visualizing numerical data

- scatterplots for paired data
- other visualizations for describing distributions of numerical variables

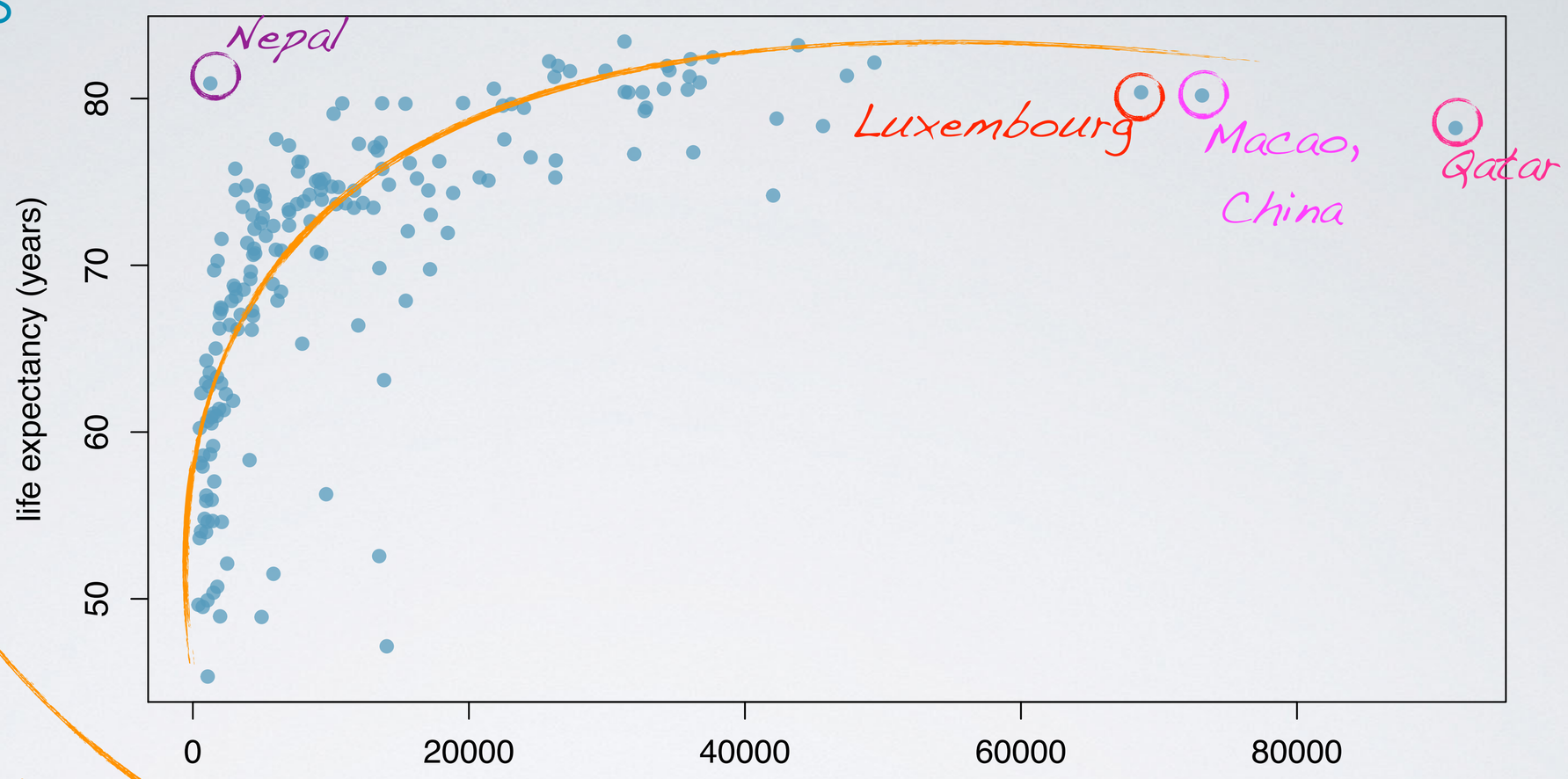
Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

data	income per person ($, 2012)	life expectancy (years, 2012)
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
…	…	…
Zimbabwe	545.3	58.142

