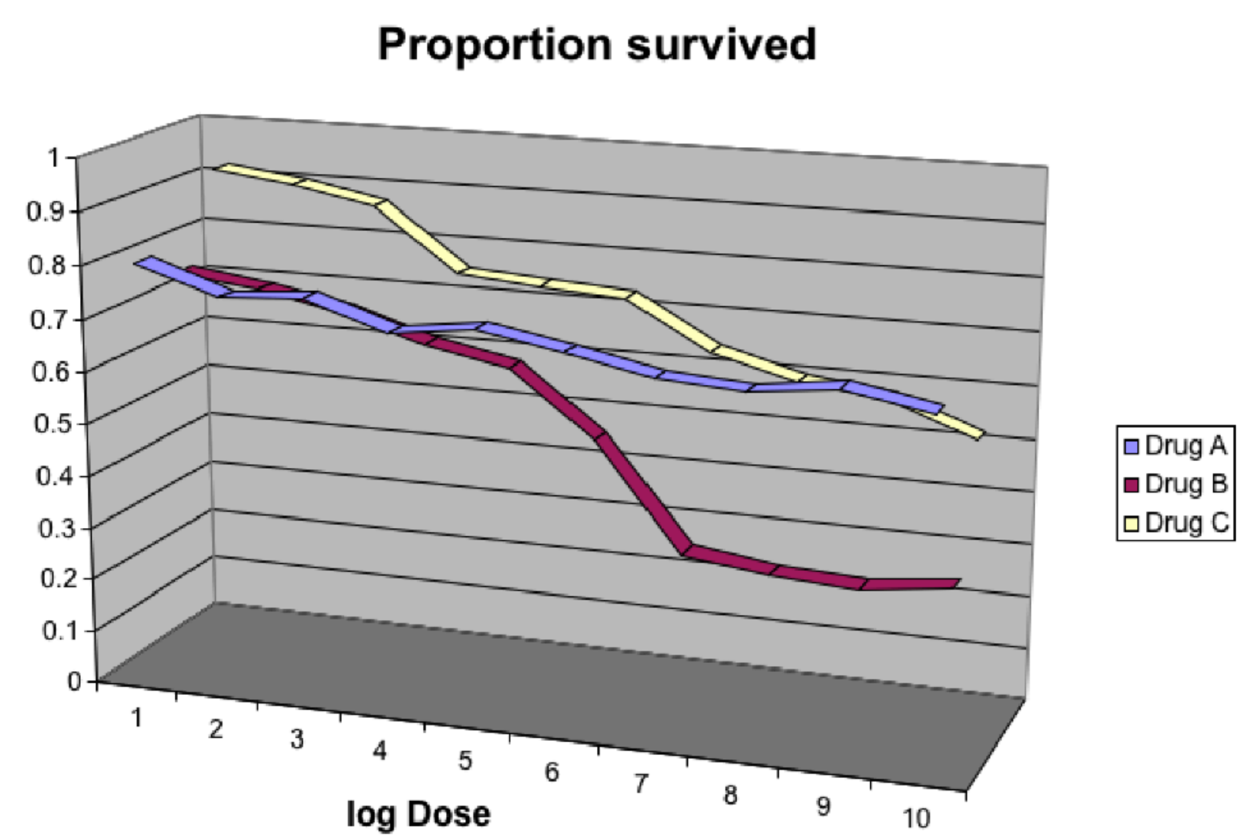
Introduction to R for data analysis

- plots -

Carl Herrmann & Carlos Ramirez
IRTG Course - December 2021

Graphical representation

- Appropriate graphical representation depends on the type of data
  - categorical
  - counts
  - continuous data
- Aim of good data graphics: **display data accurately and clearl**y (Karl Broman https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/)
- **Bad practice:**
  - as little information as possible
  - make things obscure through inappropriate graphics
  - pseudo 3D
  - poor scales



Example of bad plot

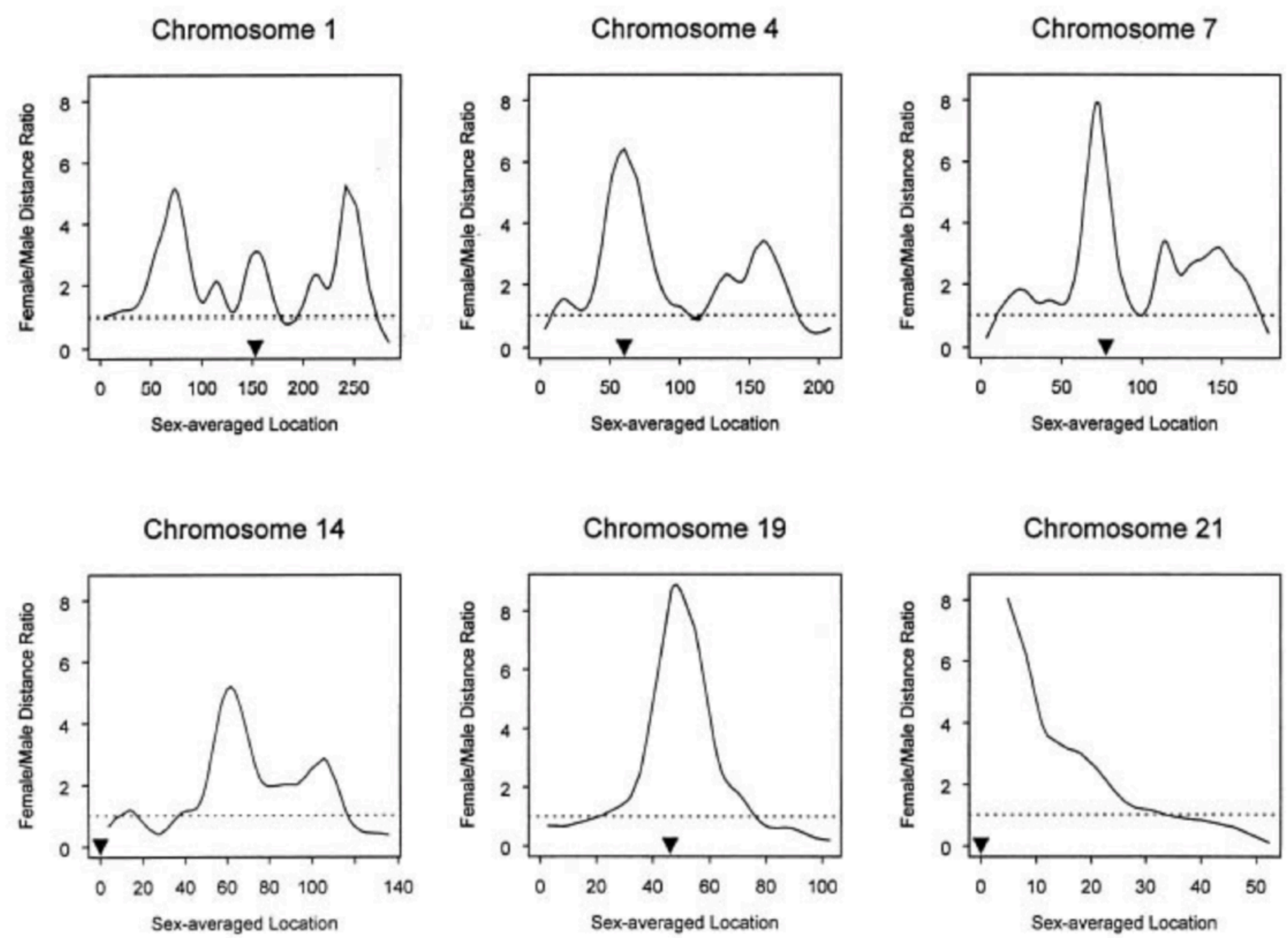


Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

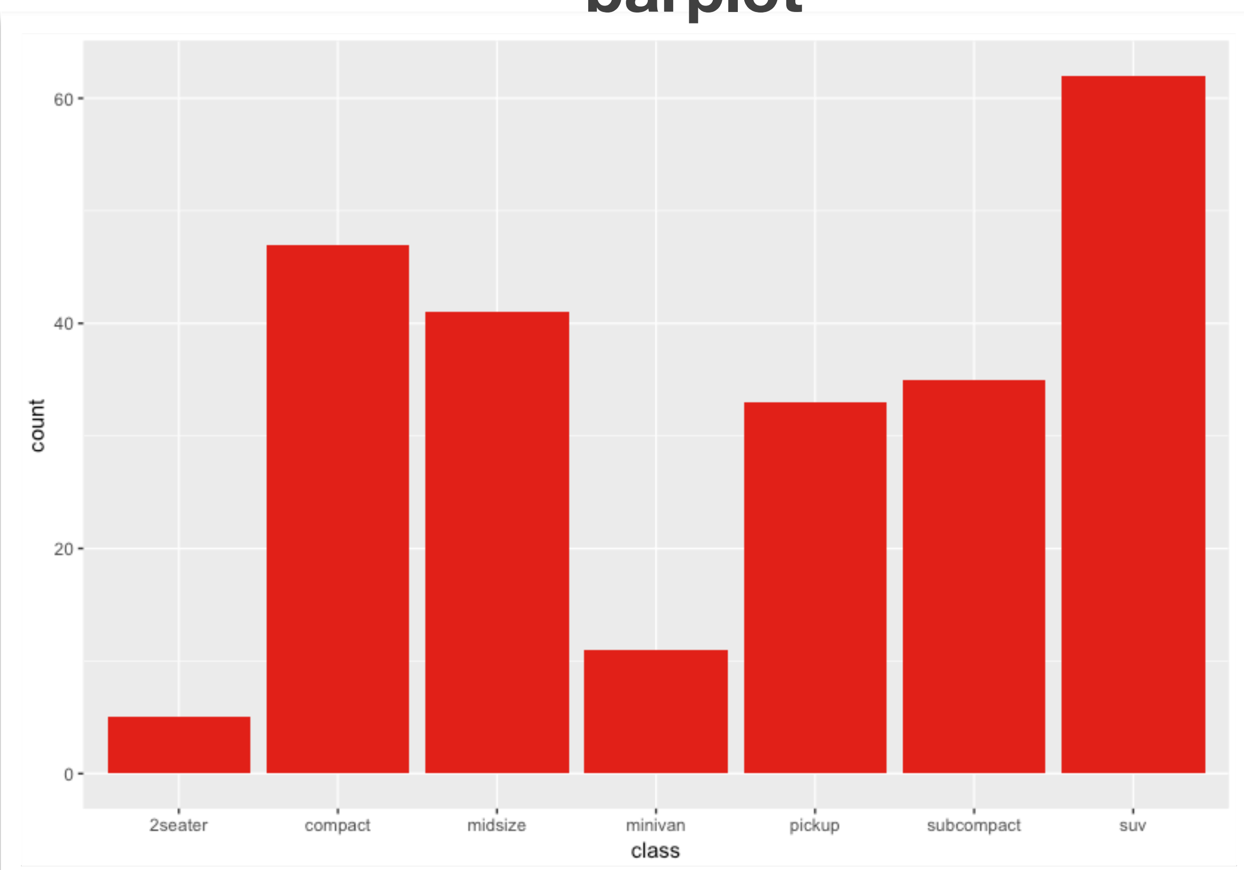
What's wrong with this plot?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/)

