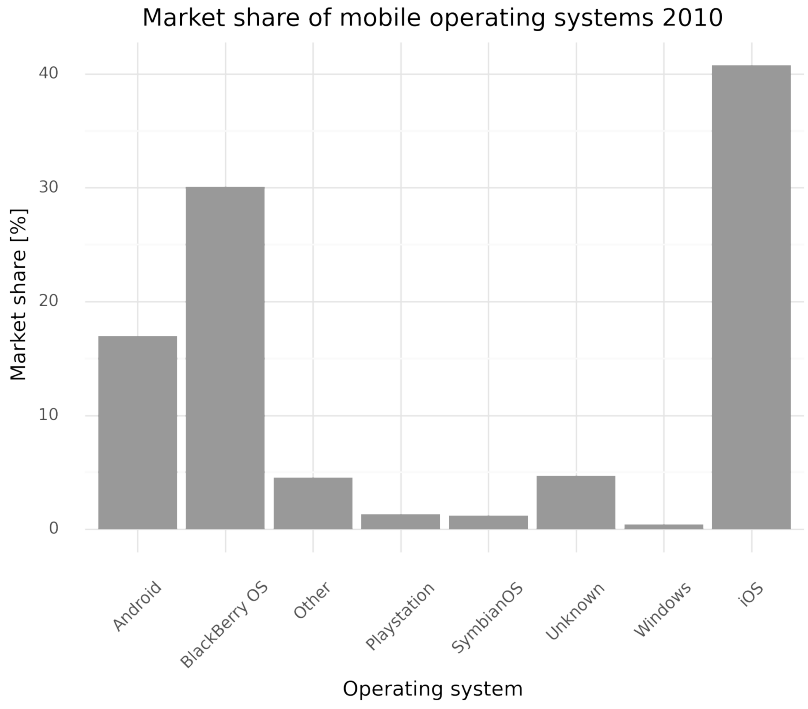
Visual noise makes it harder to interpret a plot

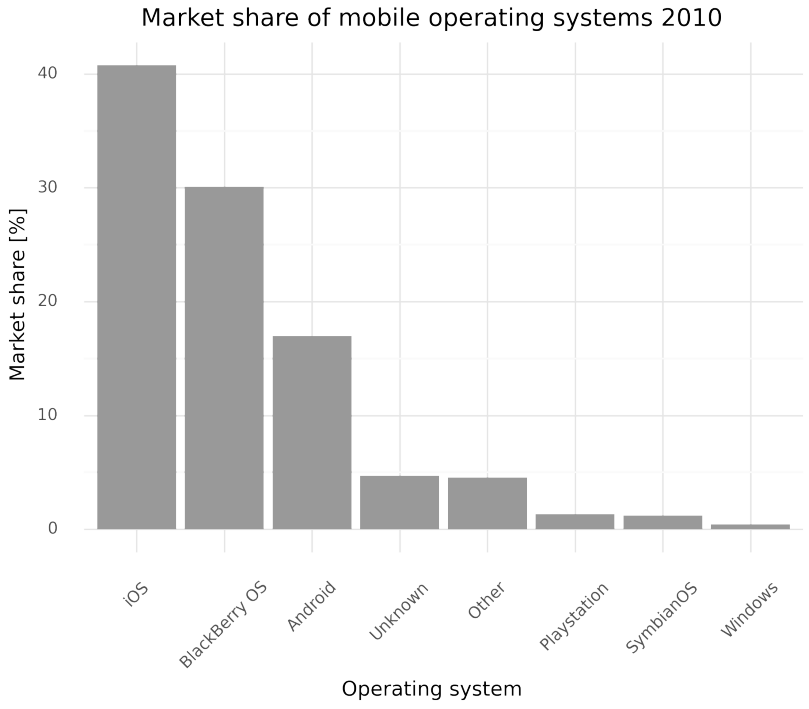
- Rainbow palette not ”unbiased” $\rightarrow$ remove



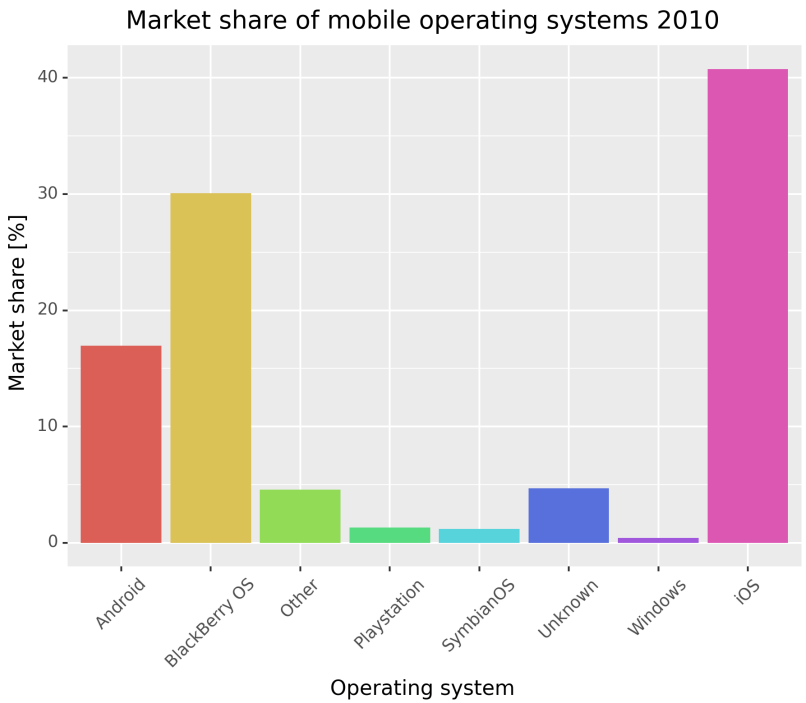
Visual noise makes it harder to interpret a plot

- Rainbow palette not ”unbiased” → remove
- We easier perceive sorted information



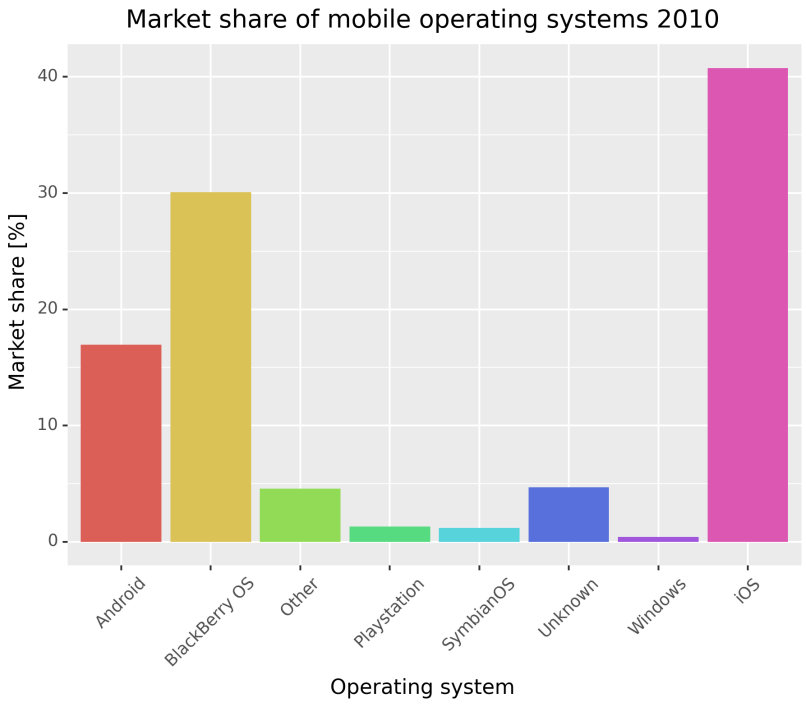


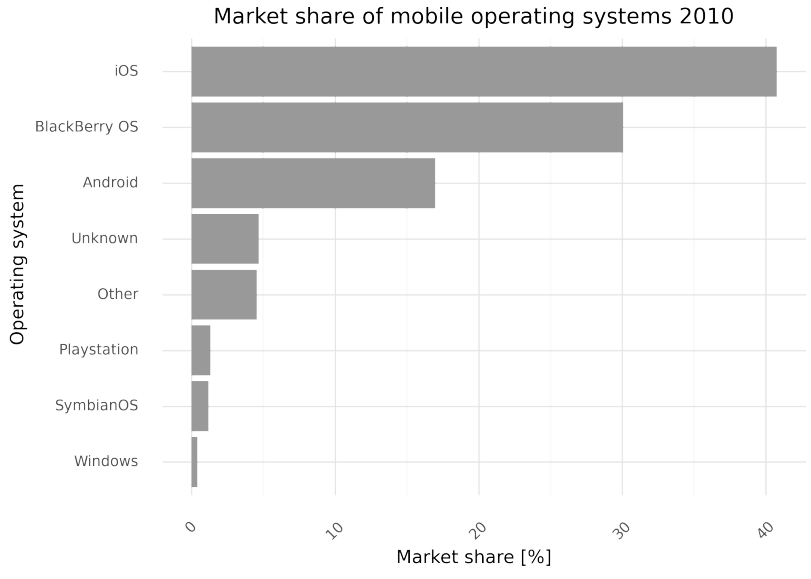
Axis should be easy to read



Axis should be easy to read

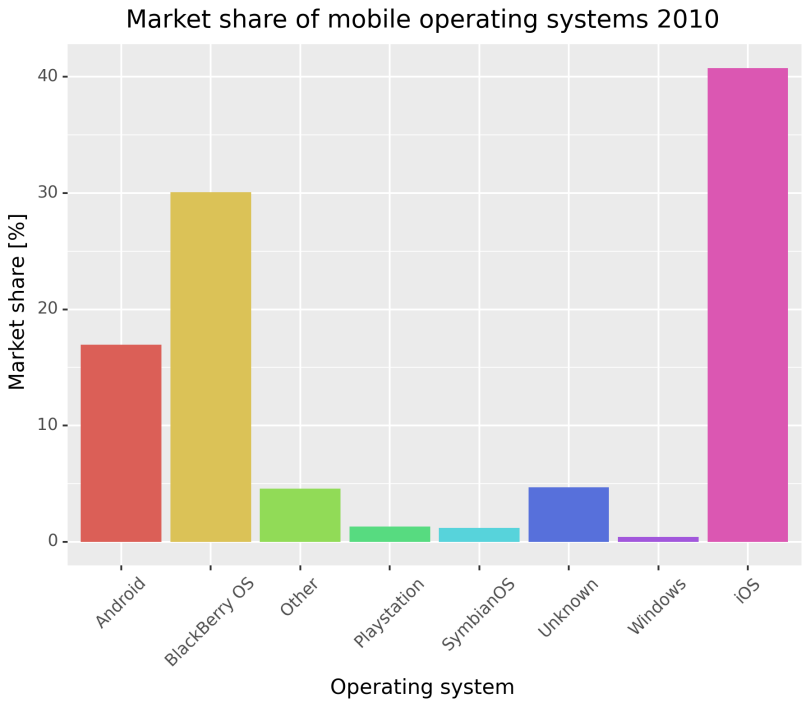
- 45-degree text awkward → flip the graph

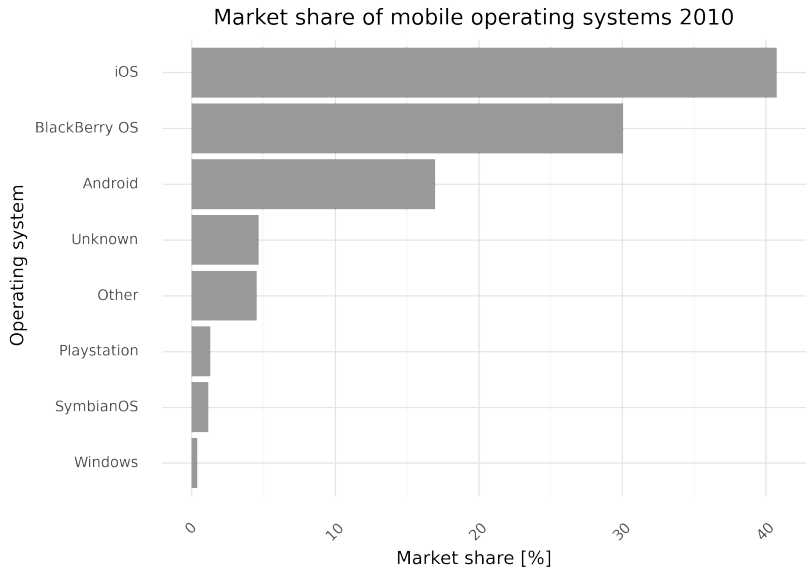




Axis should be easy to read

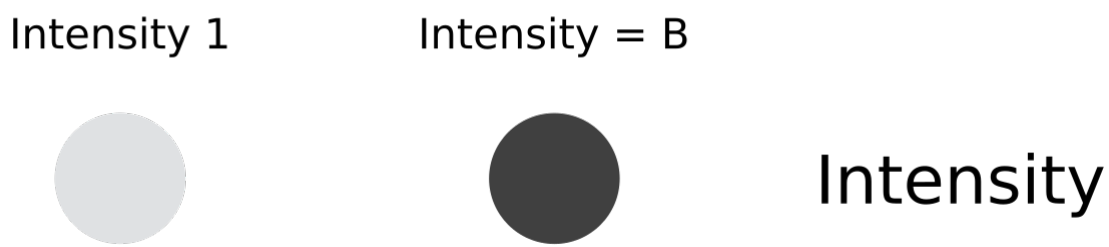
- 45-degree text awkward → flip the graph
- Only sort bars when categories do not have inherent ordering