Source: gapminder.com

scatterplots



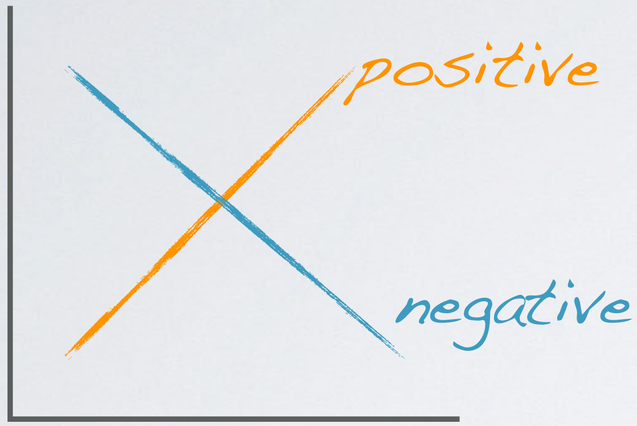
response

explanatory

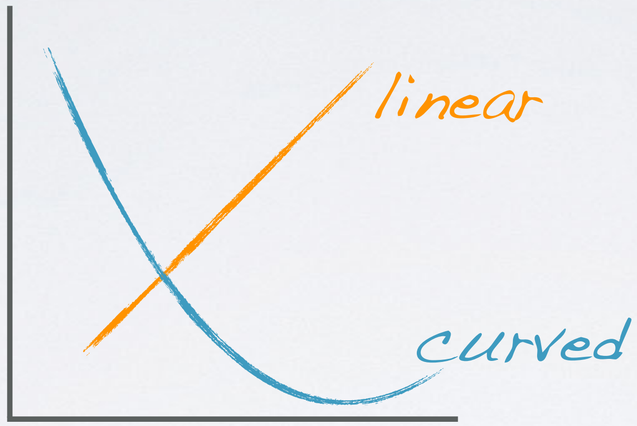
correlation, not causation

evaluating the relationship

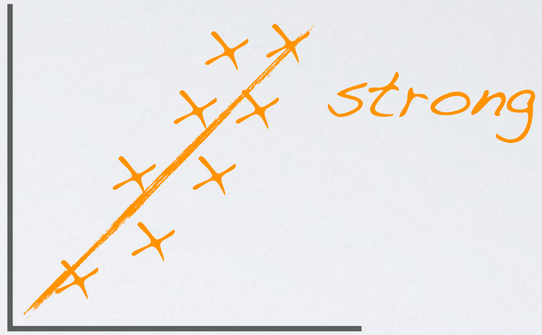
direction

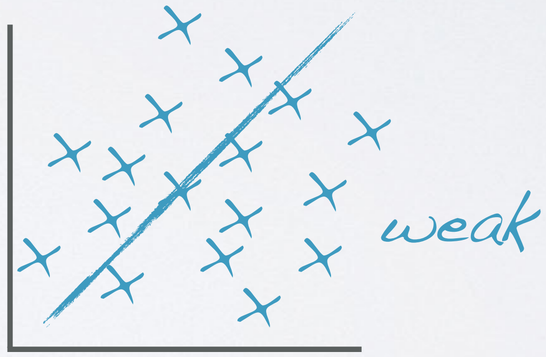


shape

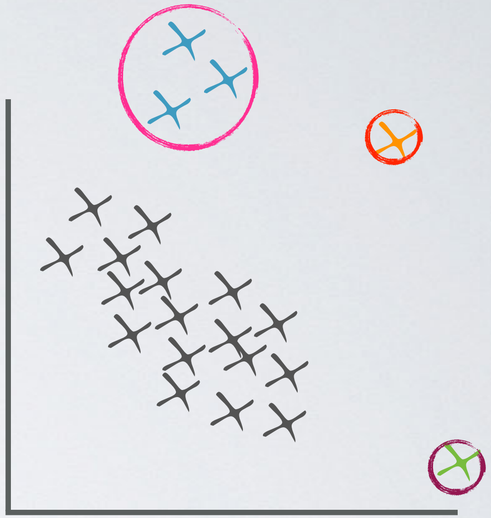