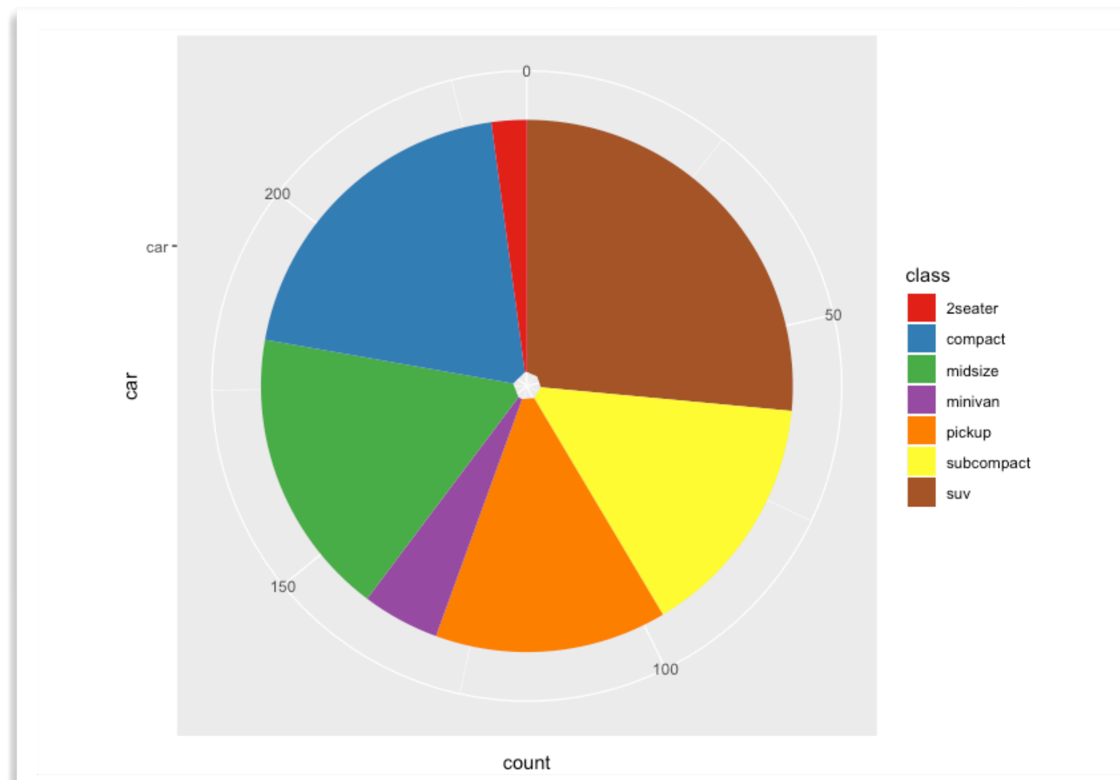
Categorical Data
Barplots

- How many instances in each category?
- Only meaningfull measure: **MODE** (= category with highest counts)
- Possible plots: **barplots**; **piecharts**

barplot



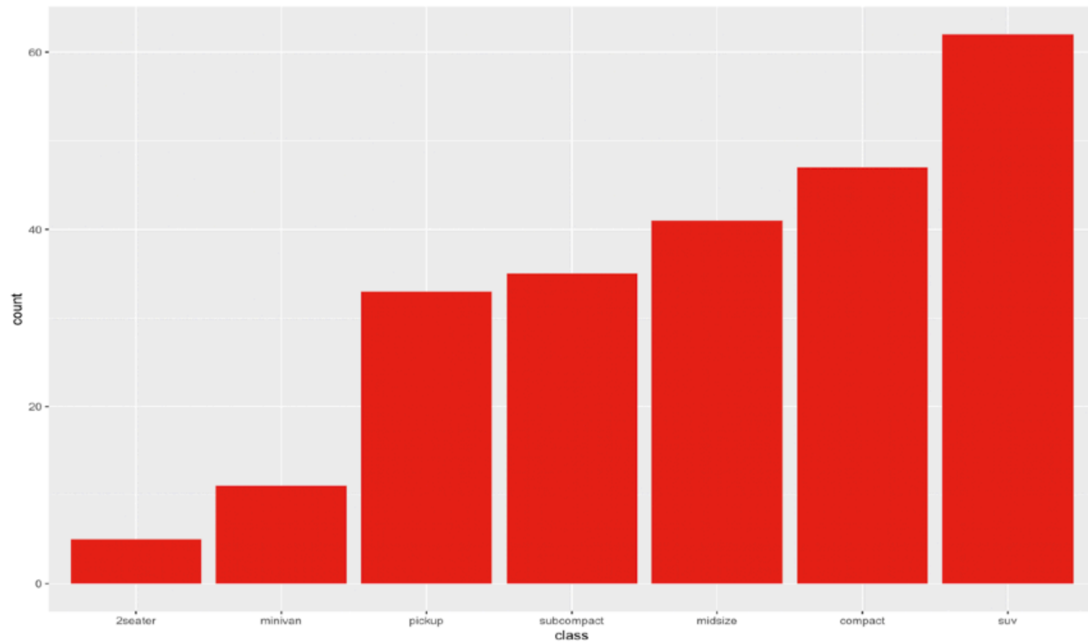
piechart

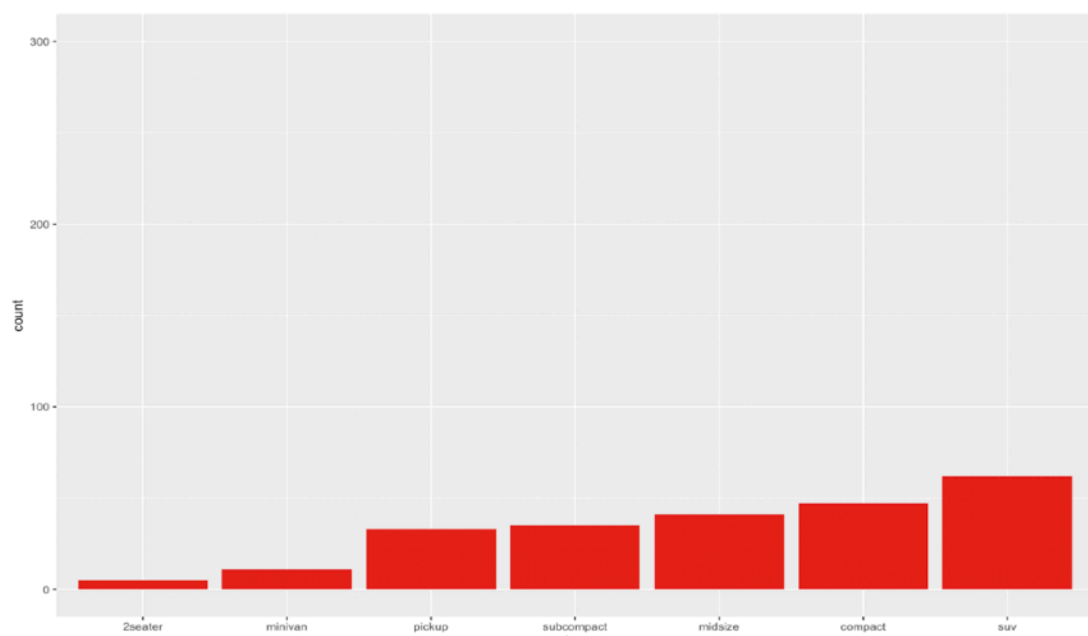


Avoid piechart : areas are more difficult to judge than length!