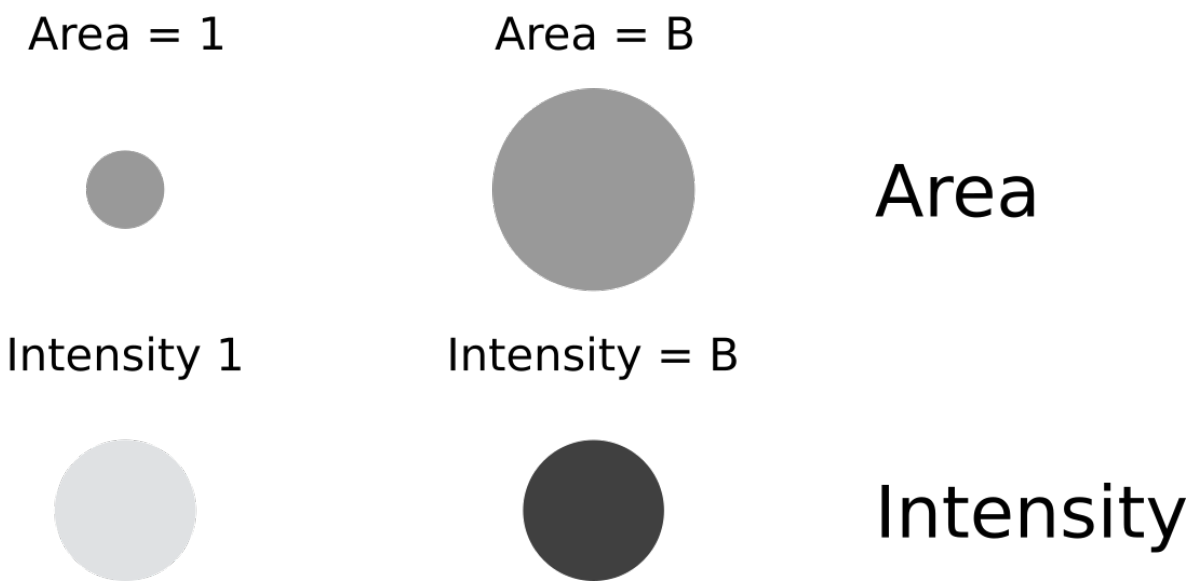
In what form do we perceive amount best?

- What is the value of B?



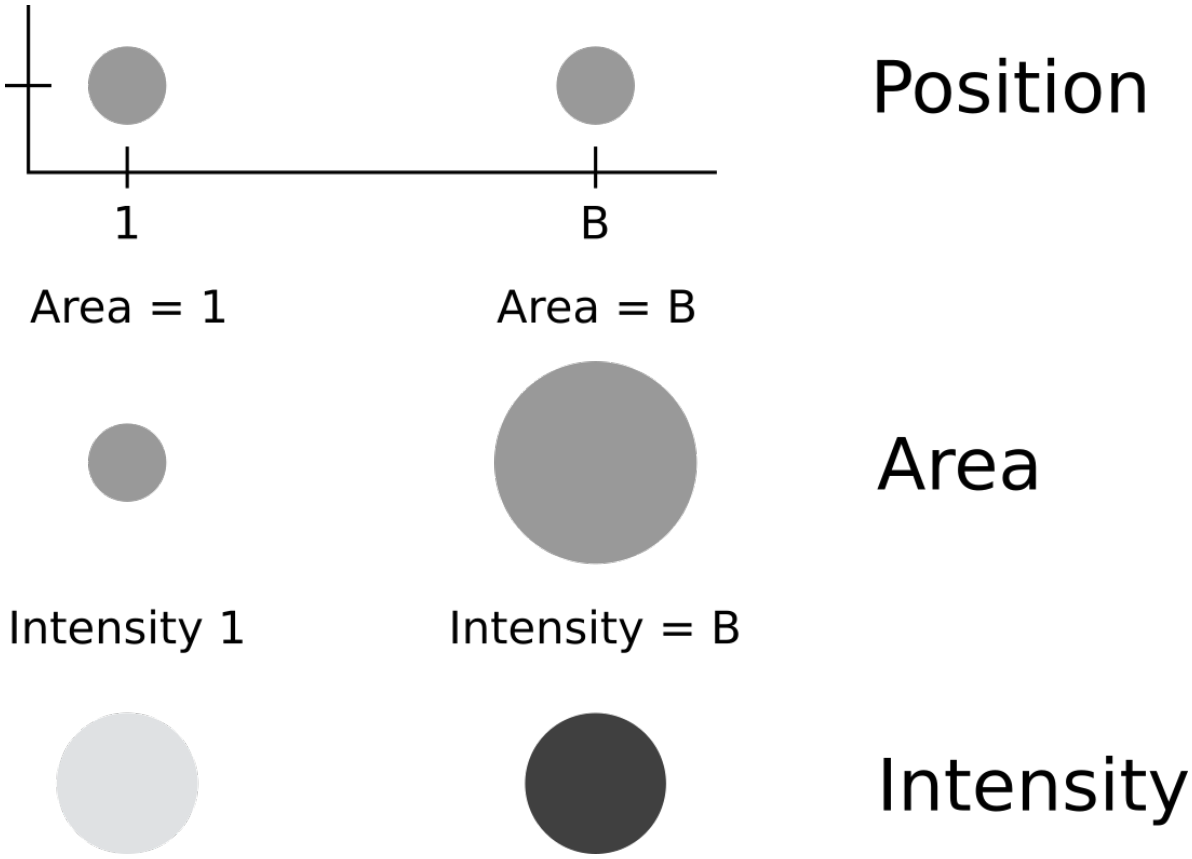
In what form do we perceive amount best?

- What is the value of B?



In what form do we perceive amount best?

- What is the value of B?



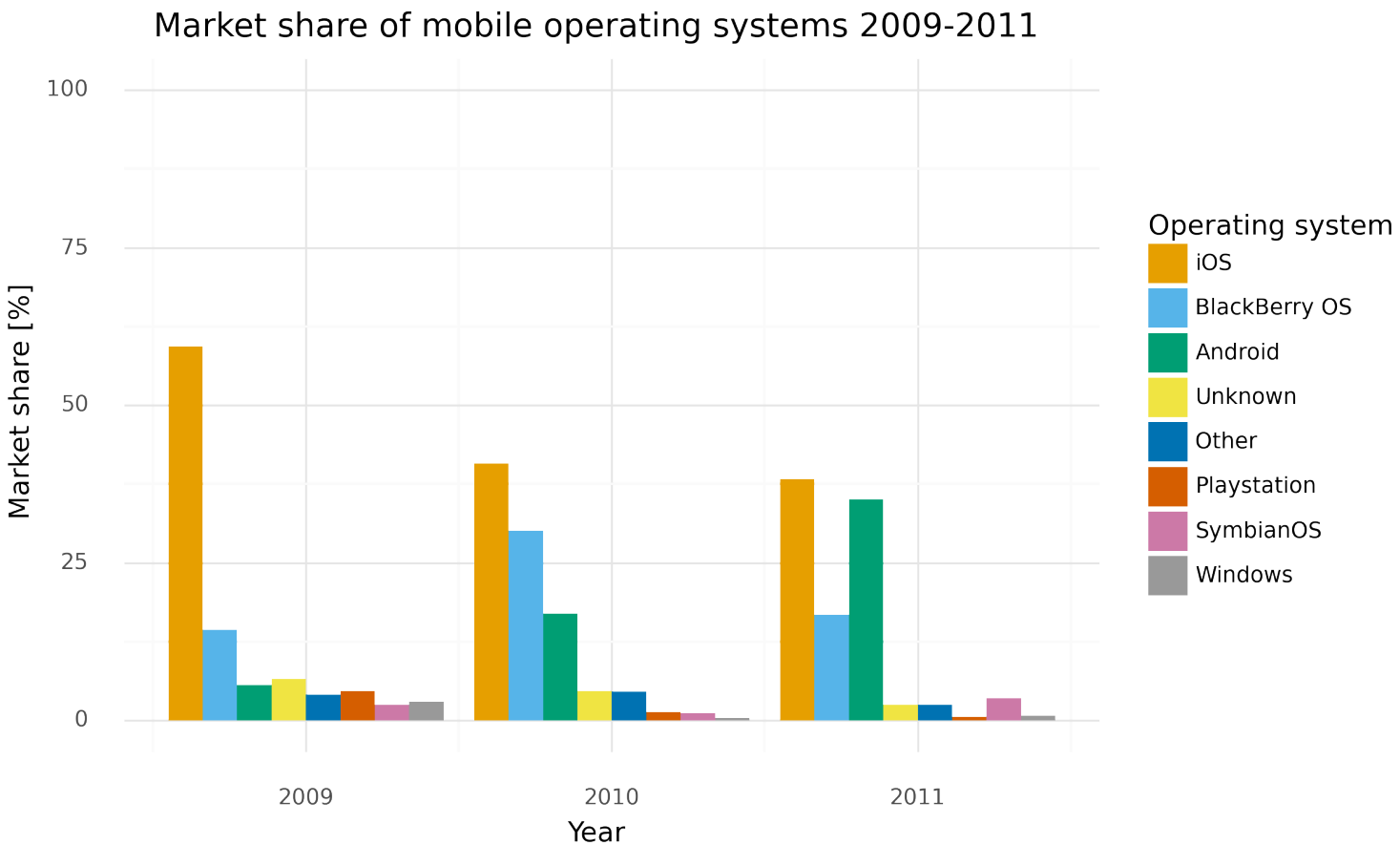
Why do we plot amounts with barplots?

We perceive ratios (1:7) best in order of:

1. Position
2. Area
3. Intensity (colour)

With visuals we control the narrative

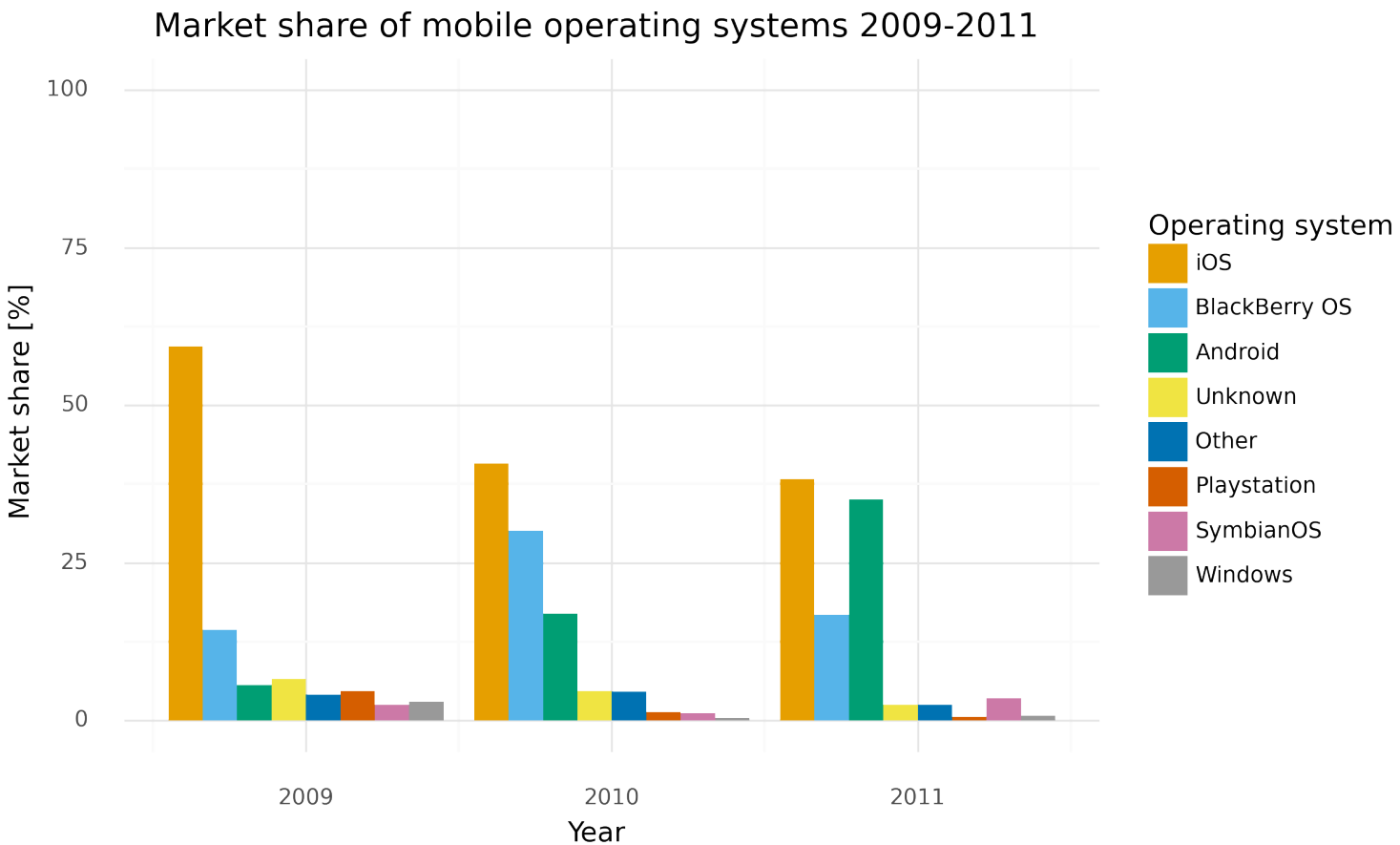
Which was the third biggest operating system 2011?