strength

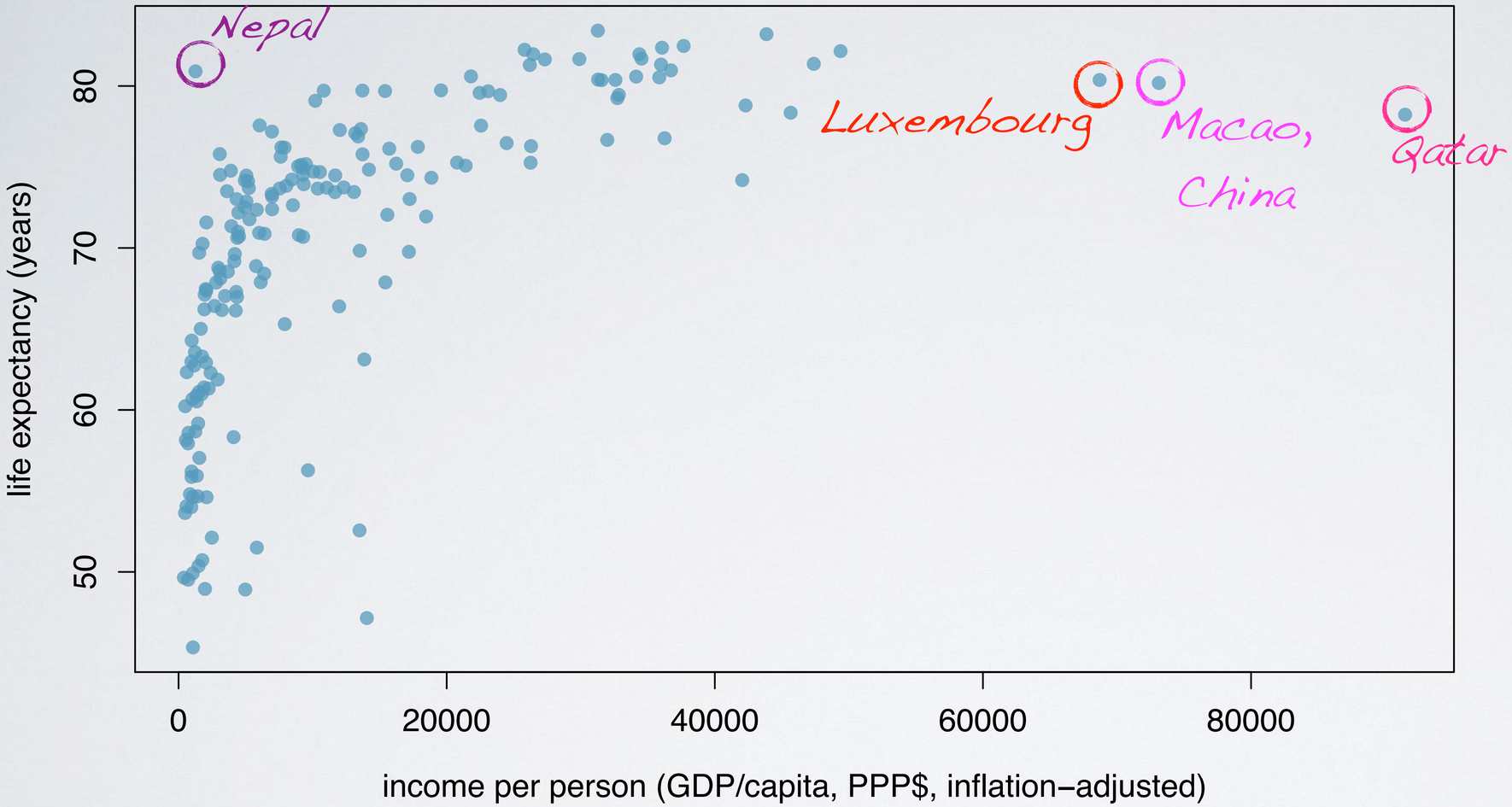


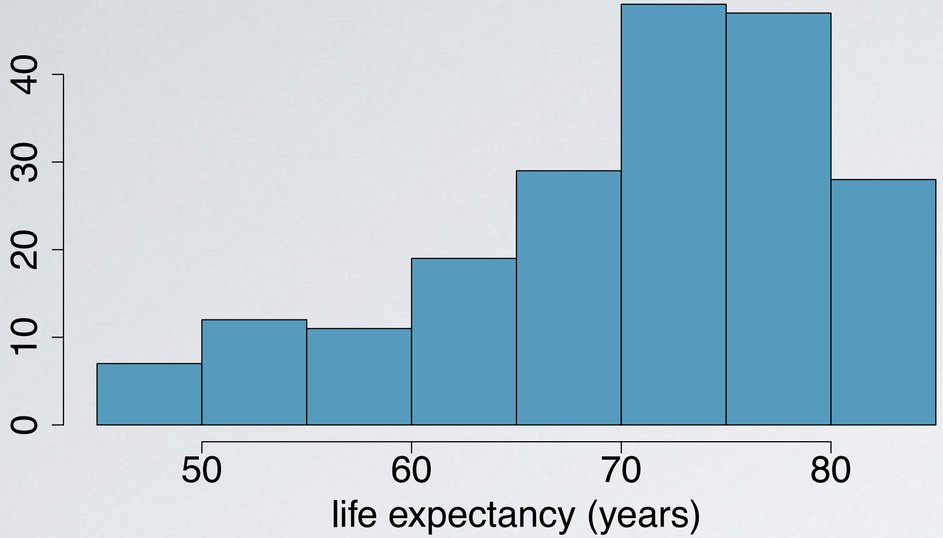


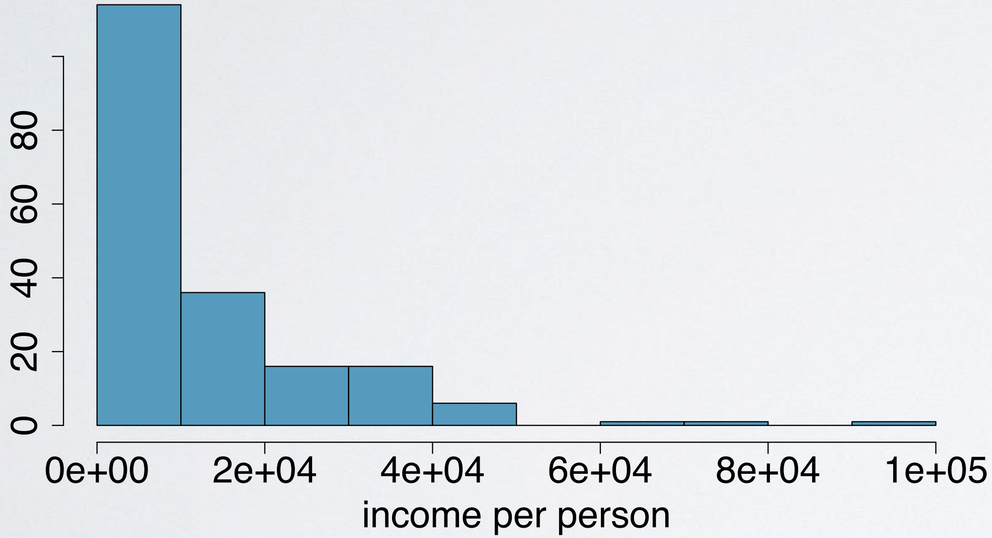
outliers



[revisit]


Nepal
Luxembourg
Macao, China
Qatar
life expectancy (years)
50
60
70
80
0
20000
40000
60000
80000
income per person (GDP/capita, PPP$, inflation–adjusted)