Categorical Data

Barplots



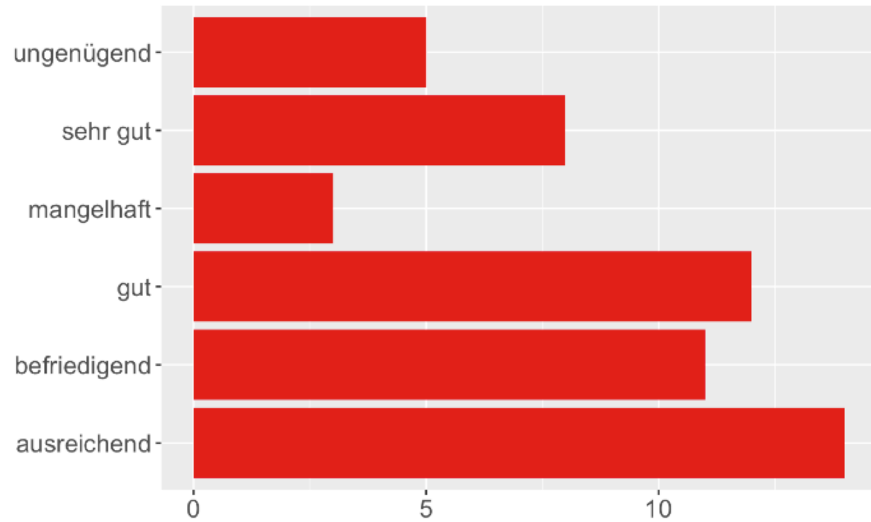


- Consider ordering the data by counts (no natural order of categories for nominal data)

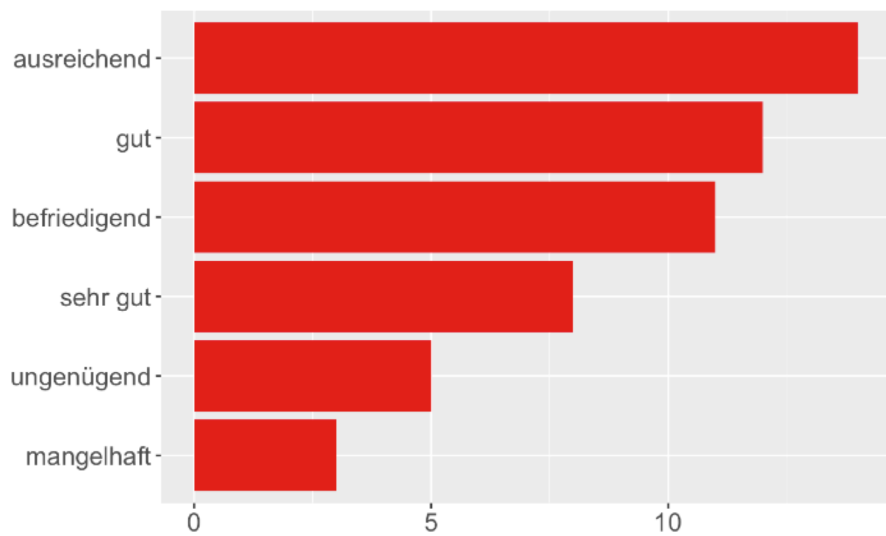
- Beware of selecting the proper scales for plotting!

Categorical Data
Barplots

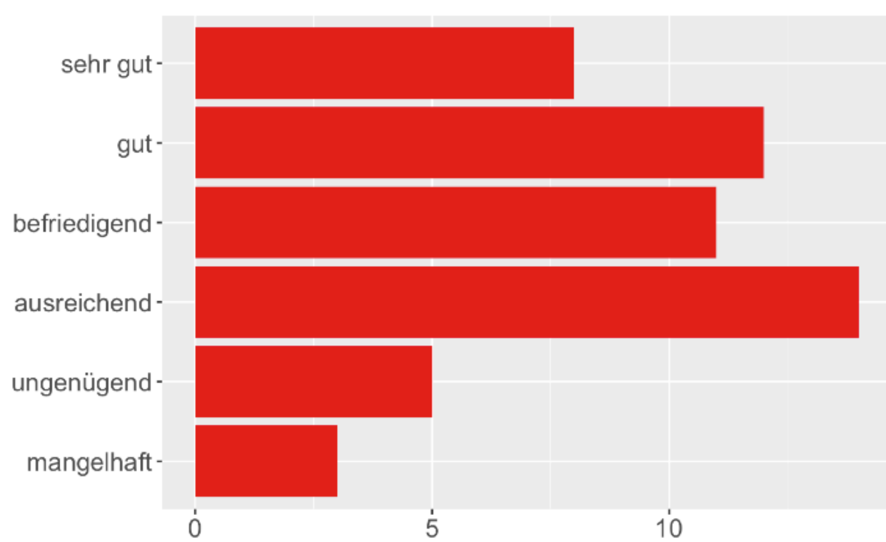
Medizinische Fakultät Heidelberg



- Random order



- Order by increasing / decreasing counts



- Natural order of ordinal data

Mode

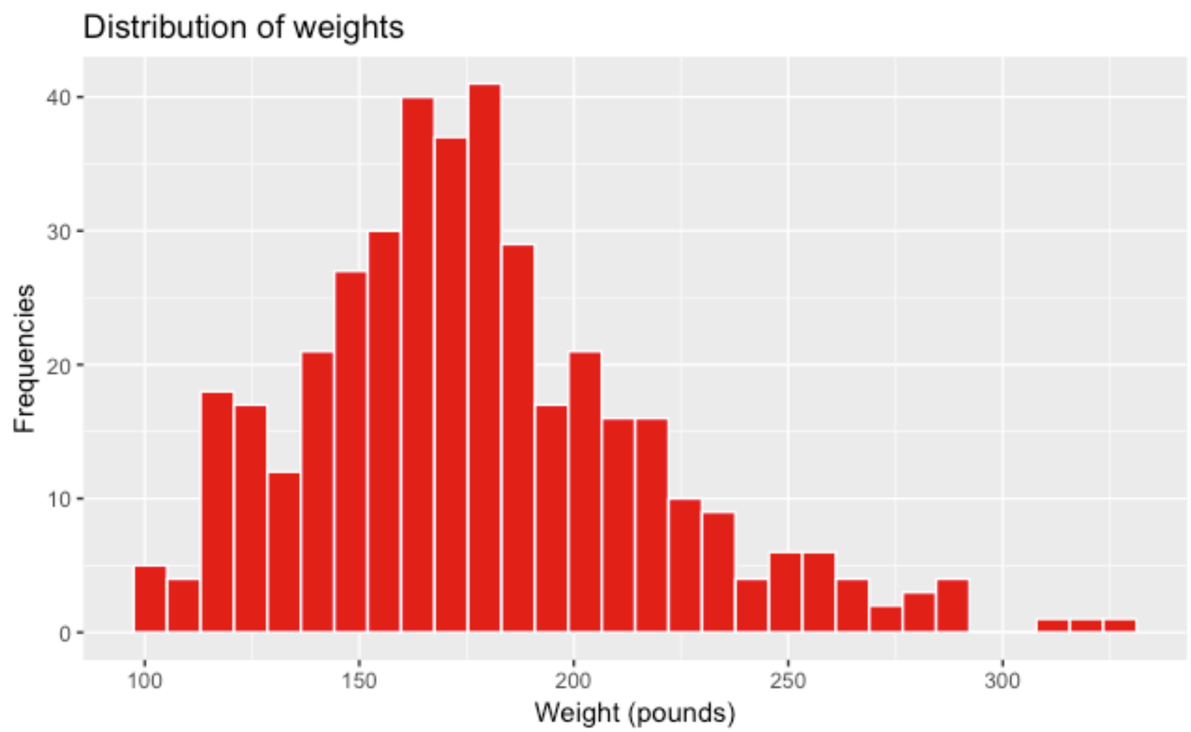
Numerical variables

- Numerical data are instances of underlying **random variable**

- Random variables X have
  - **Density** distributions $p(X)$
  - **Expectation values** $E(X)$
  - **Variances** $Var(X)$

Numerical data
Histograms

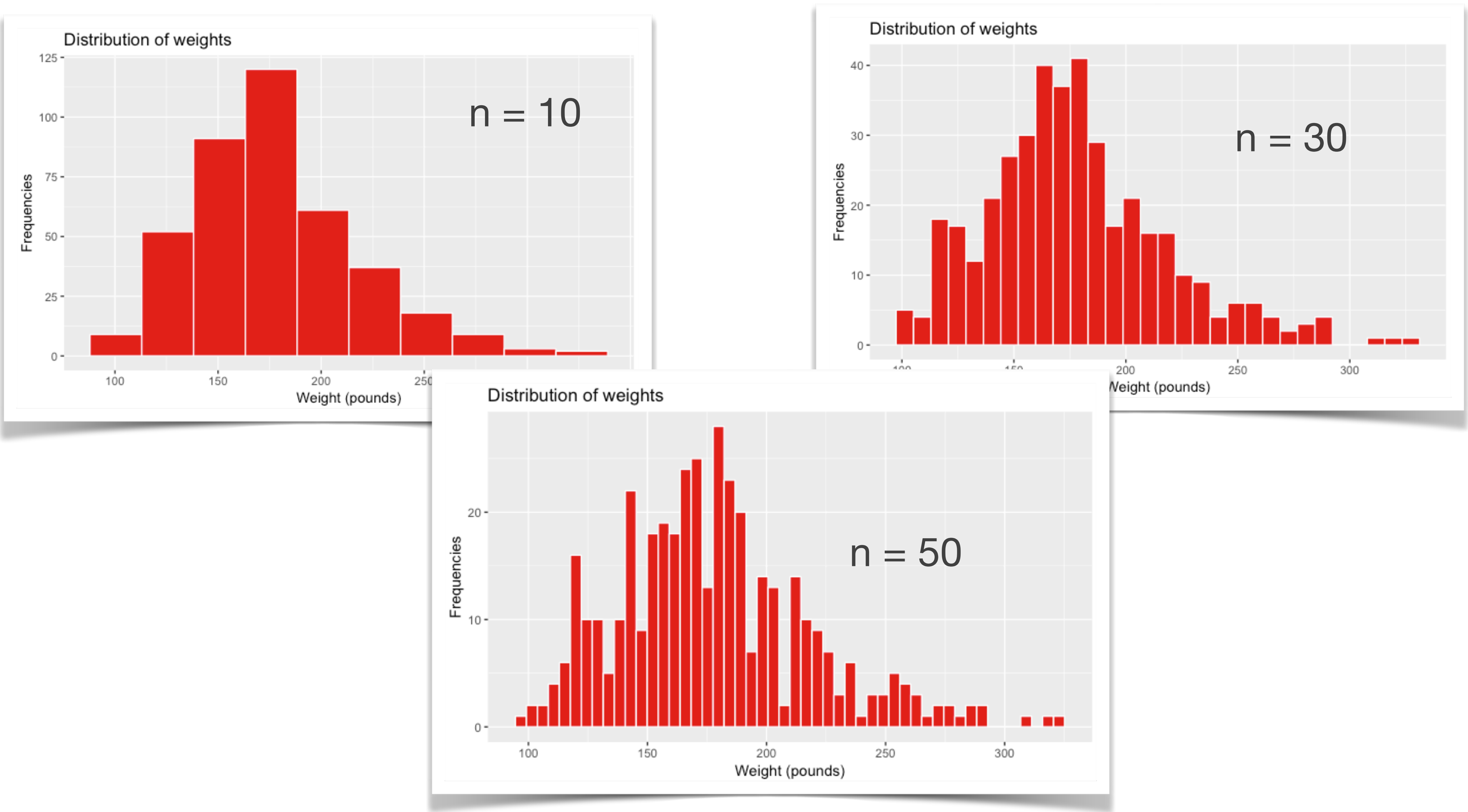
- Categorical data → counts (**barplot**)
- Numerical data → counts within intervals (**histogram**)
  - define **discrete intervals** for numerical variable → **ordinal variable**
    e.g. [0,10), [10,20), [20,30), …
  - count occurences within intervals and plot
- histograms represent the **distribution of the variable**