With visuals we control the narrative

Which was the third biggest operating system 2011?

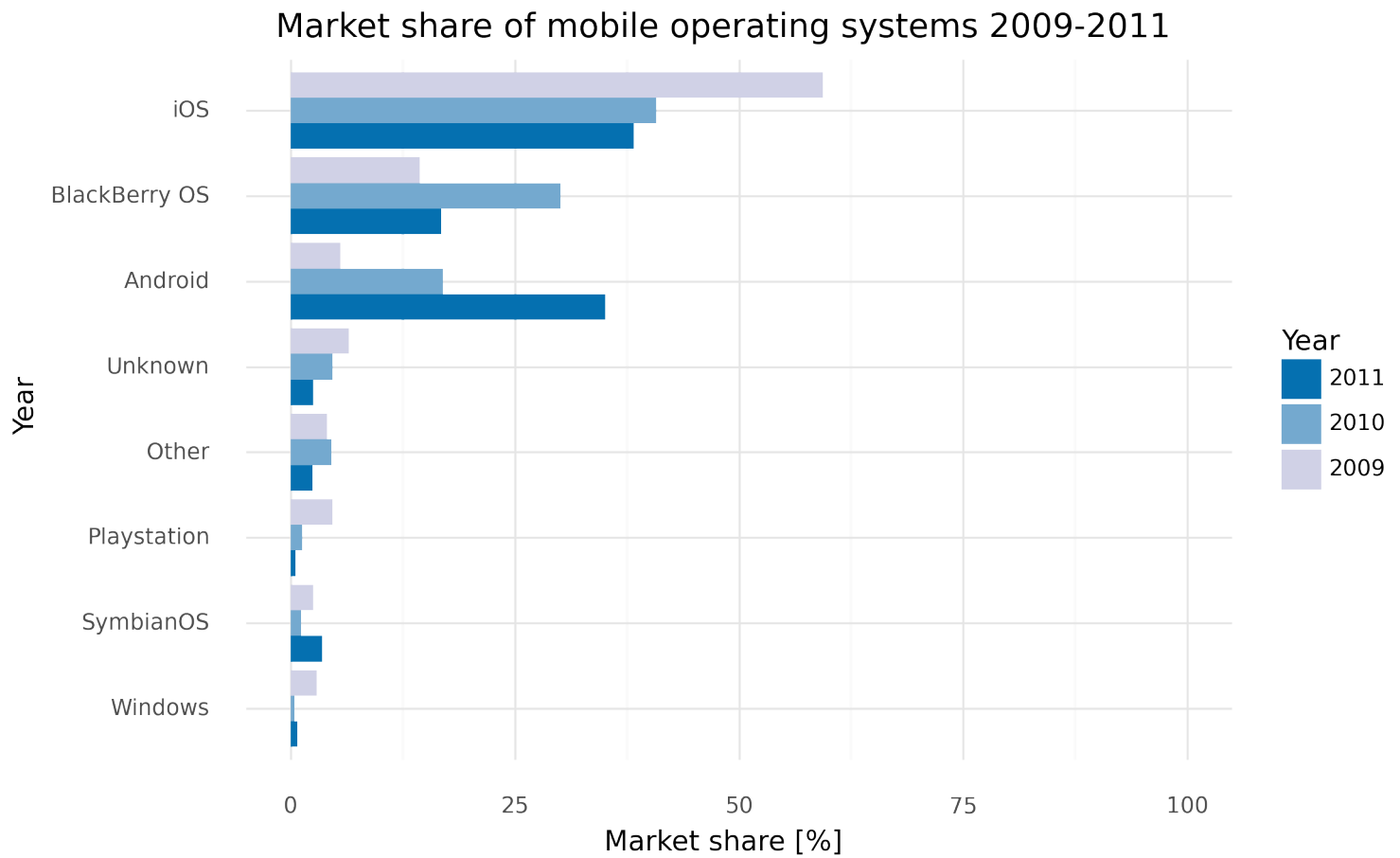
Which was the second best year for BlackBerry?



With visuals we control the narrative

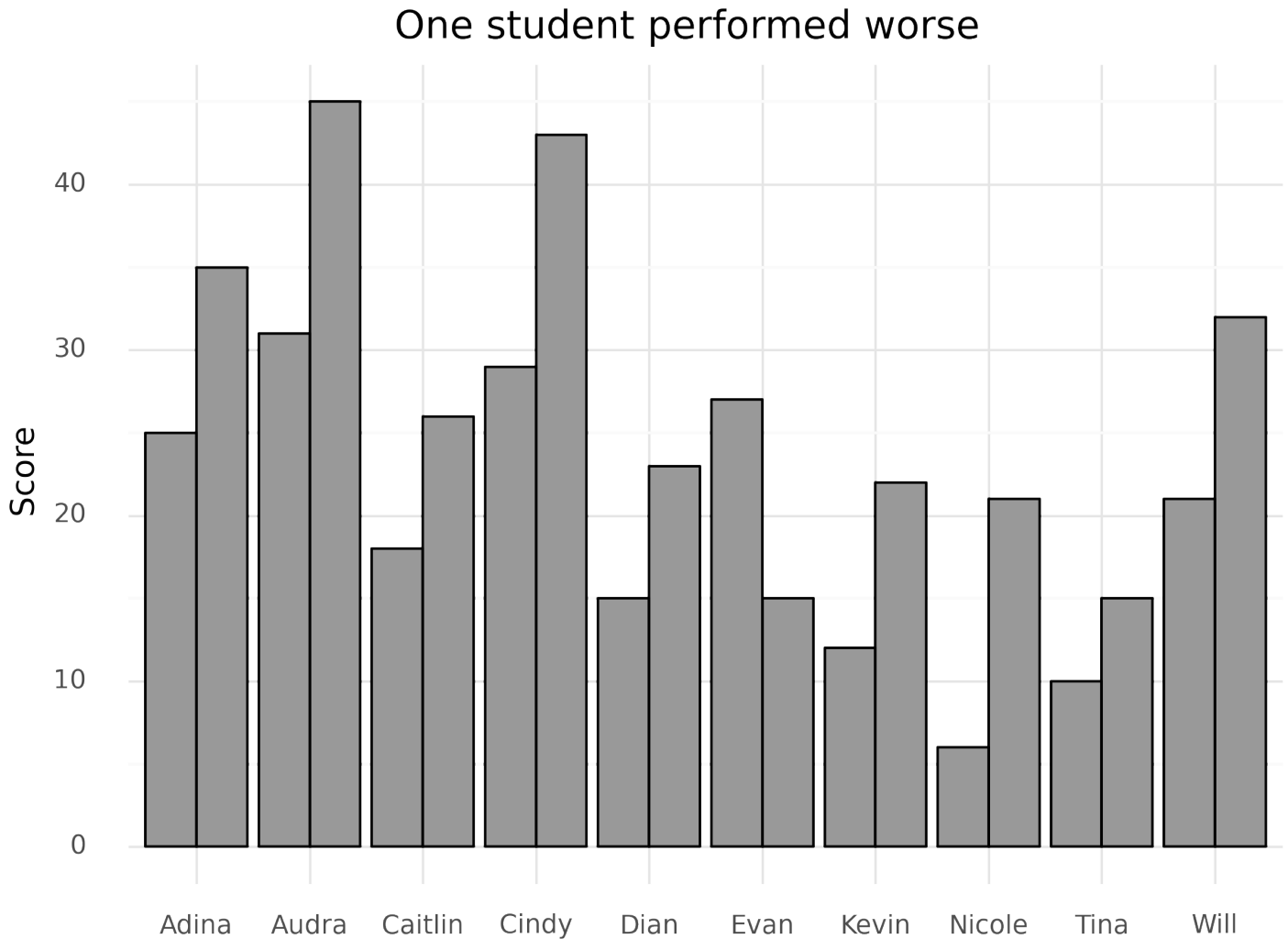
Which was the third biggest operating system 2011?

Which was the second best year for BlackBerry?



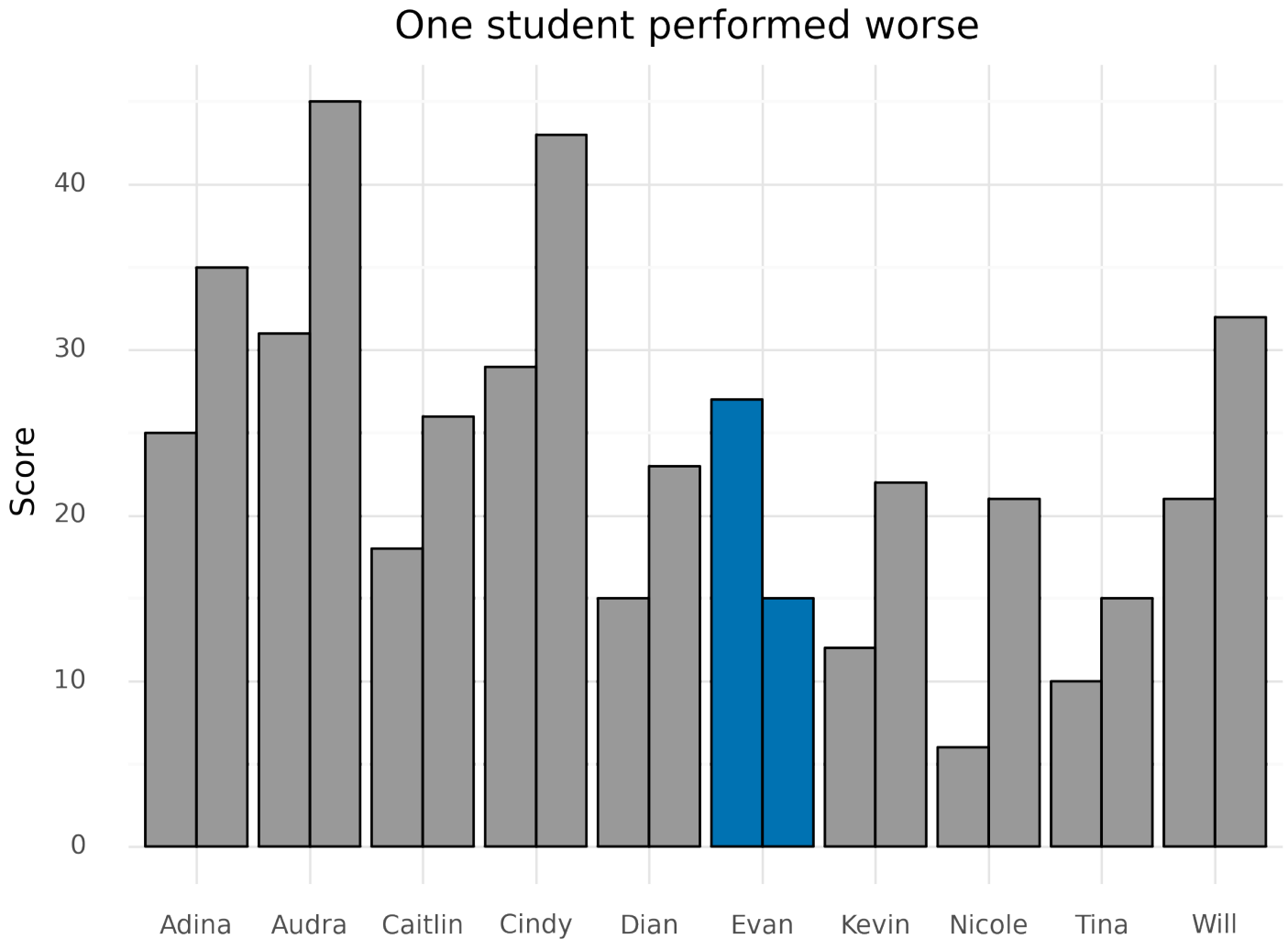
Querying a visual is demanding

Which student got worse?



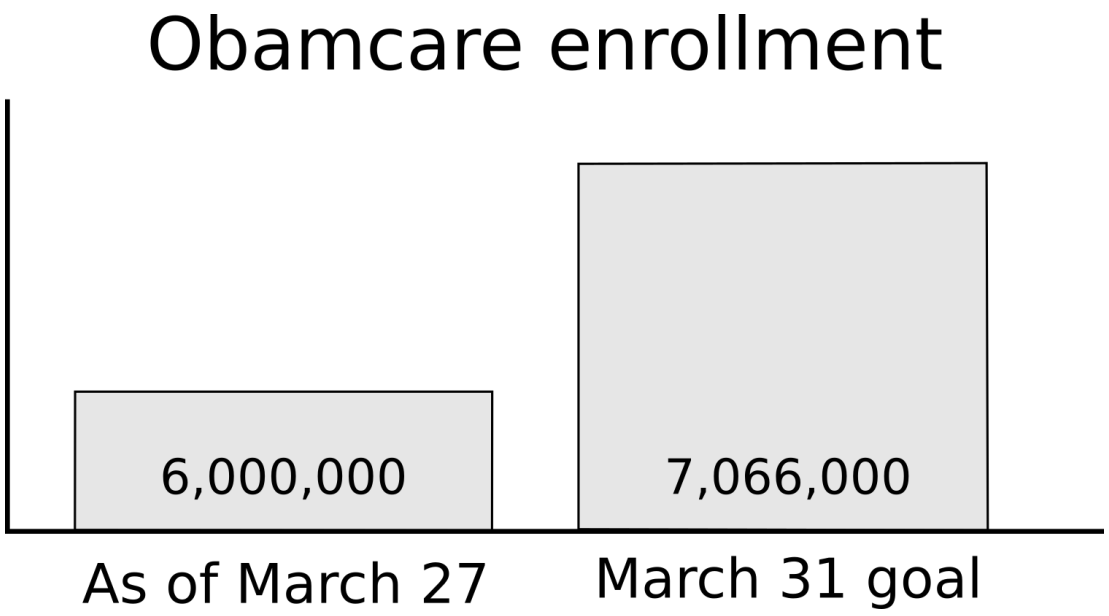
Querying a visual is demanding

Which student got worse?