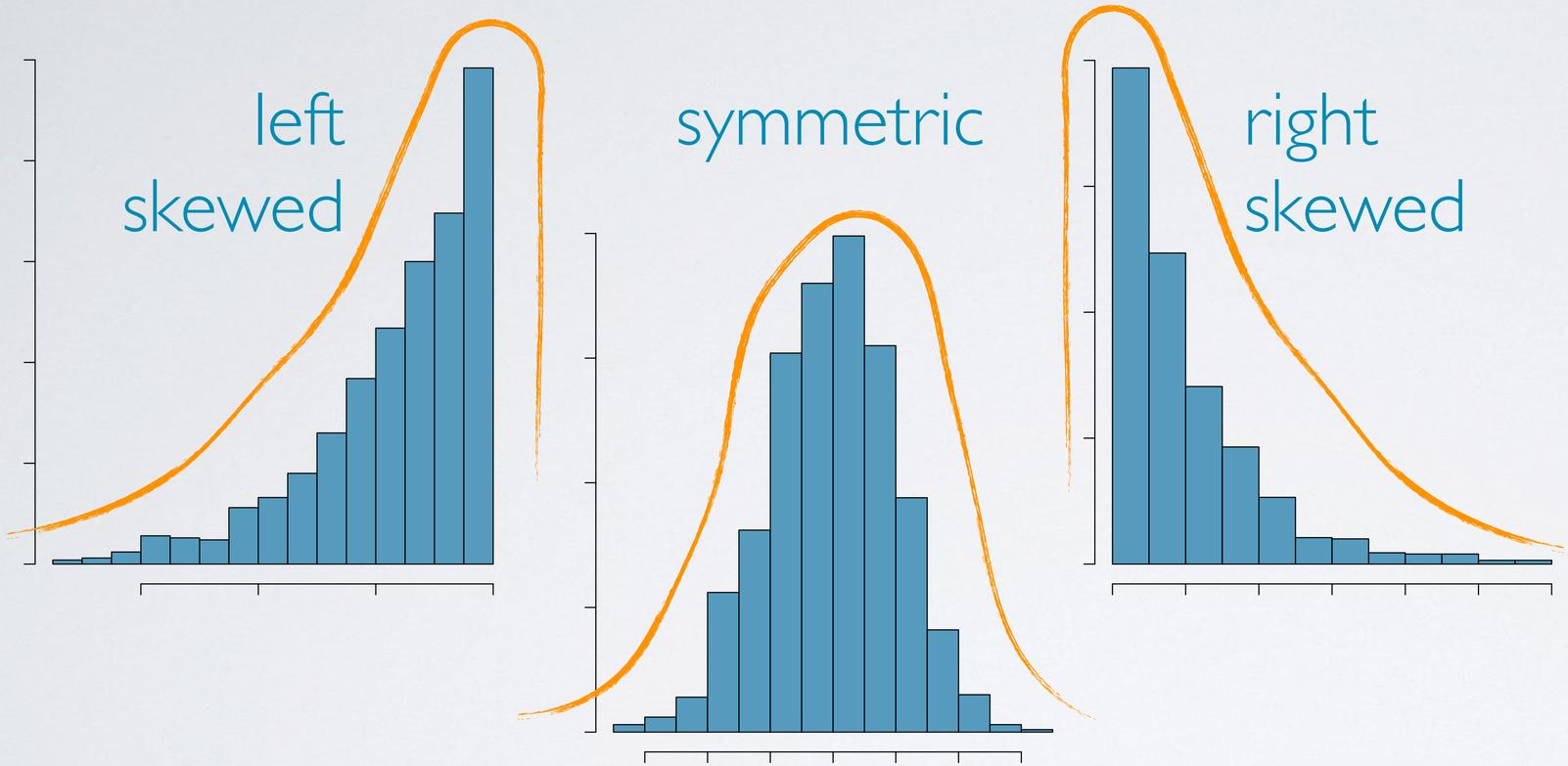


histogram

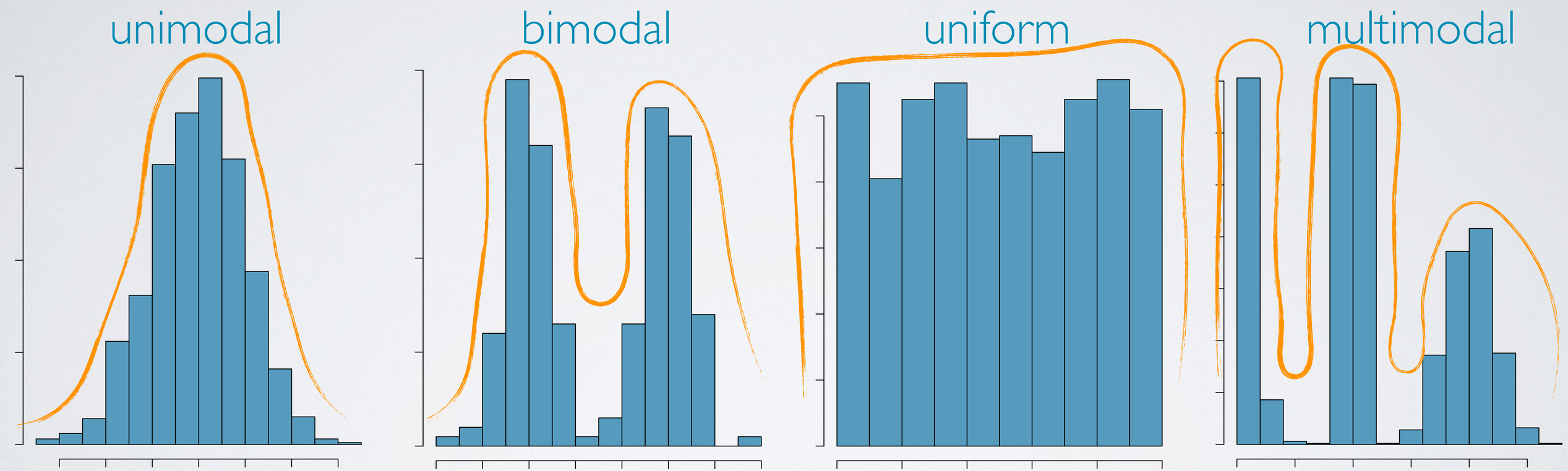
- provides a view of the data density
- especially useful for describing the shape of the distribution