Numerical data
Histograms

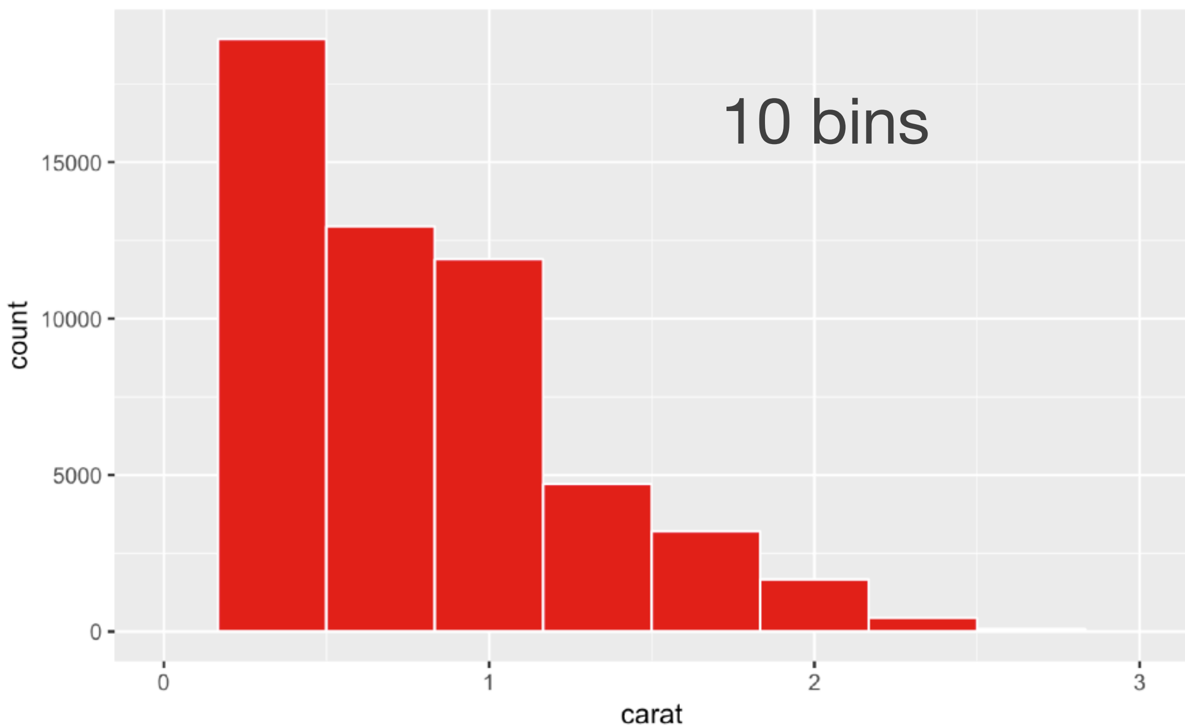
Medizinische Fakultät Heidelberg

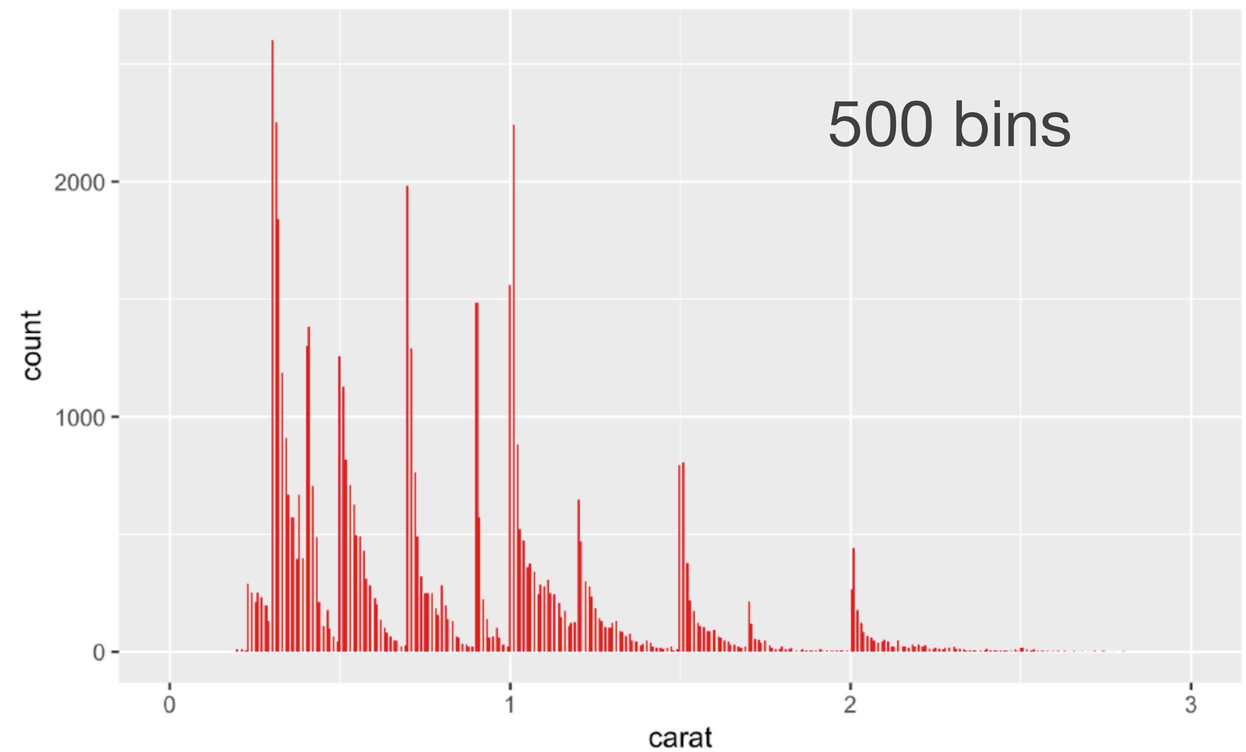
- Right choice of interval depends on the data type
- **Number of bins has a strong impact on appearance of plot!**