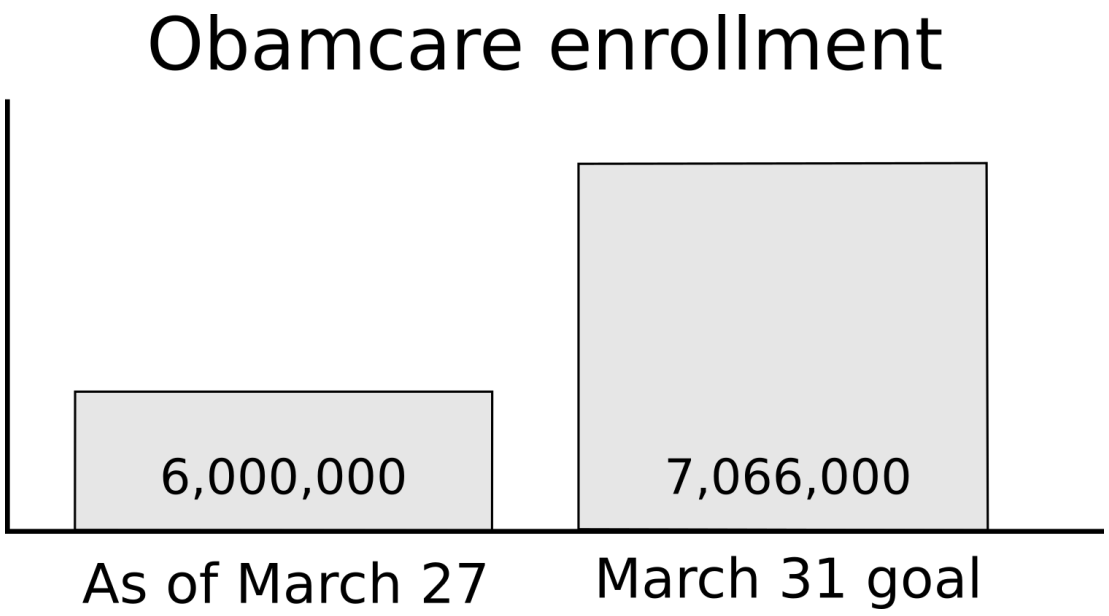
With great narrative power comes great responsibility

- Always start barplots from zero



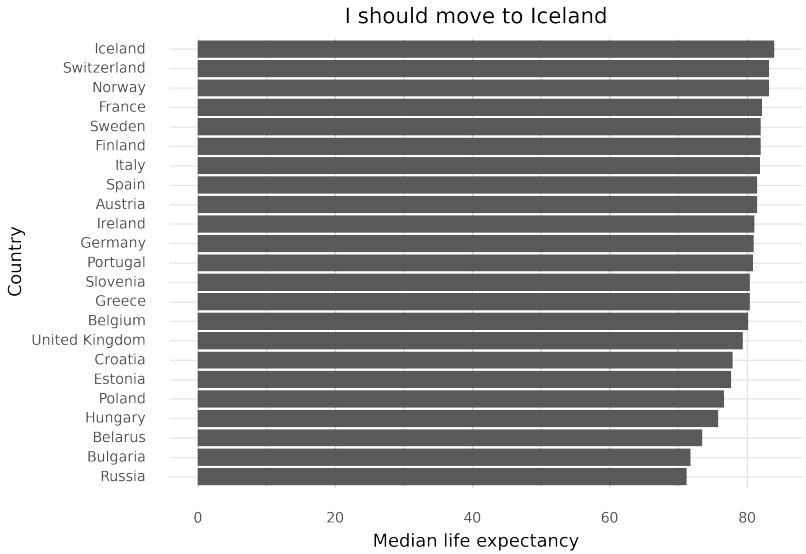
With great narrative power comes great responsibility

- Always start barplots from zero
- Arrange your visual to convey the main message
  - Later we cover how to represent complex data



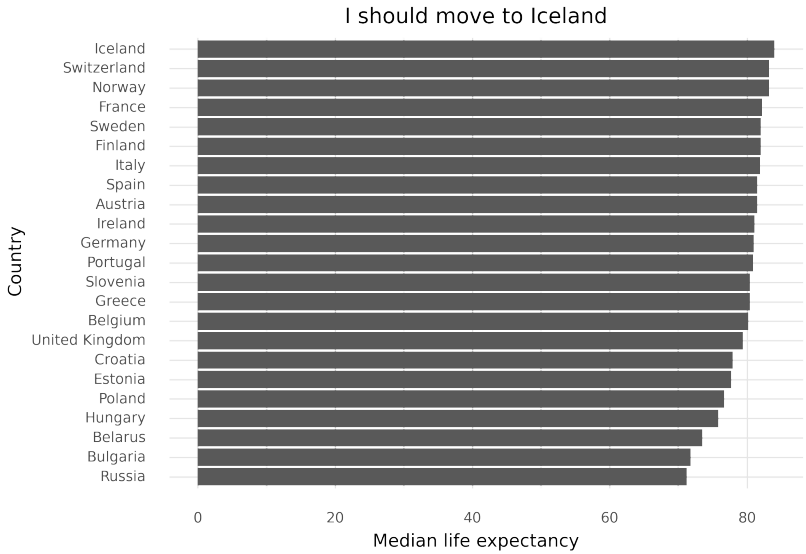
When starting from zero is a bad option

- Barplot introduces substantial visual noise



When starting from zero is a bad option

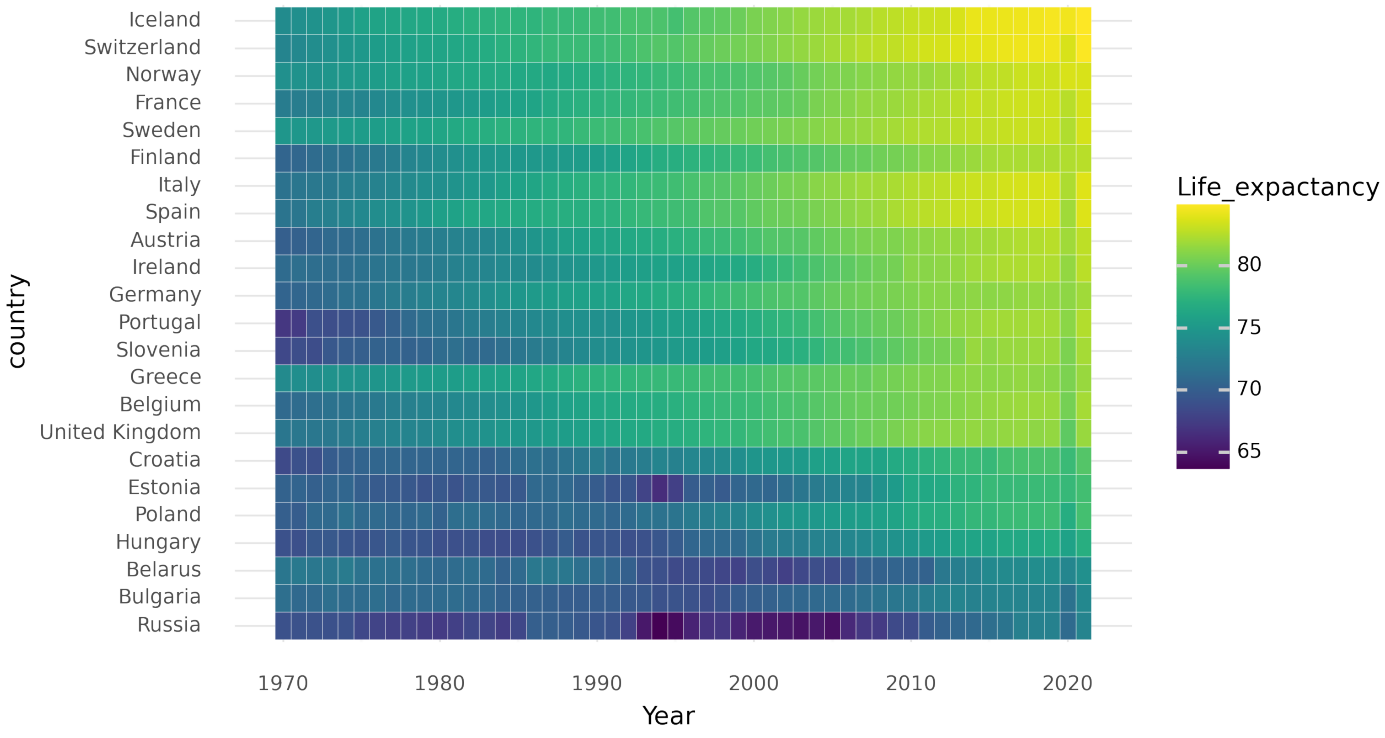
- Barplot introduces substantial visual noise
- **Solution** : The dot-plot (geom_bar → geom_point)



Iceland or Mordor?

When looking for trends

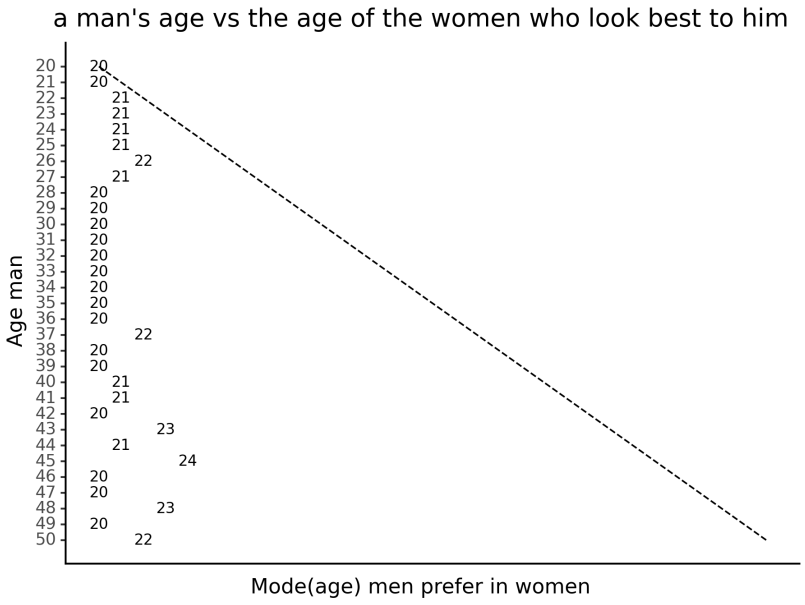
- For many variables (e.g time and country) heatmaps
  - Not as accurate as barplots
  - Sort to decrease visual noise
  - Choose suitable colormap (next lecture)



Distributions - should we care?