skewness

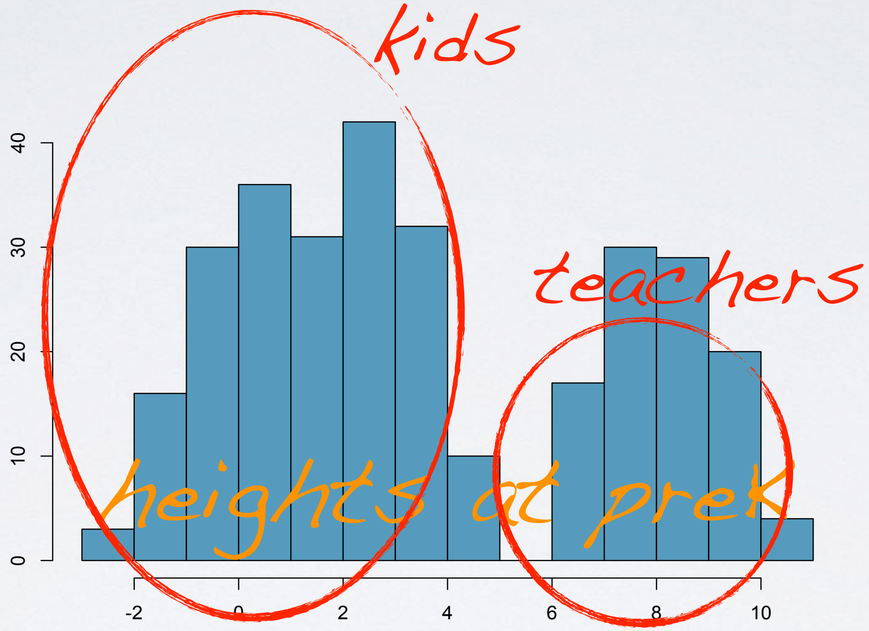
distributions are skewed to the side of the long tail