Numerical data
Histograms

Distribution of carat values for diamonds

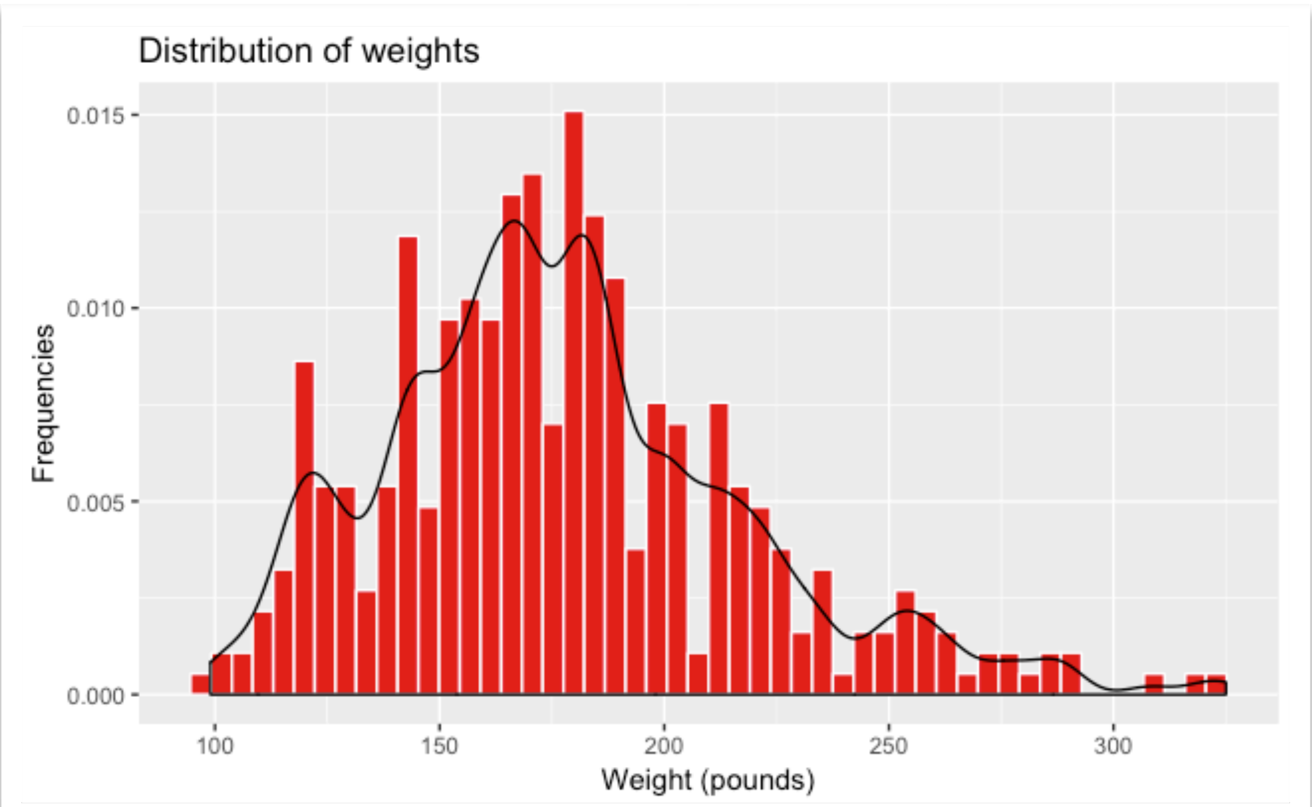




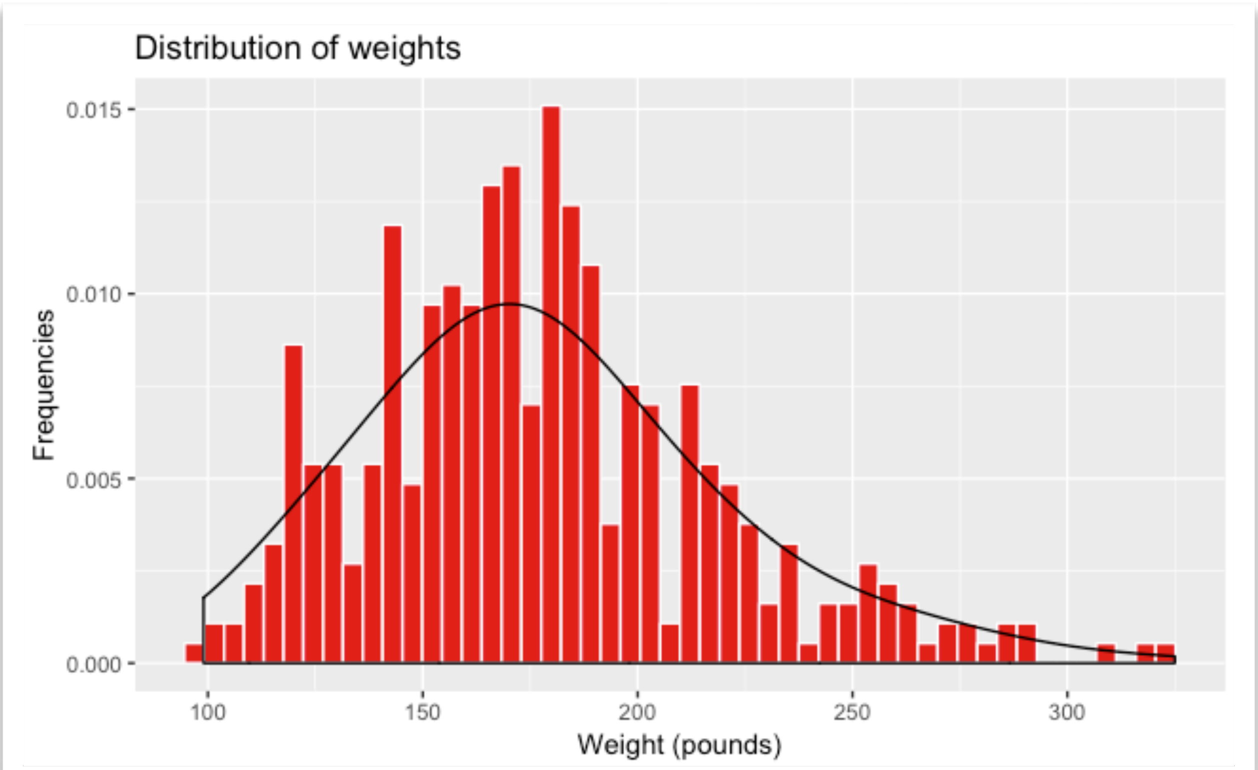
- pattern becomes visible at high resolution
- peaks around integer values (why?)
- tail on the right of integer values (why?)

Numerical data
Histograms

- Frequency distributions (= histograms) can be shown using a smoothed **density curve**
- Smoothing depends on the bandwidth (~size of the interval over which to smooth)



bandwith = 5

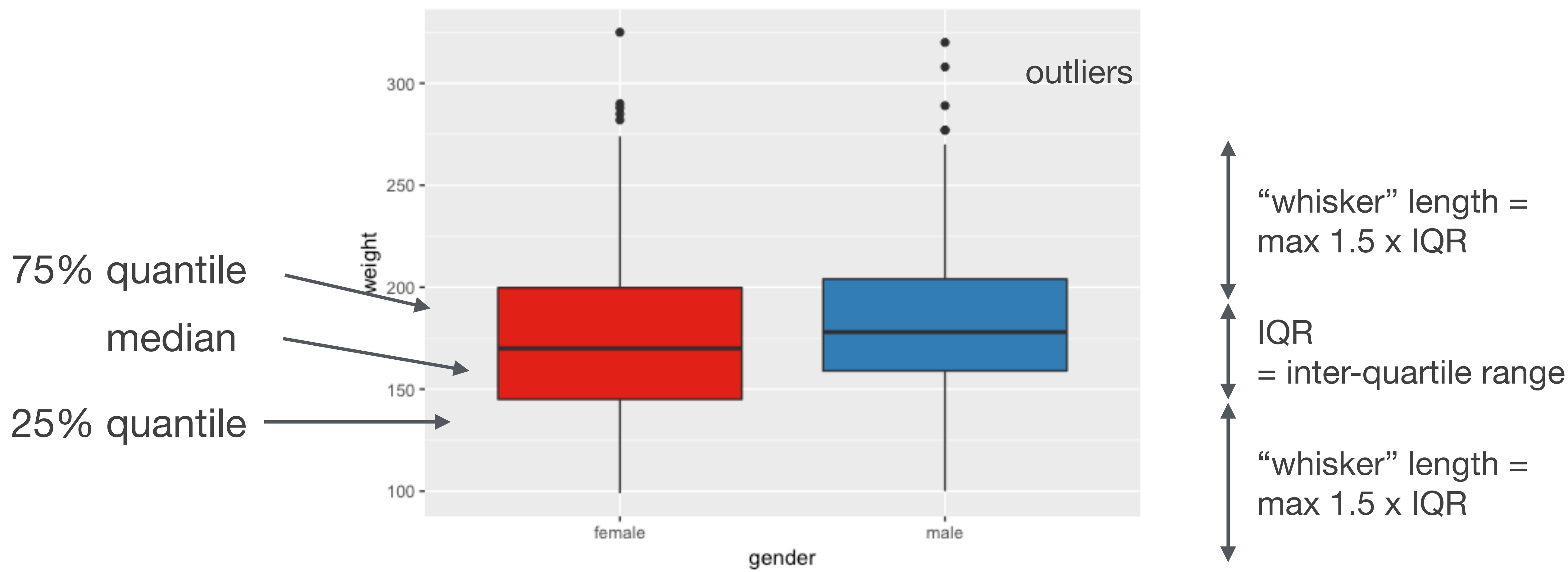


bandwith = 20

Numerical values

boxplots

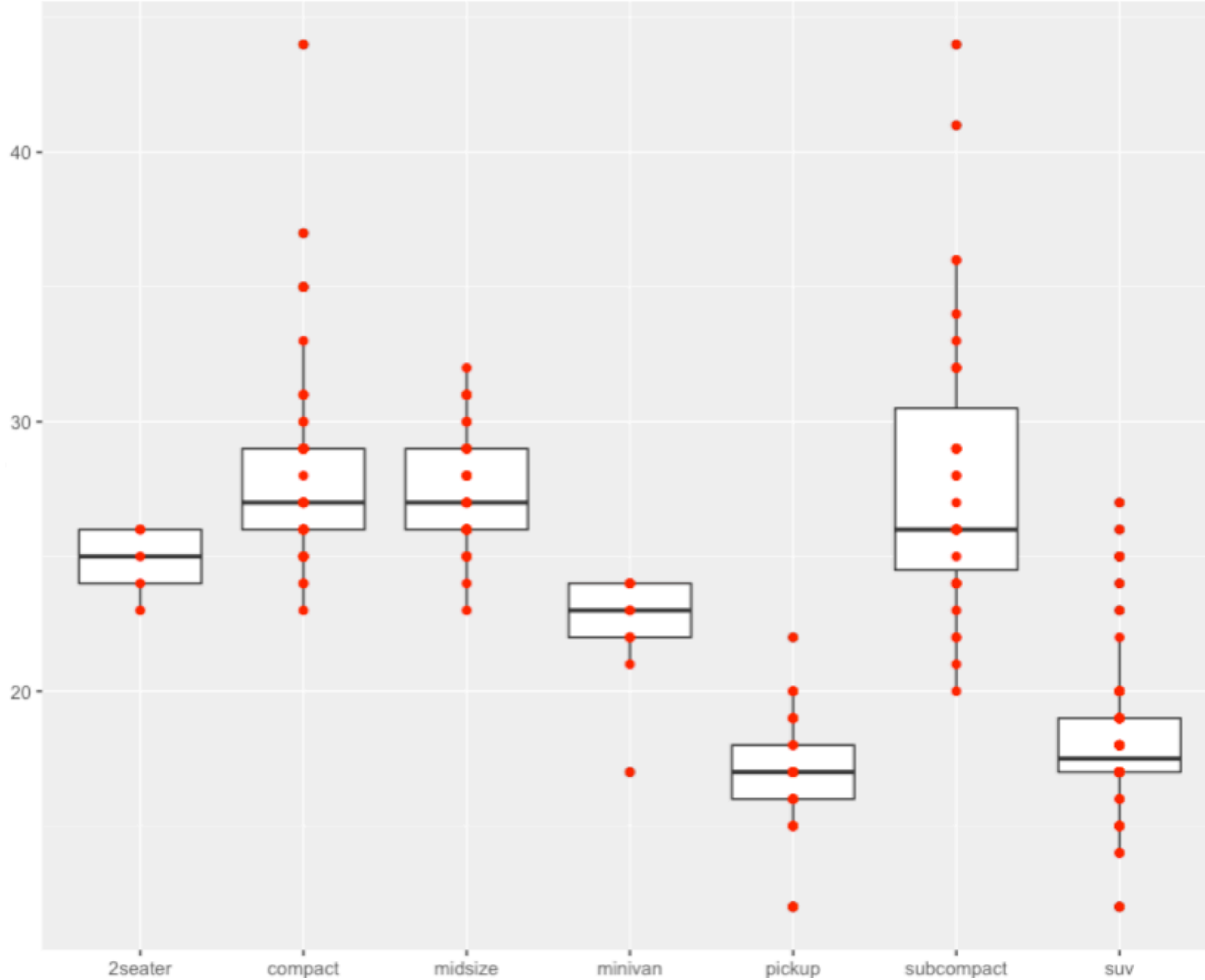
- Boxplot give an indication on the shape of the distribution (median / symmetry / outlier / …)