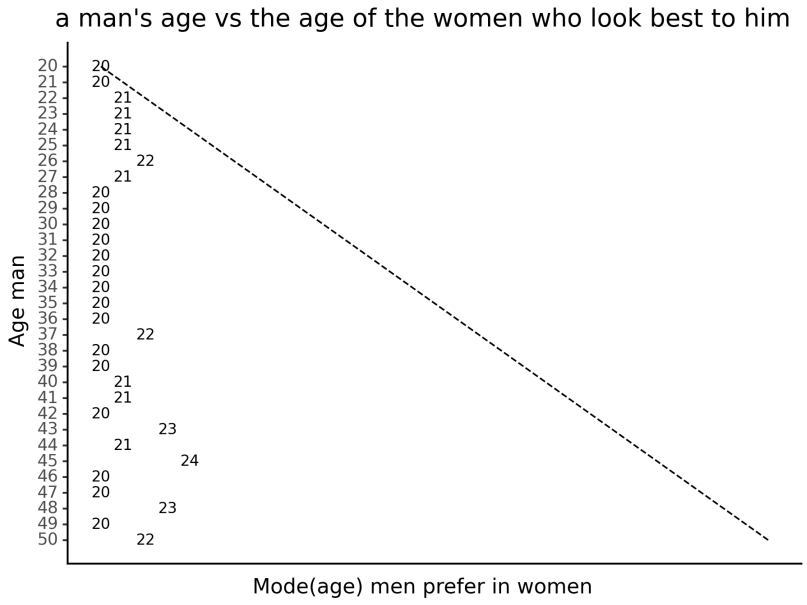
Dating in the modern world

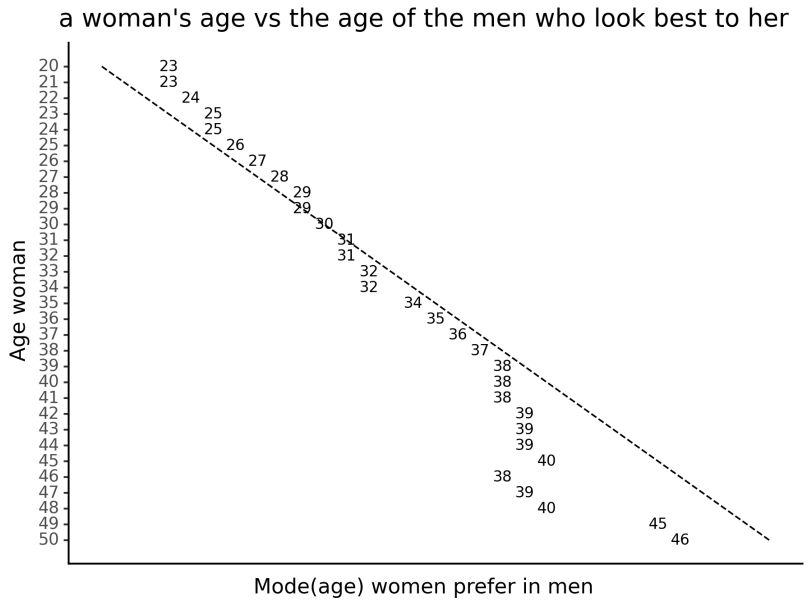
- Preferences men and women from *okcupid*



Dating in the modern world

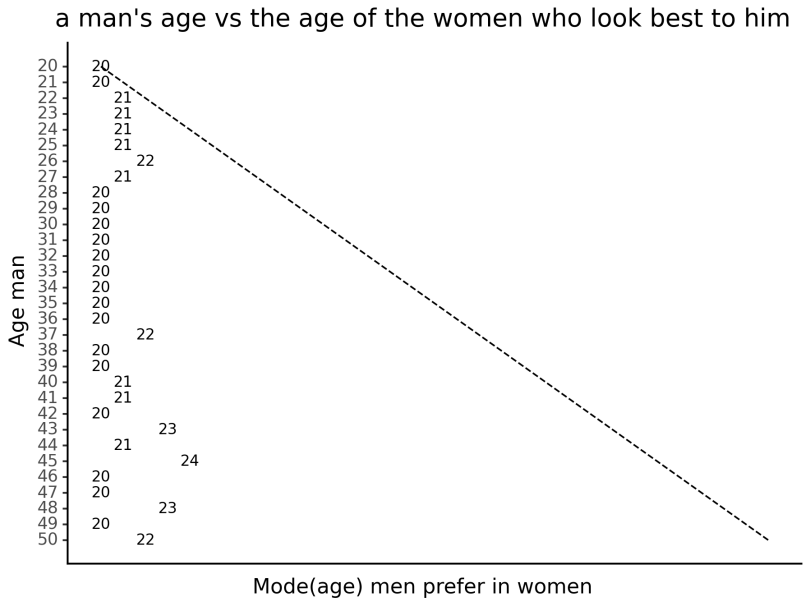
- Preferences men and women from *okcupid*

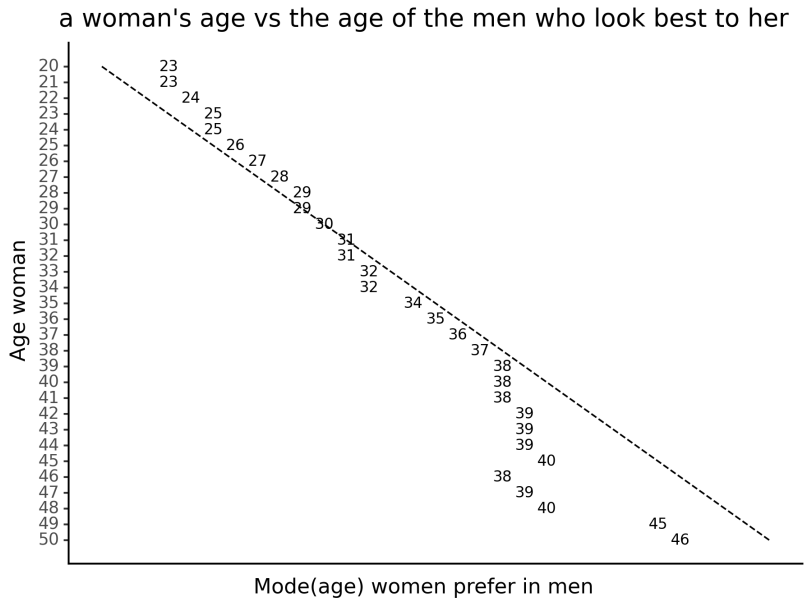




Dating in the modern world

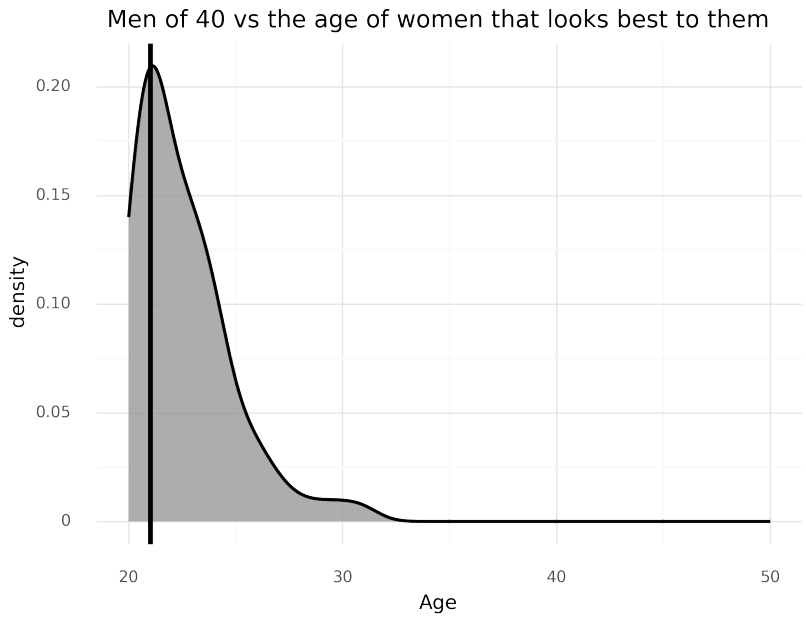
- Preferences men and women from *okcupid*
- Do you see any potential problem?

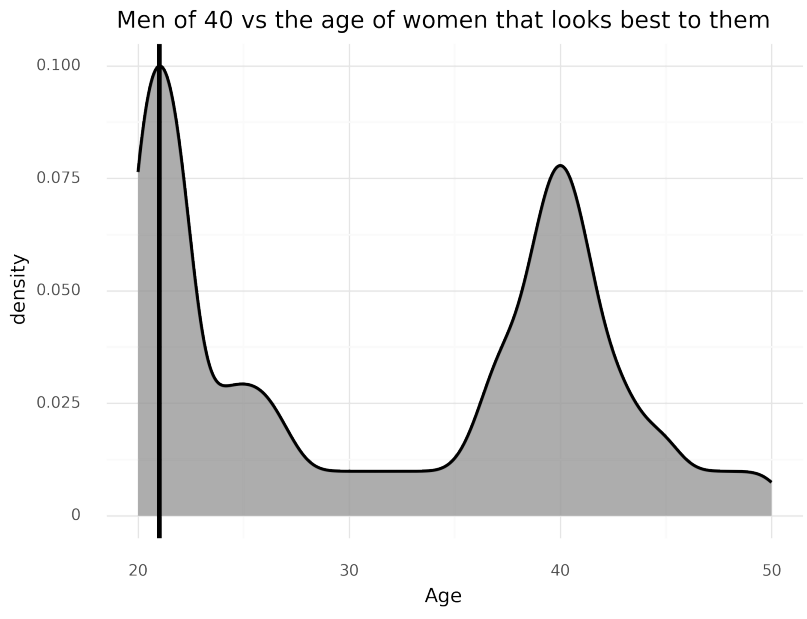




Summary statistics can be too sensational

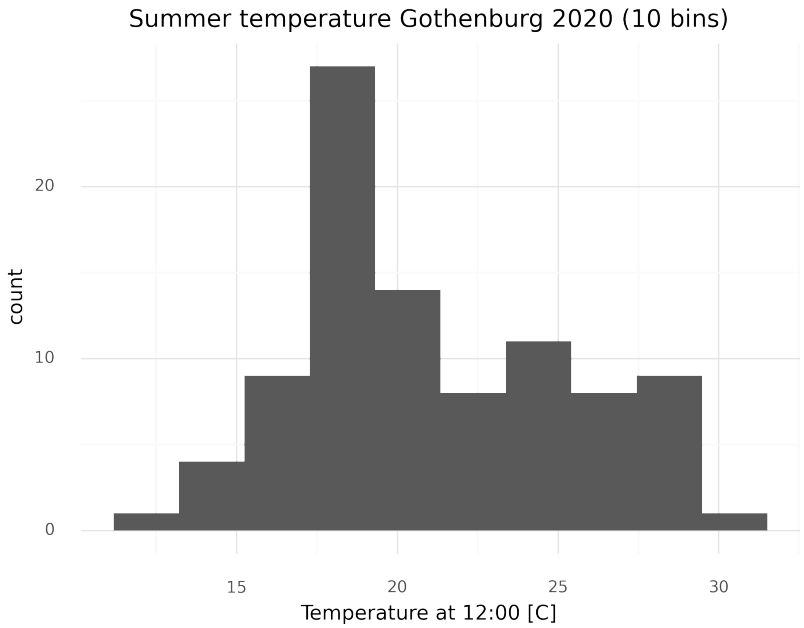
- Mode - the value which appears most frequently in a dataset
  - Below fictional data below





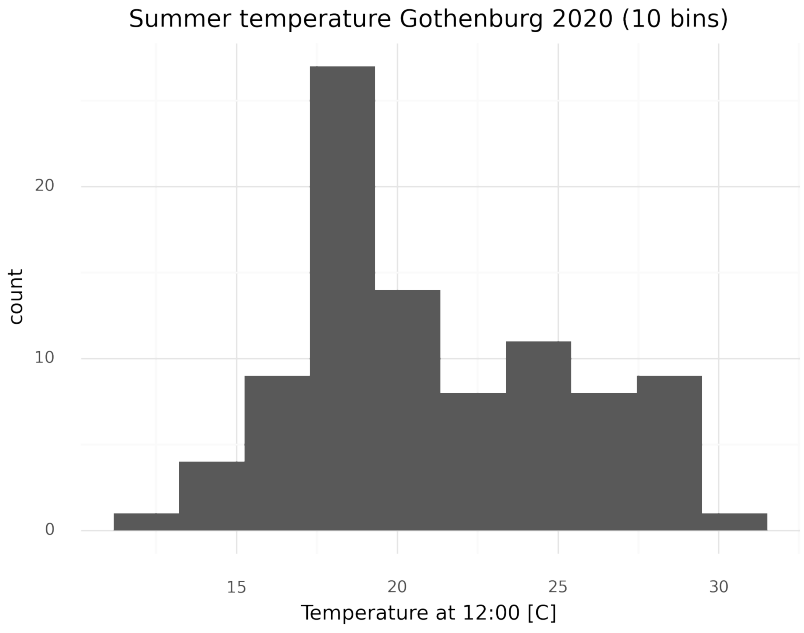
Histograms - the workhorse of distributions

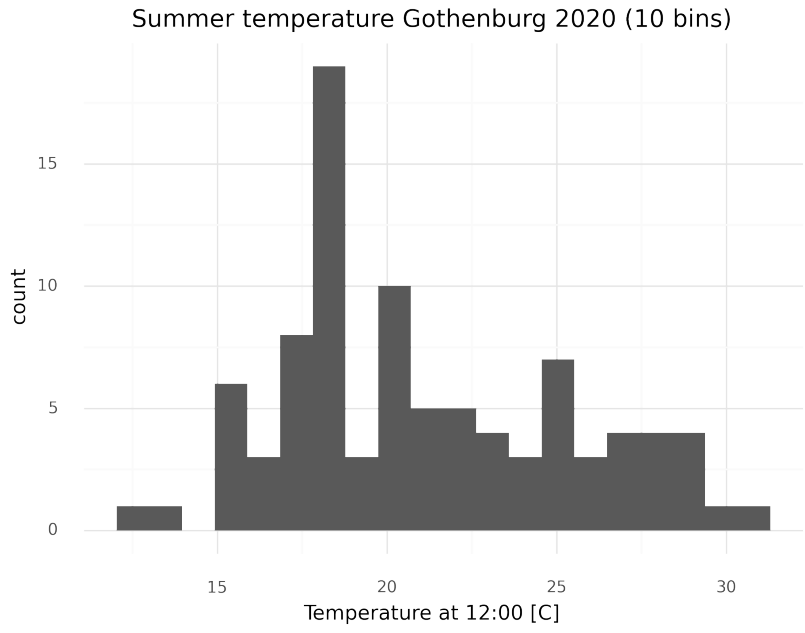
- Extremely common
- Tuning parameter - bin size
  - Test different sizes