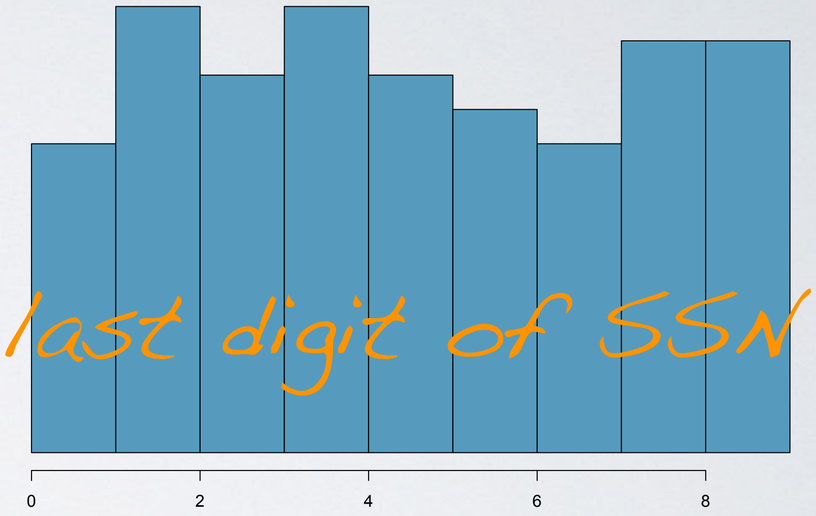


modality

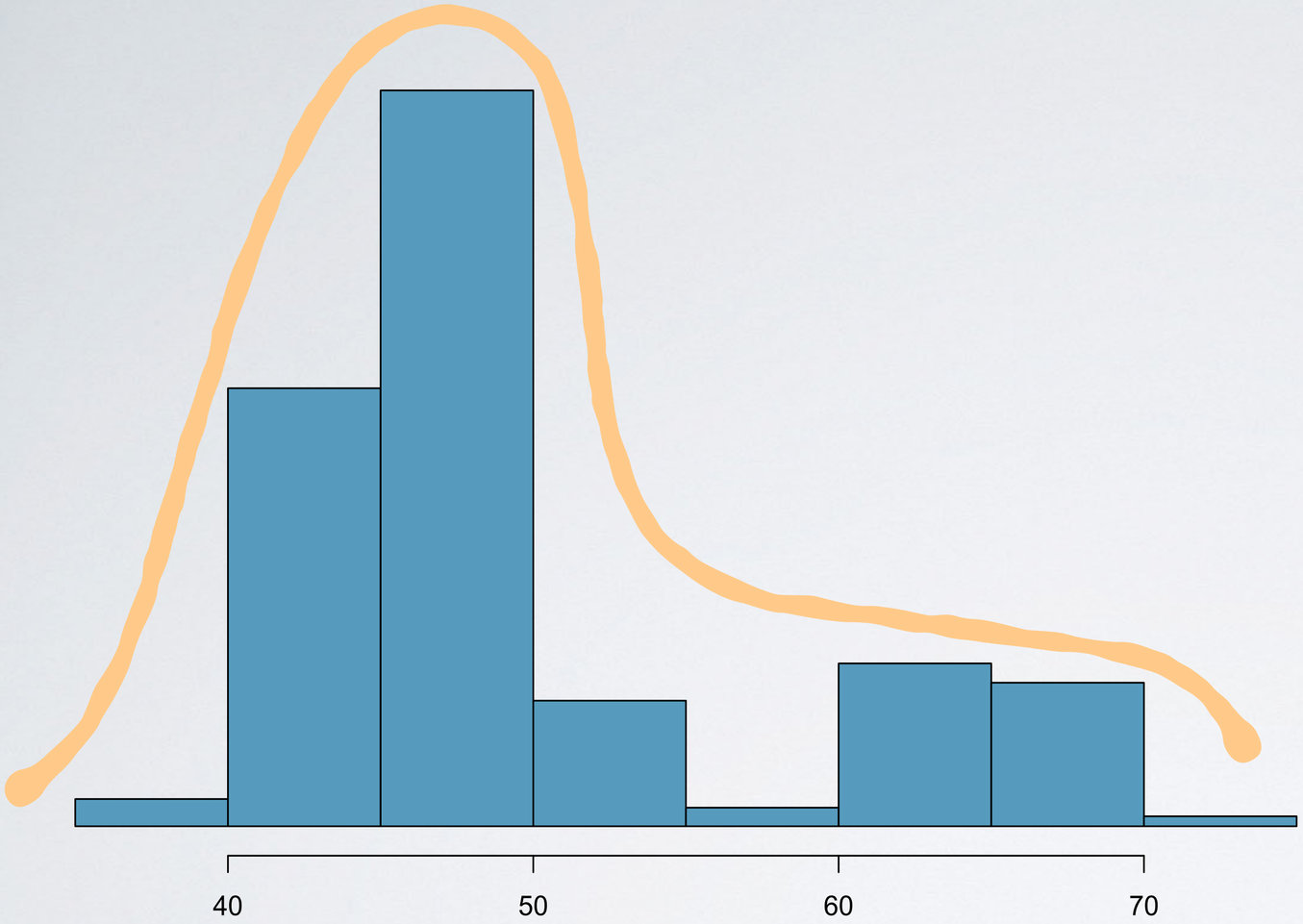


modality (cont.)