- Upper Whisker extend to the last point that is not larger than Q75 + 1.5*IQR
- Lower Whisker extends to the last point that is not smaller than Q25-1.5*IQR
- Whisker does not go not beyond maximun or minimum value ! (Hence both whisker can have different length < 1.5 x IQR)

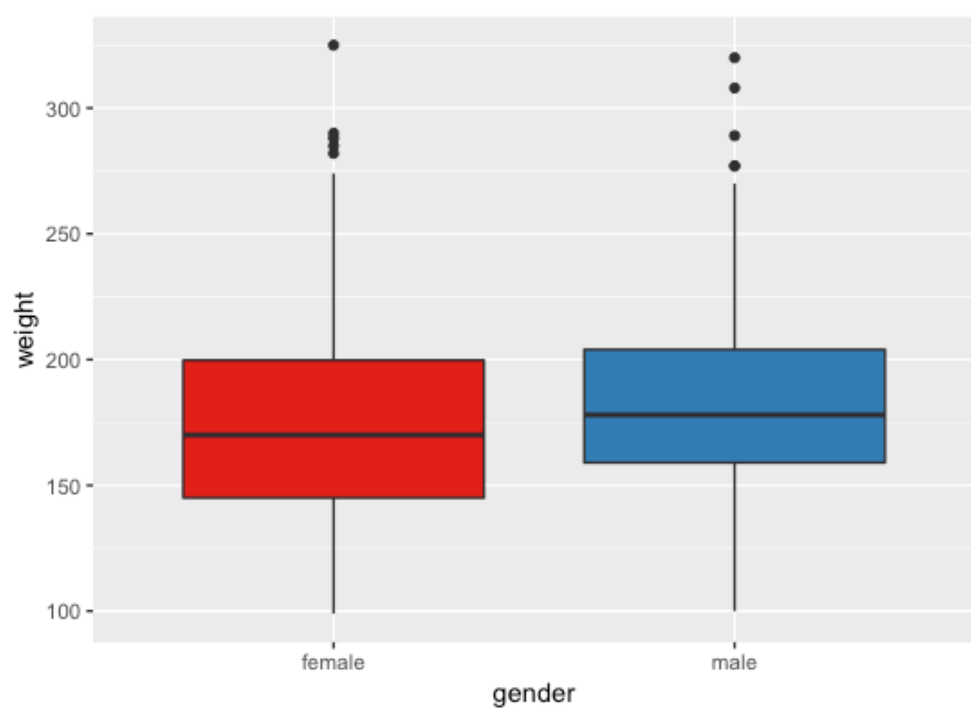
Numerical values

boxplots

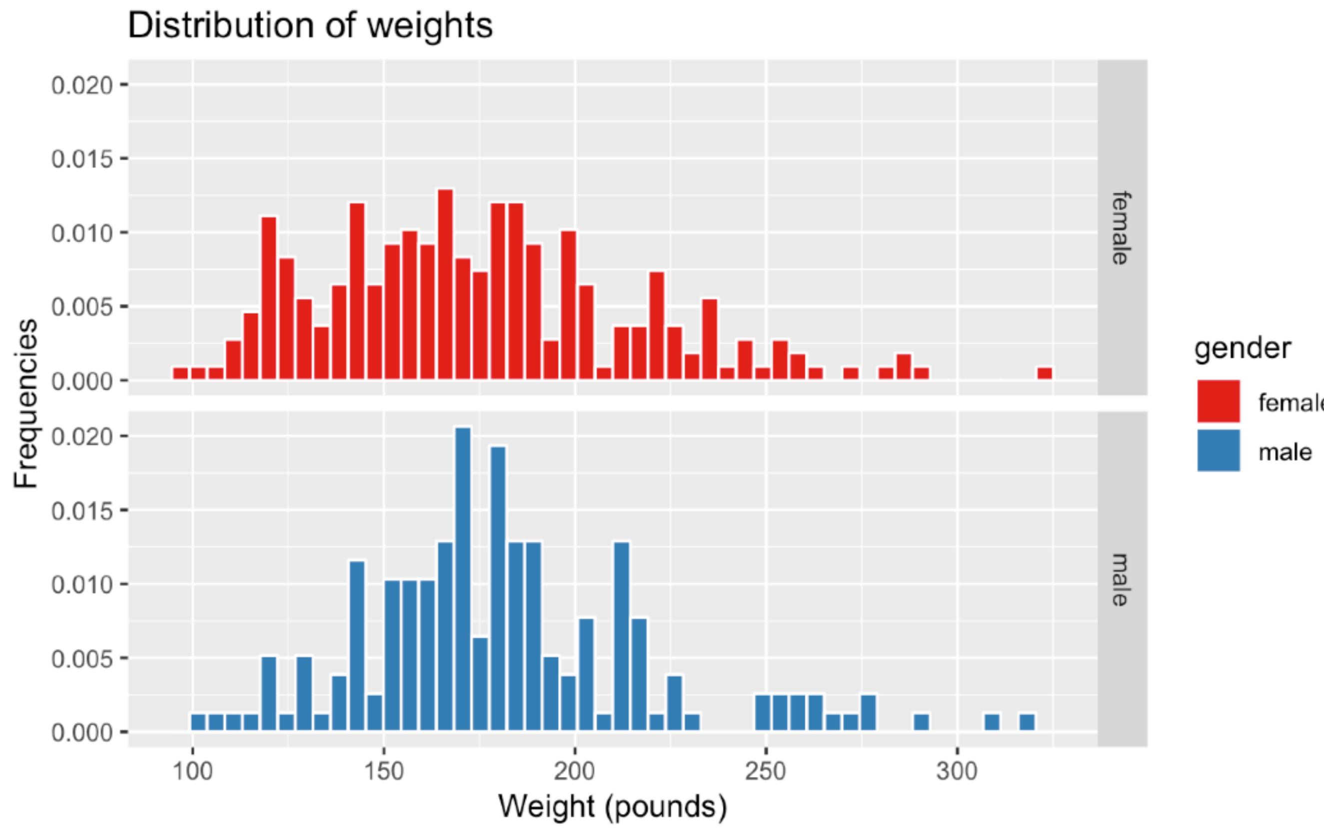


Numerical values
boxplots

Boxplot

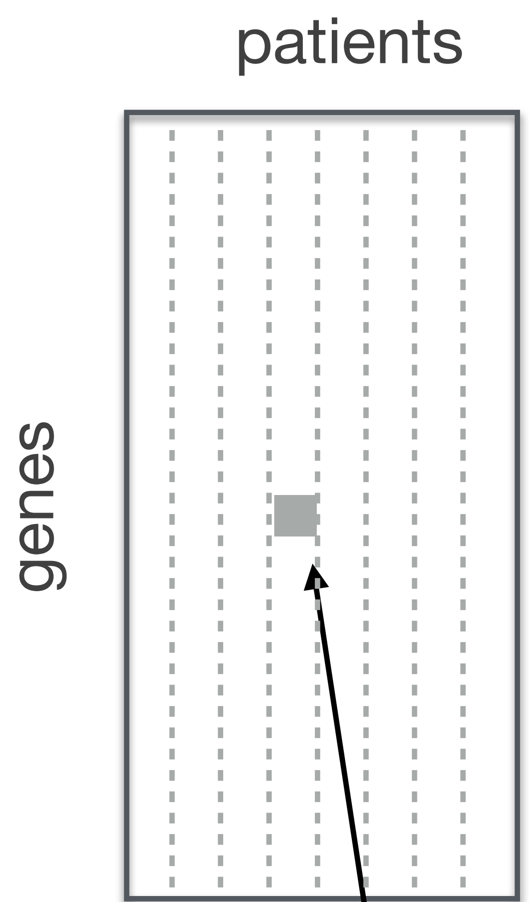