Histograms - the workhorse of distributions

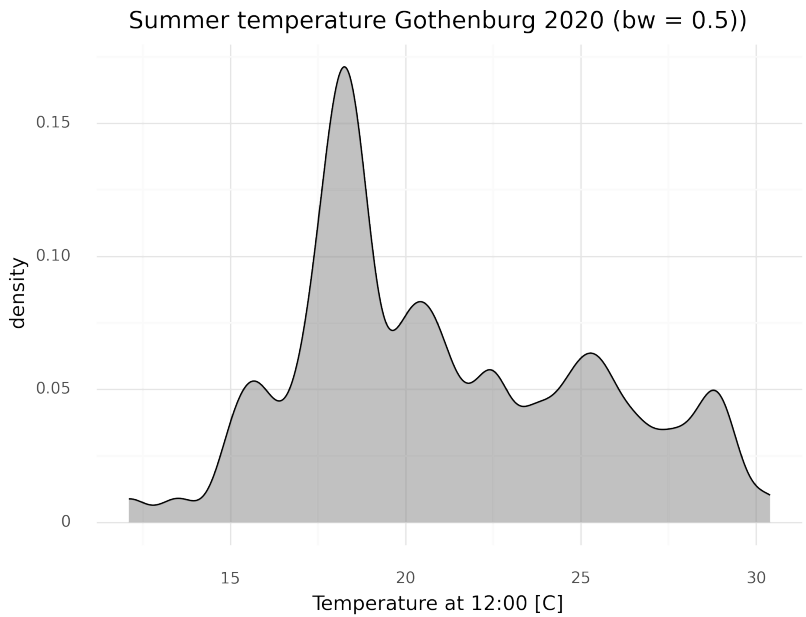
- Extremely common
- Tuning parameter - bin size
  - Test different sizes





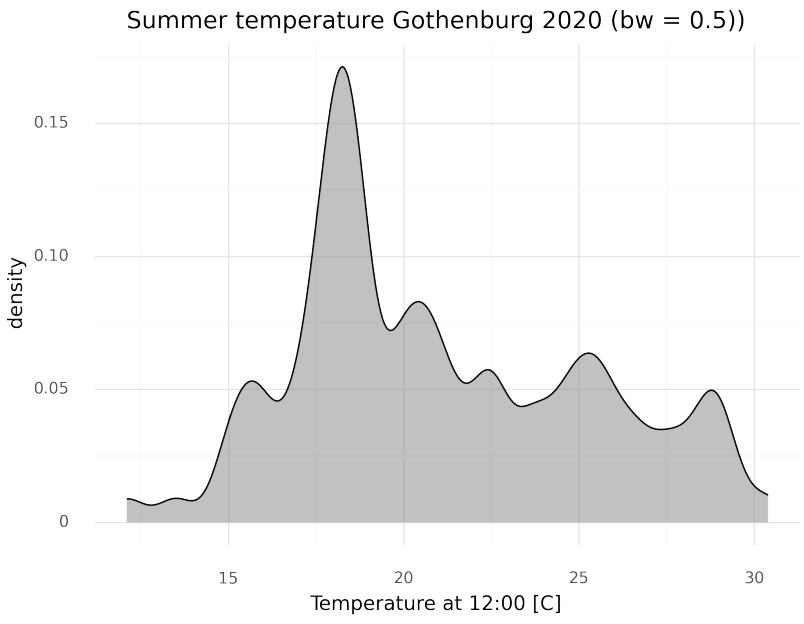
Density plots - a newer fresher approach

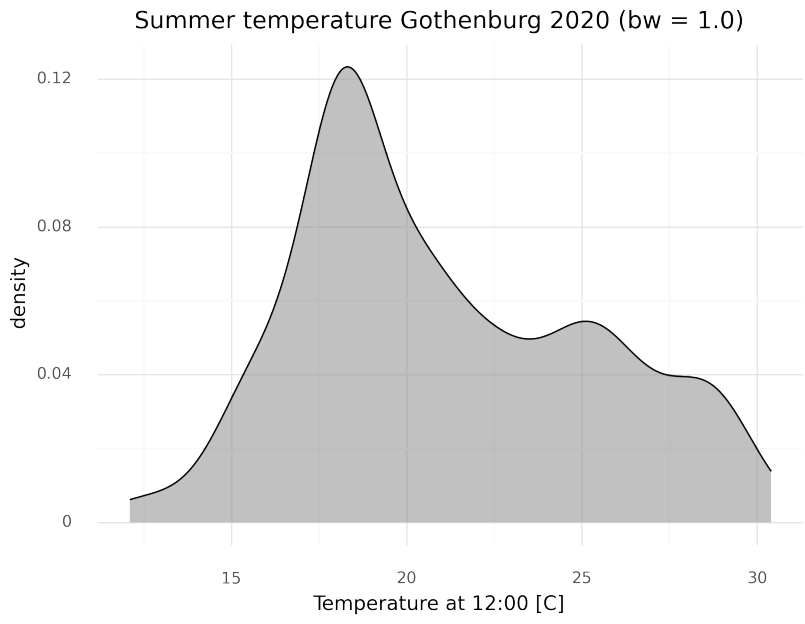
- Kernel density estimate
- Tuning parameter - bin width
  - Sensitive when there are few data points



Density plots - a newer fresher approach

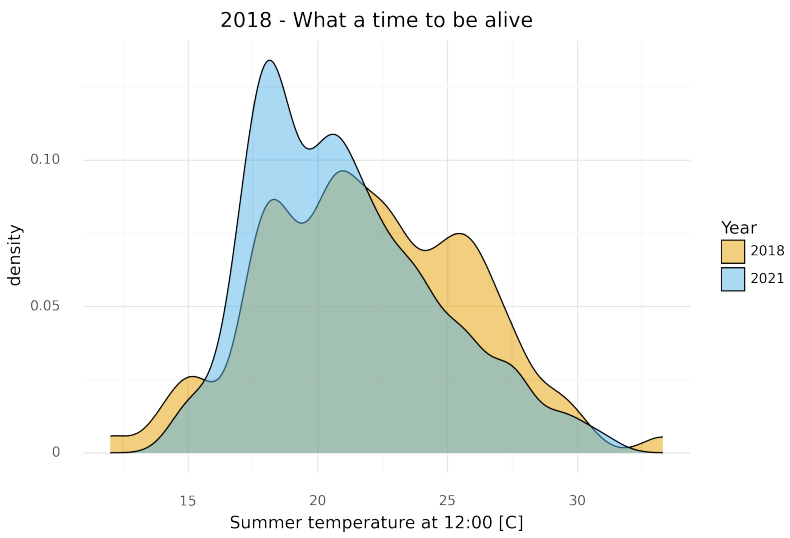
- Kernel density estimate
- Tuning parameter - bin width
  - Sensitive when there are few data points

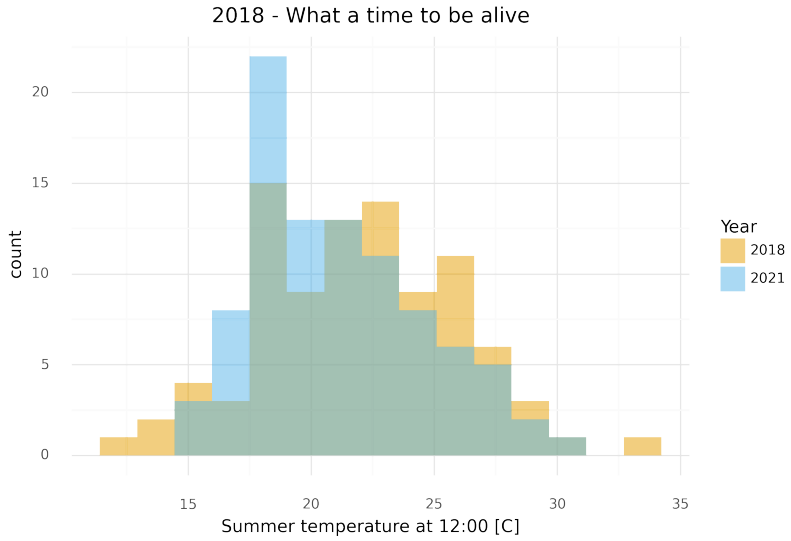




More than one distribution → density plots

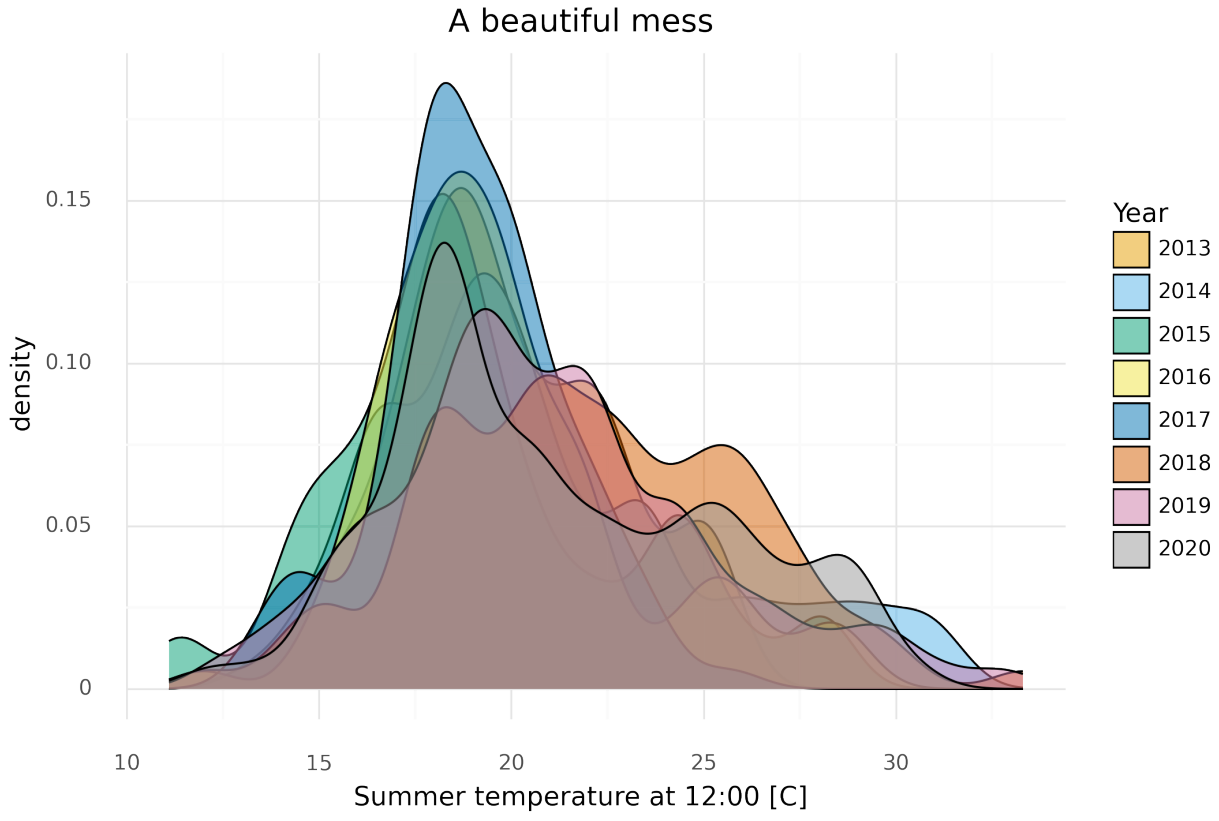
- Histograms hard to interpret for several categories
- Density plots often easier
  - Use transparent filling (alpha=0.6)





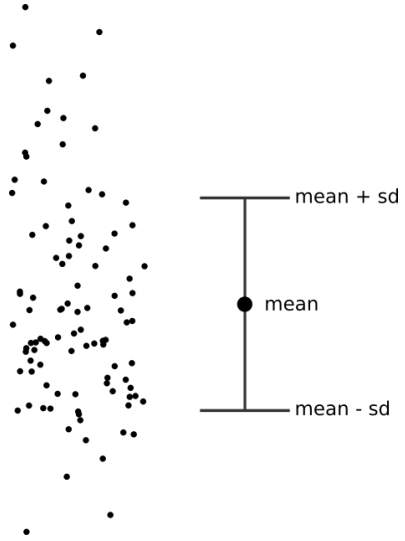
More than 4 distributions - density plots?

- For more than 4 distribution density plots often become a mess, and interpreting them becomes non-trivial.



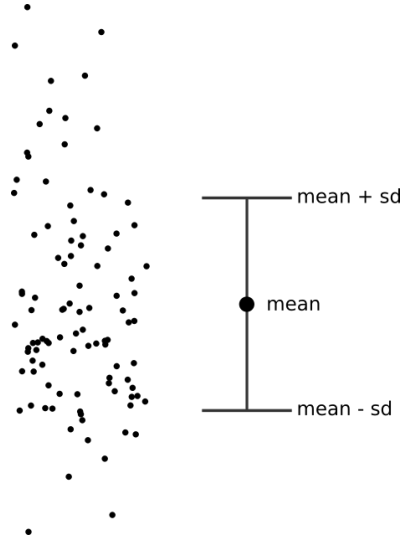
More than 4 distributions - summary statistics?