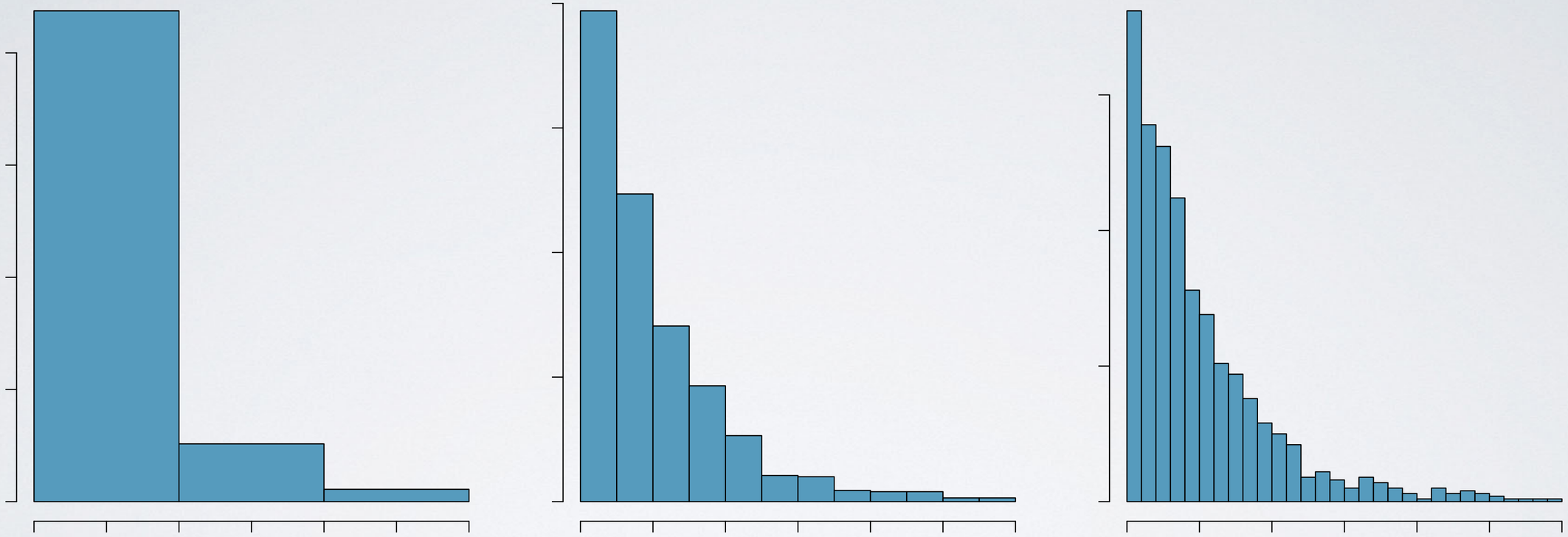


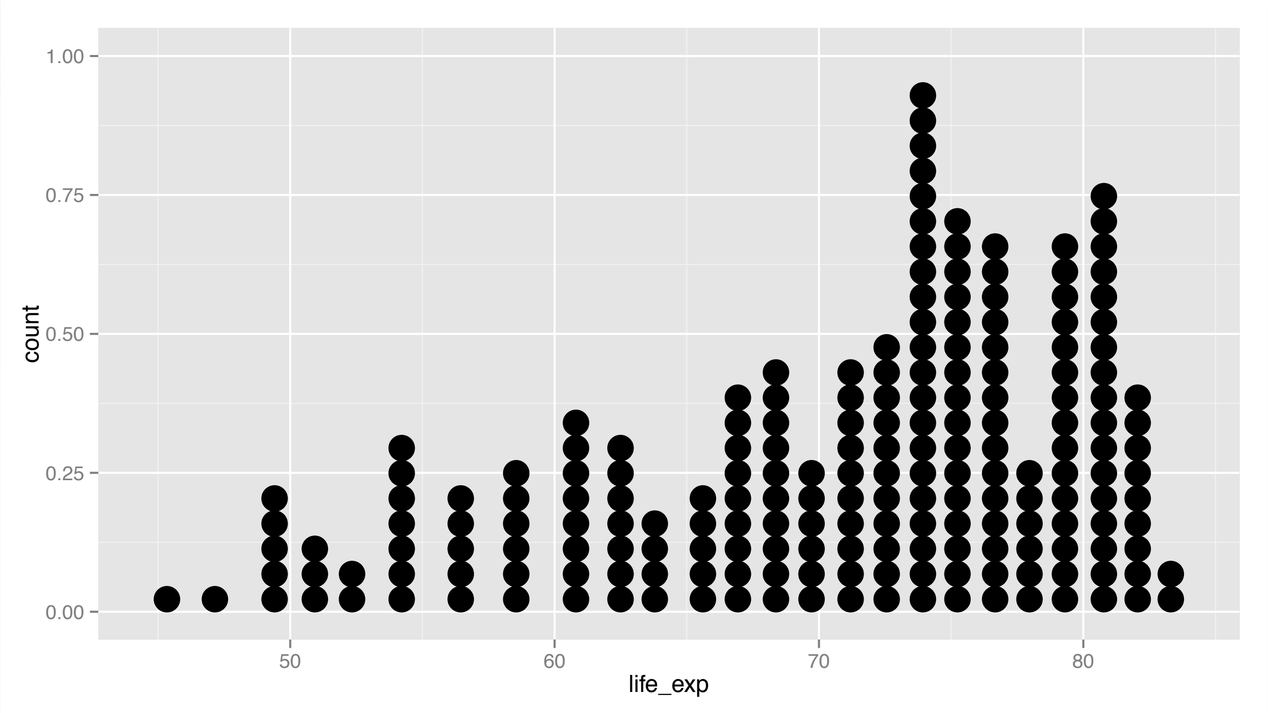
modality (cont.)



histogram & bin width

The chosen bin width can alter the story the histogram is telling.