Histogram

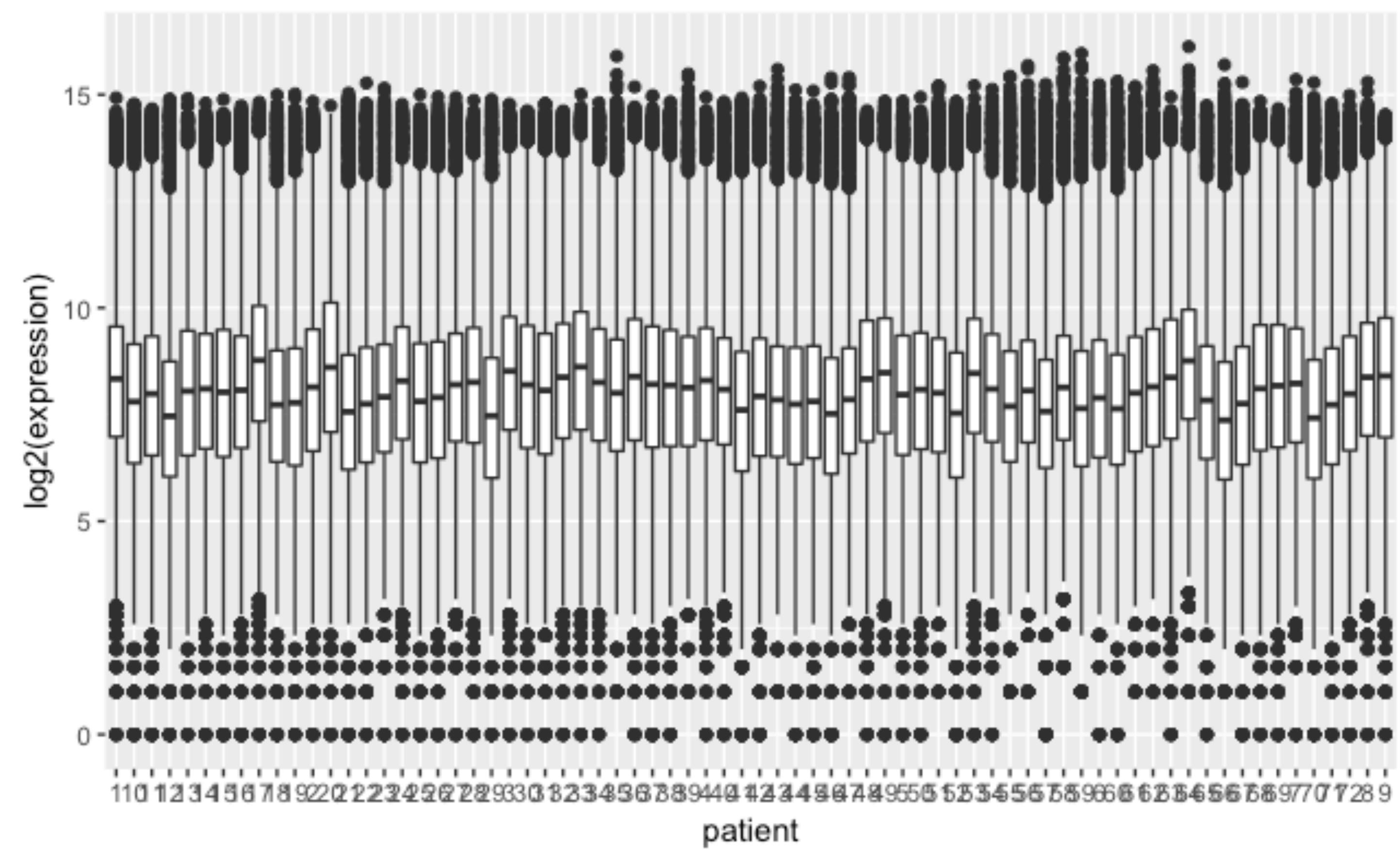


Boxplots summarize the properties of the distribution
Usefull to compare many distributions side-by-side

Numerical values
boxplots



Expression of gene *i*
in patient *j*



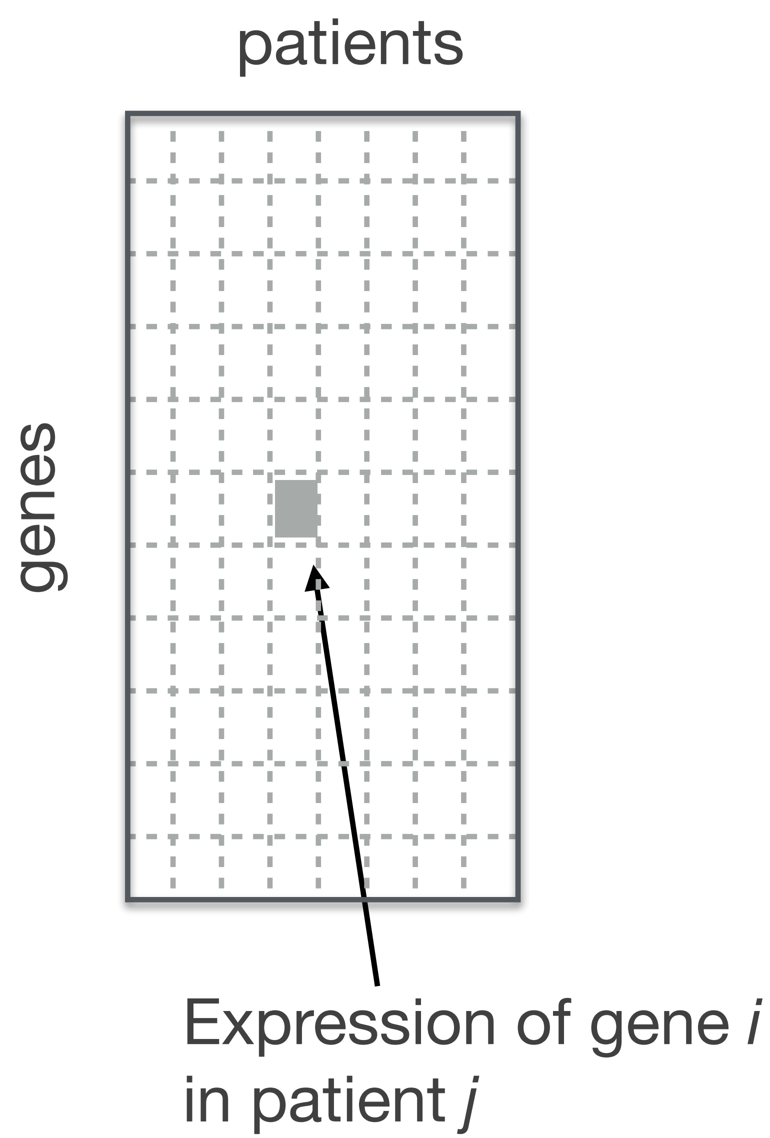
Question: do some patients have
a different median gene expression?

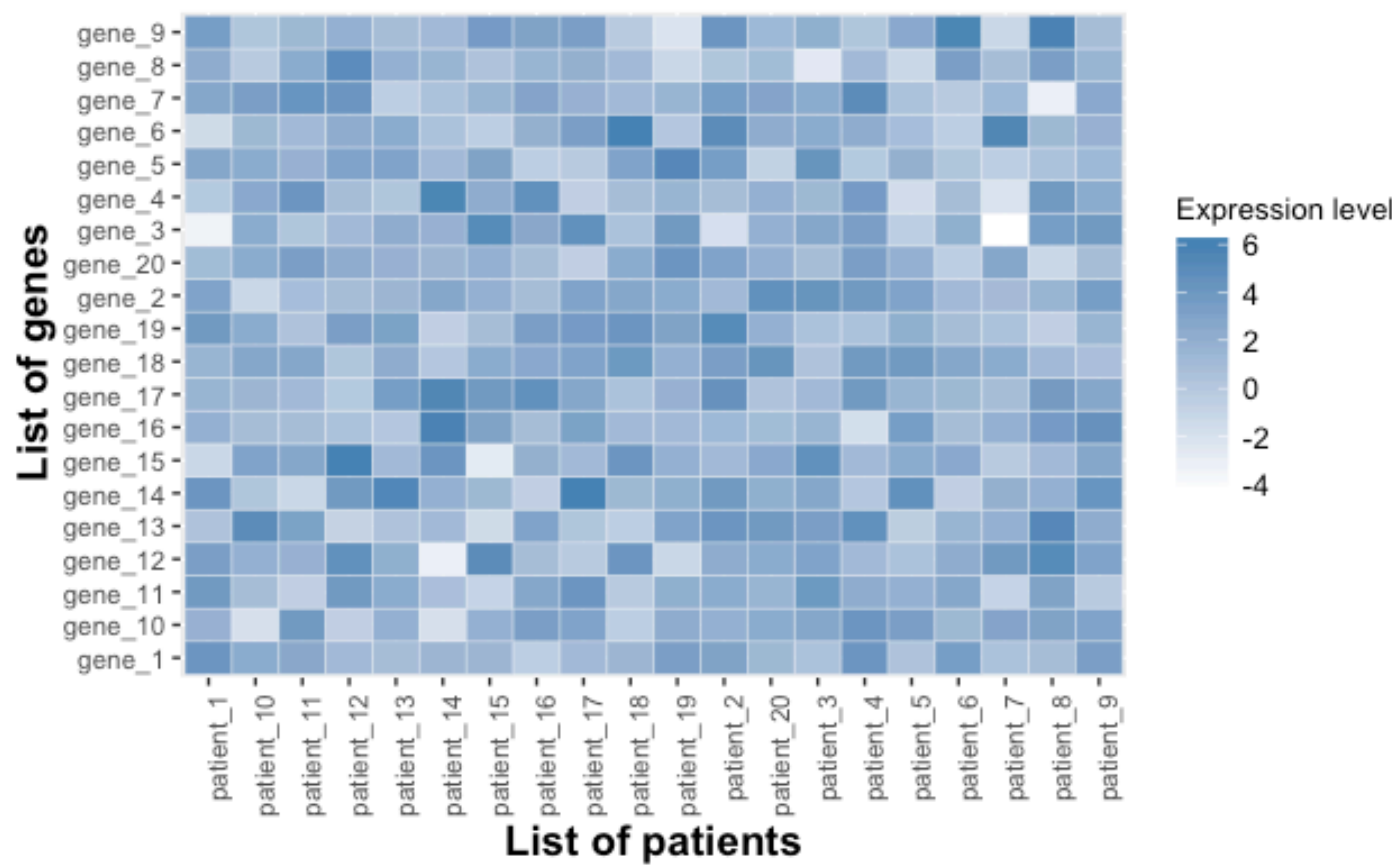
→ *values for individual genes are lost in this type of plot!*