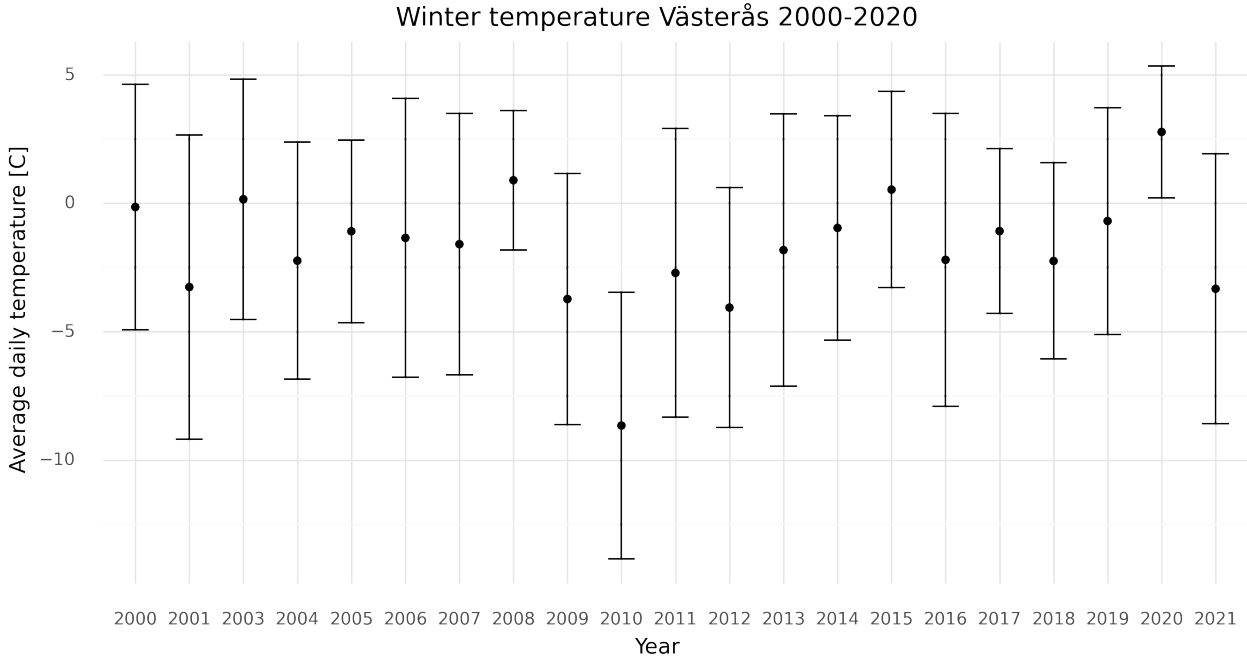
- Mean $\pm$ standard deviation



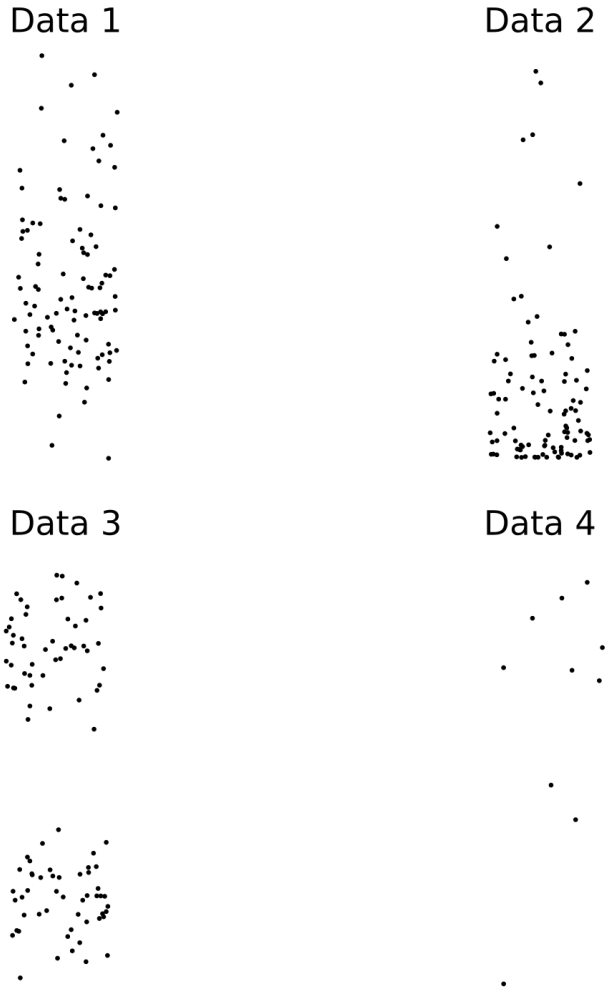
More than 4 distributions - summary statistics?

- Mean ± standard deviation

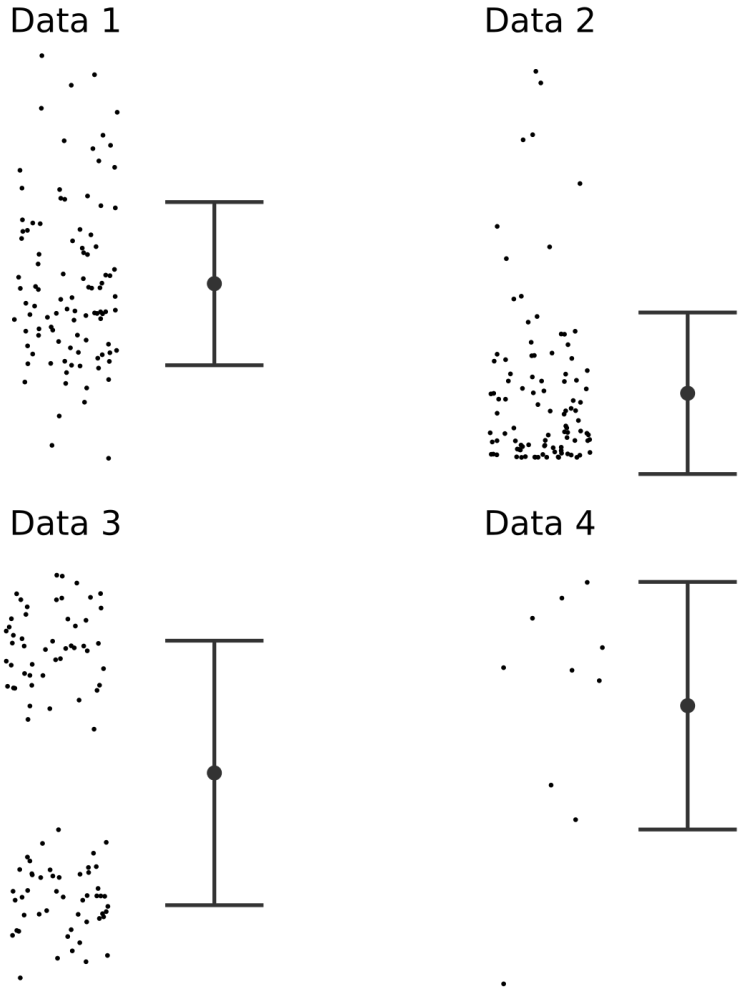




When do summary statistic break?

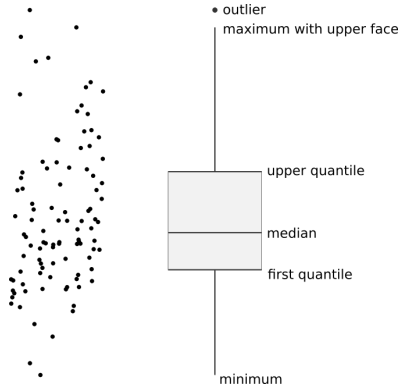


When do summary statistic break?



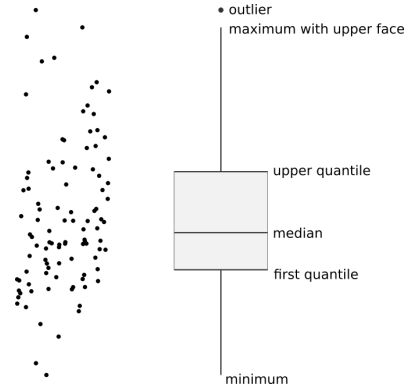
More than 4 distributions - boxplots?

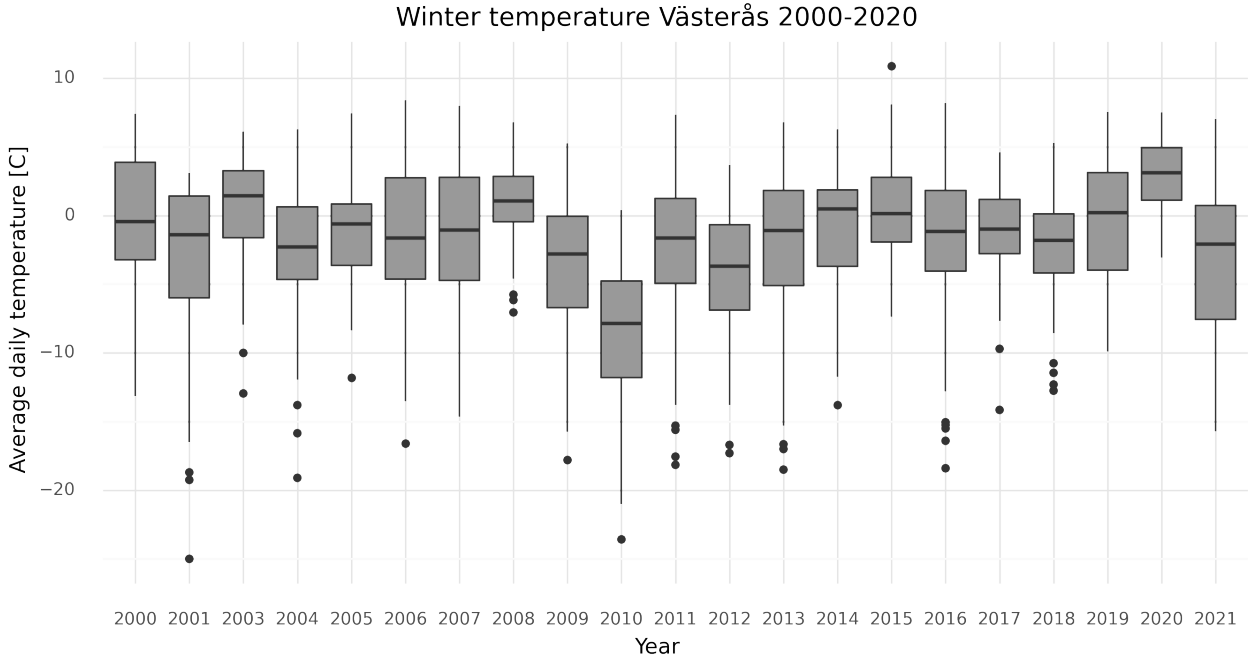
- Quantiles + max/min



More than 4 distributions - boxplots?

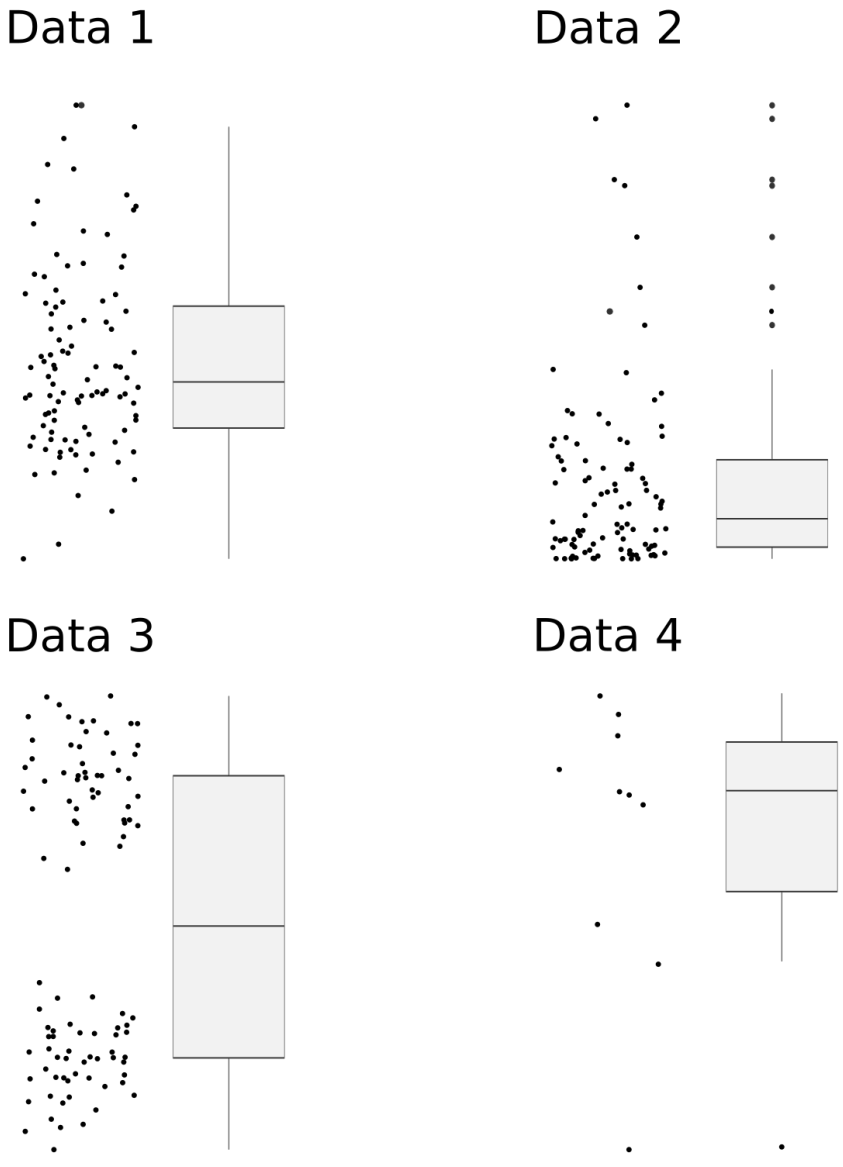
- Quantiles + max/min





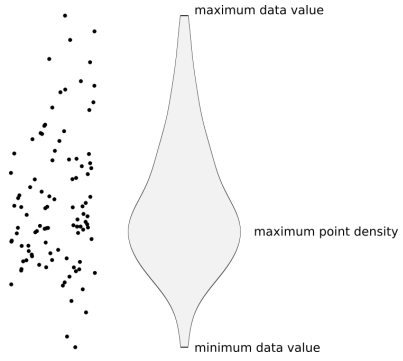
When do boxplots break?