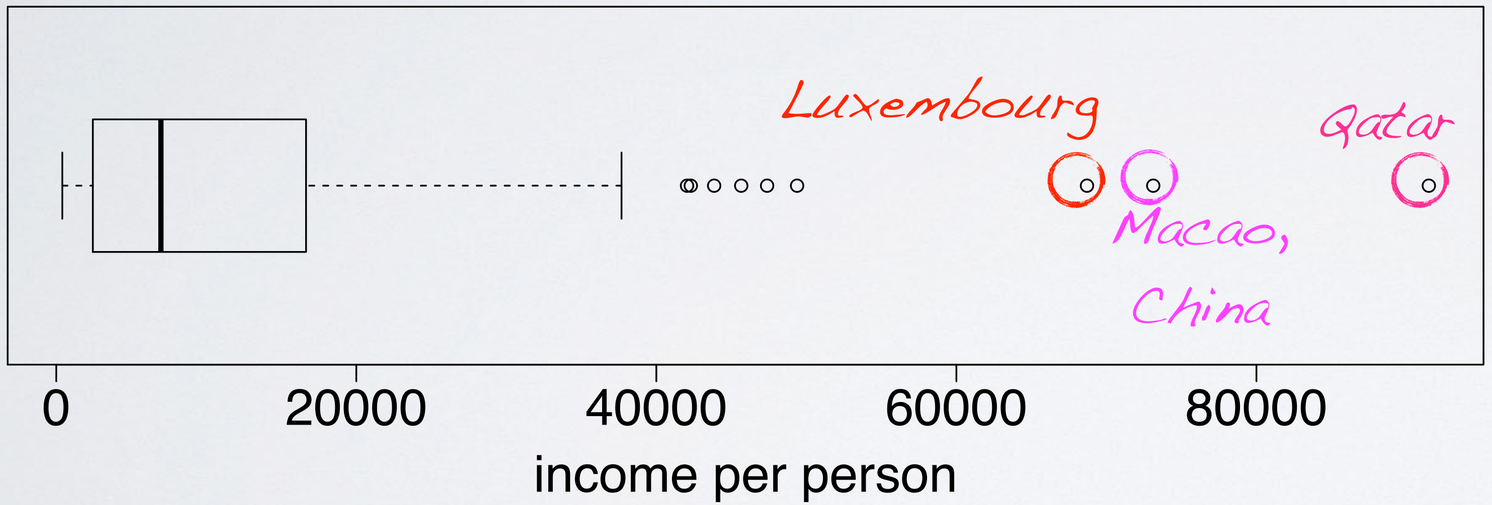


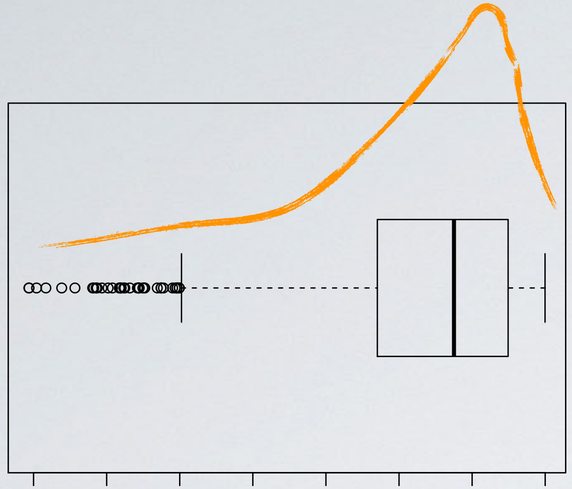
dotplot

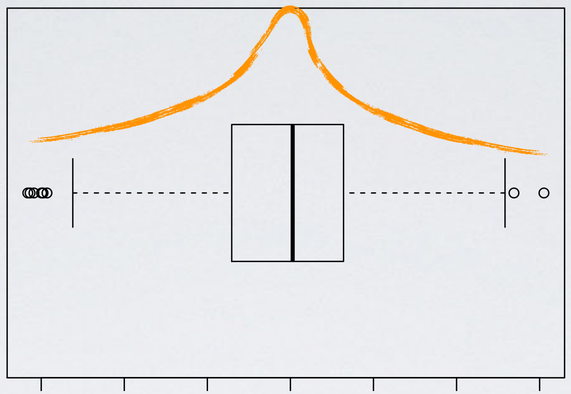
- useful when individual values are of interest
- can get busy as the sample size increases

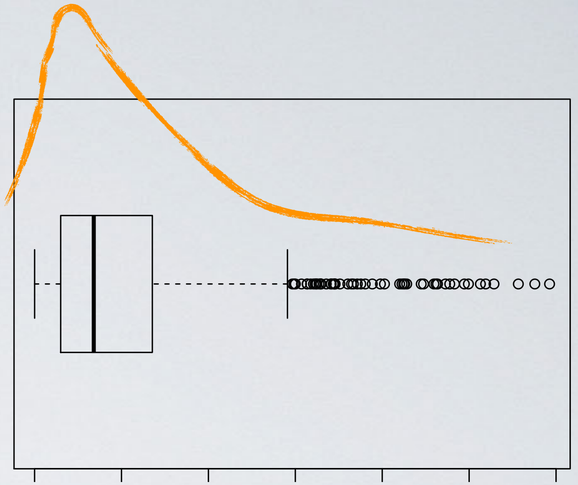
box plot

useful for highlighting outliers, median, IQR



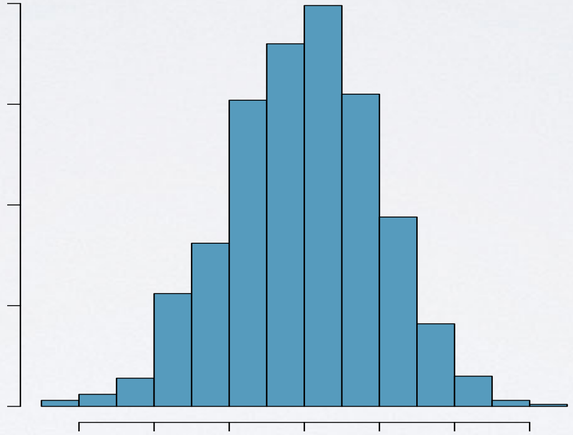




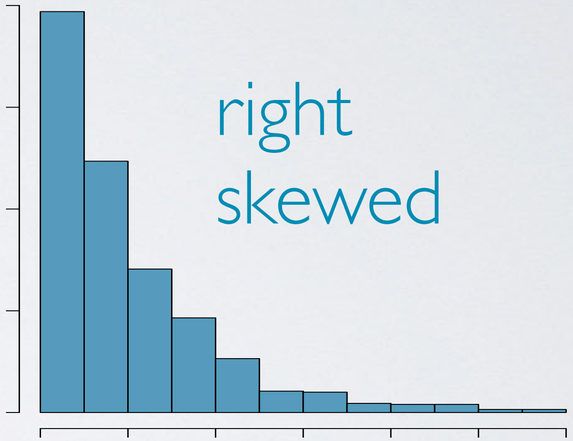


left
skewed

symmetric



right
skewed



intensity map

- Useful for highlighting the spatial distribution.