When do boxplots break?



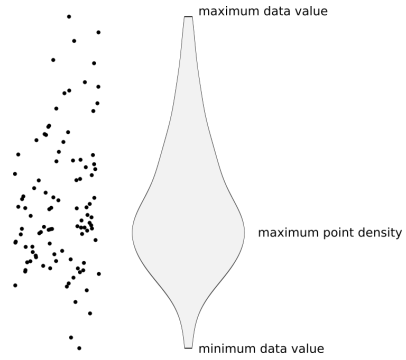
More than 4 distributions - violin plots?

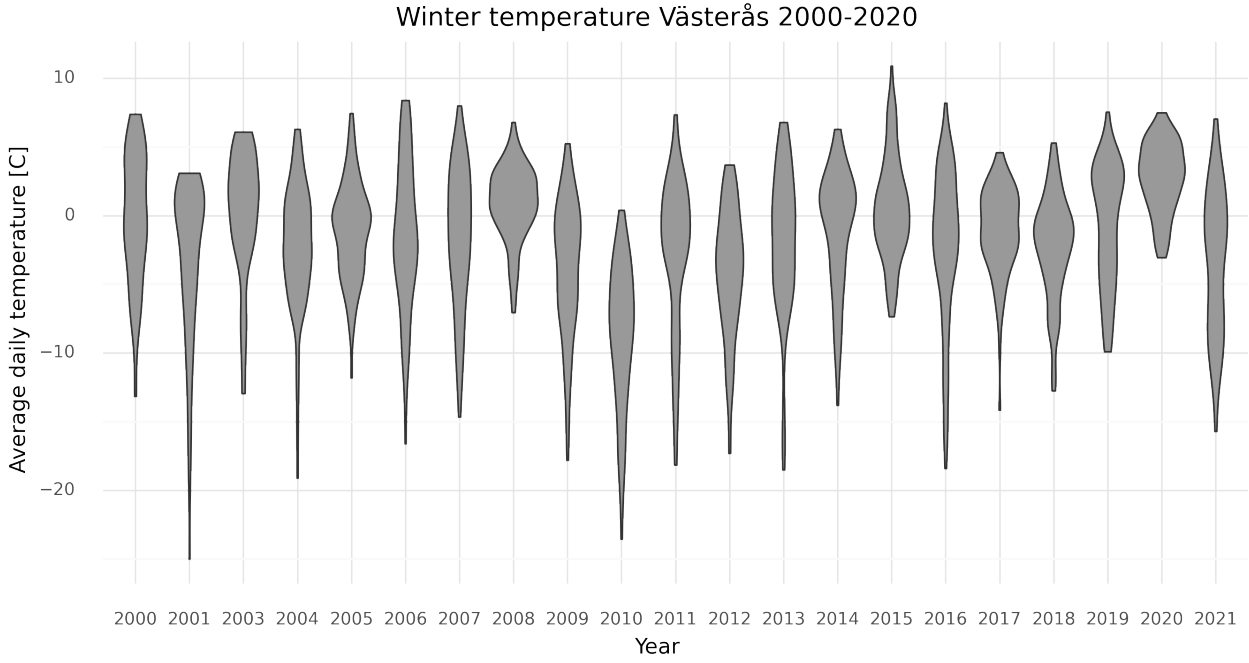
- Kernel density estimate



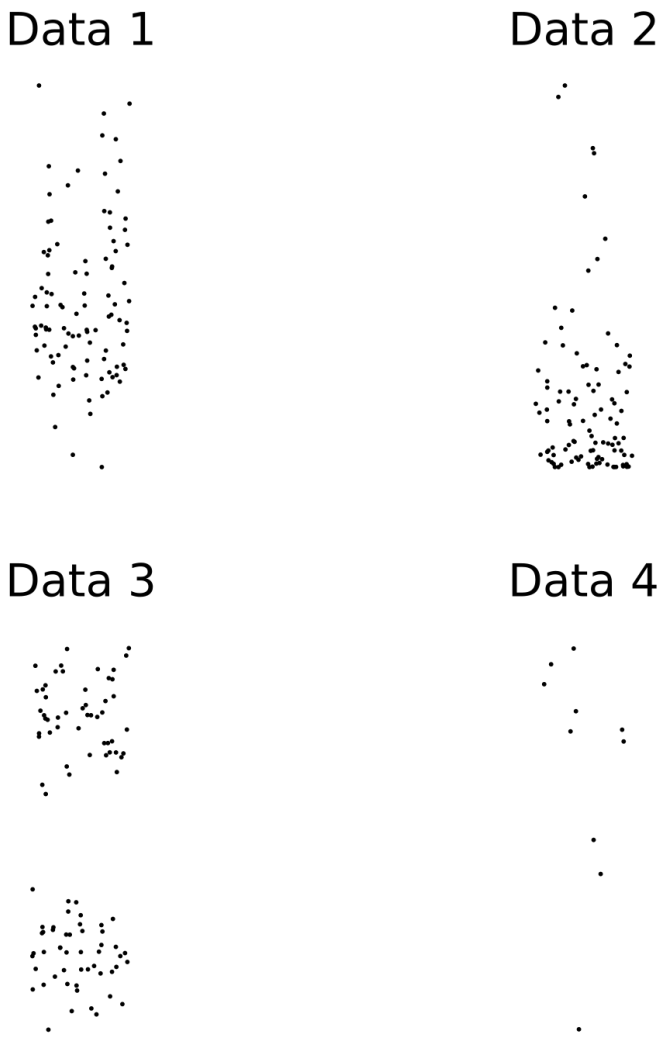
More than 4 distributions - violin plots?

- Kernel density estimate

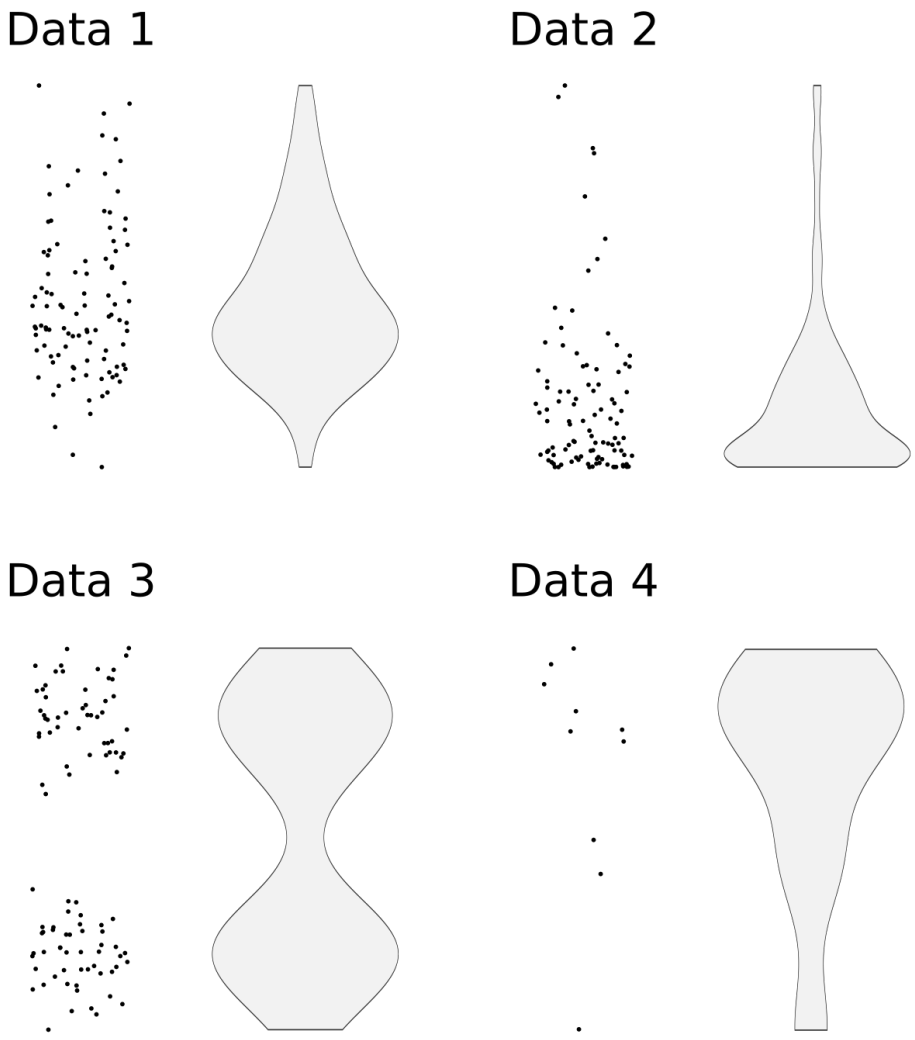




When do violin plots break?

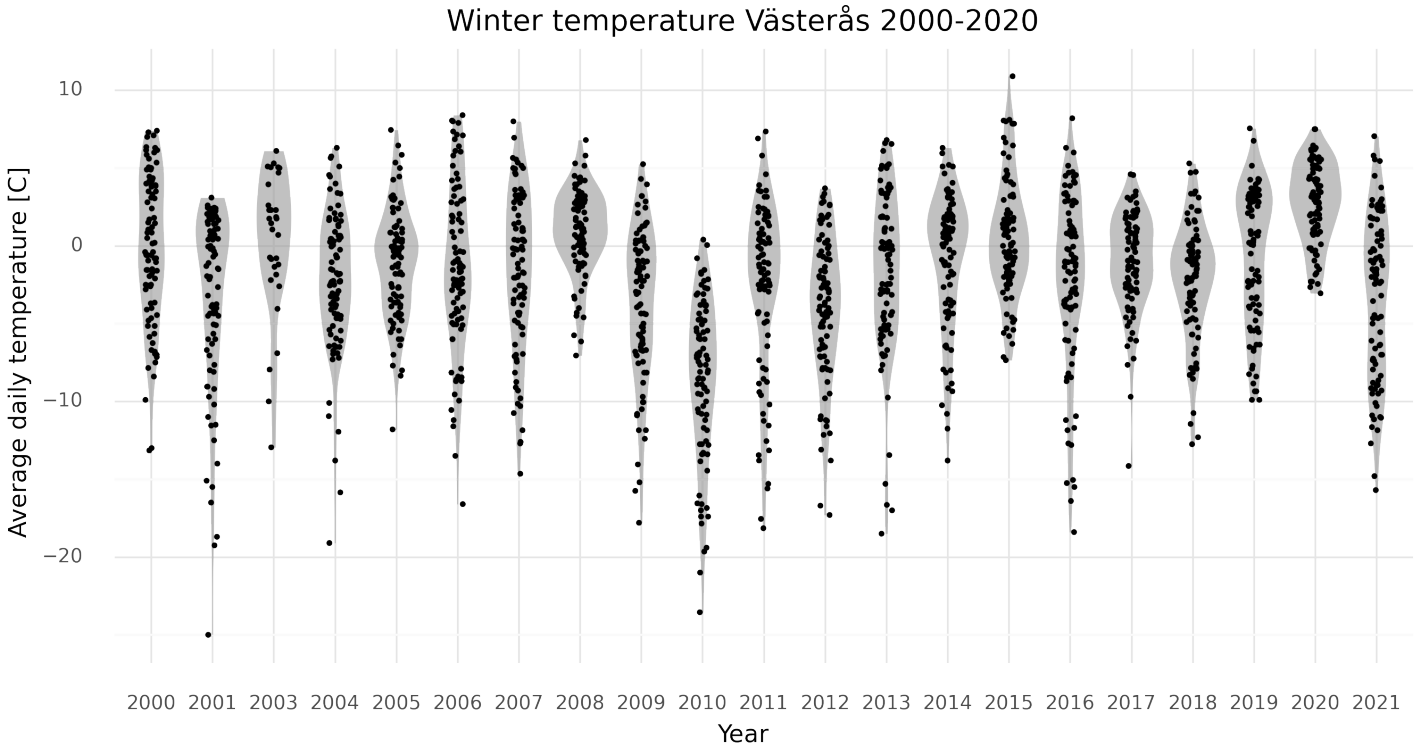


When do violin plots break?



More than 4 distributions - violin + data points?

Violin and boxplot can make it appear there are many data points than observed → plot the data points



Take home messages

- Avoid visual noise and hard to read labels
  - For barplots sorting and flipping can do a lot
- With great narrative power comes great responsibility
  - Make it easy to extract main take-home message
- Summary statistics can be too sensational
- Violin plots with underlying data points are a powerful combination

For next lecture...

Why is the colour palette here bad?