Numerical Data :
heatmaps

Medizinische Fakultät Heidelberg

Heatmaps display numerical values in a data matrix using a color scheme

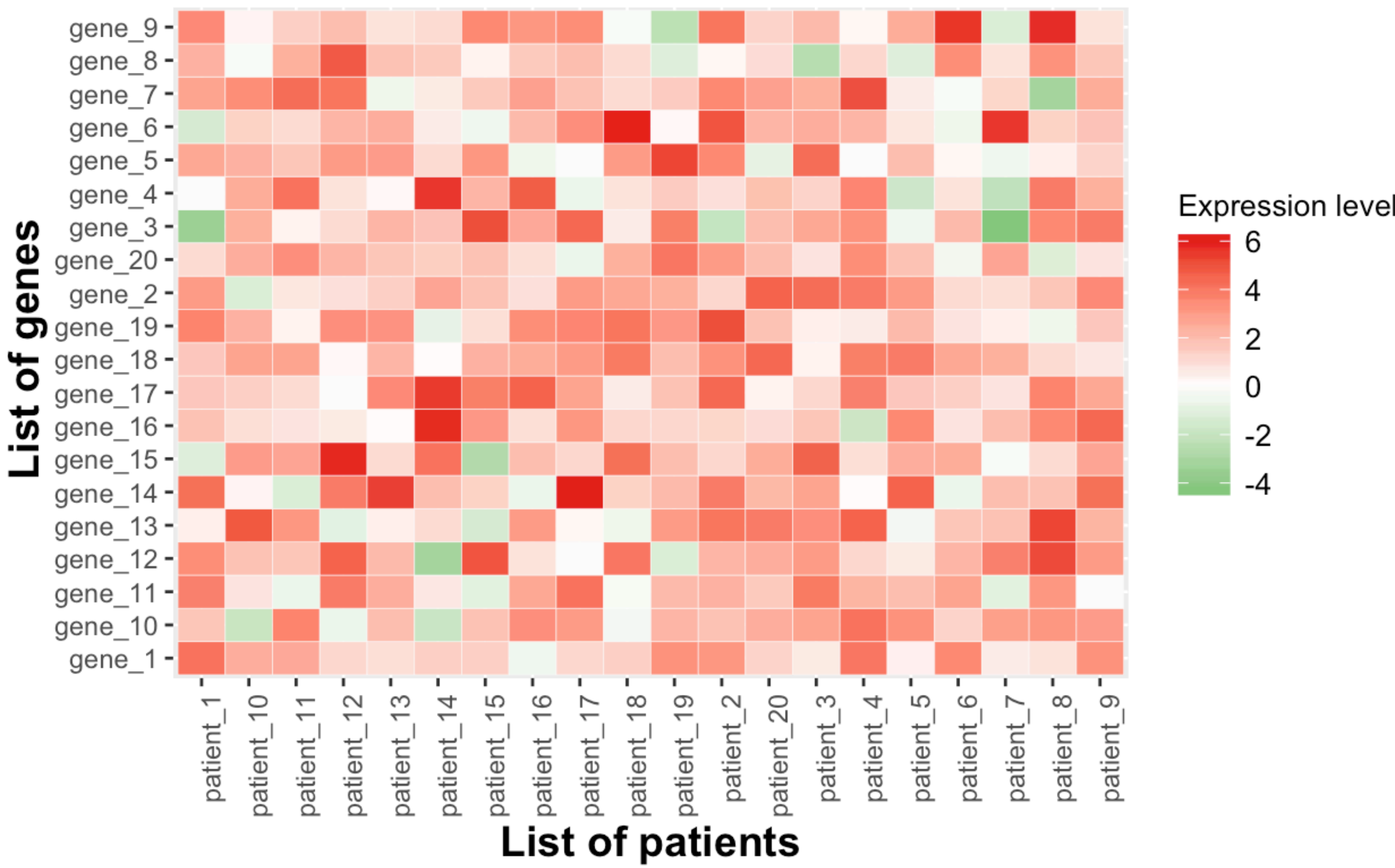




Question: do some genes in some patients have a different gene expression?

Numerical Data :
heatmaps

use symmetrical color scales for symmetrical ranges



Numerical data

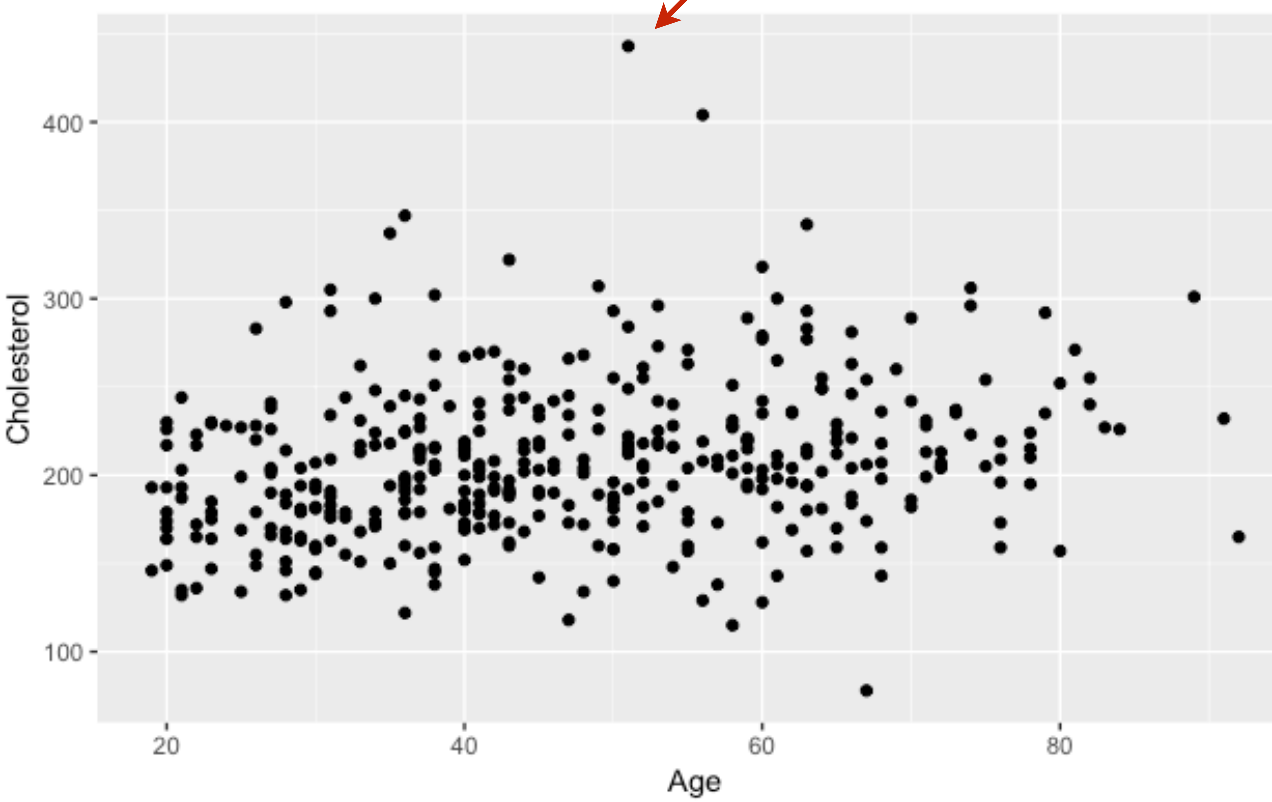
comparing variable with scatter plots

any relation between age and cholesterol?

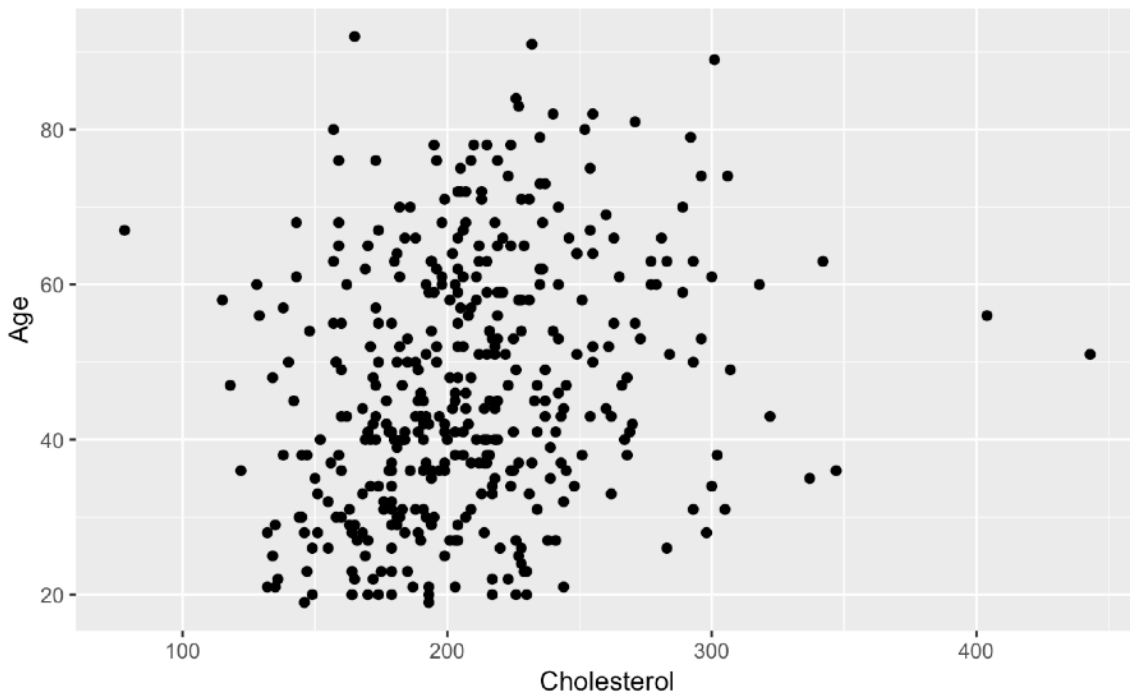
id	chol	stab.glu	hdl	ratio	glyhb	location	age
1000	203	82	56	3.60	4.31	Buckingham	46
1001	165	97	24	6.90	4.44	Buckingham	29
1002	228	92	37	6.20	4.64	Buckingham	58
1003	78	93	12	6.50	4.63	Buckingham	67
1005	249	90	28	8.90	7.72	Buckingham	64
1008	248	94	69	3.60	4.81	Buckingham	34
1011	195	92	41	4.80	4.84	Buckingham	30
1015	227	75	44	5.20	3.94	Buckingham	37
1016	177	87	49	3.60	4.84	Buckingham	45
1022	263	89	40	6.60	5.78	Buckingham	55
1024	242	82	54	4.50	4.77	Louisa	60
1029	215	128	34	6.30	4.97	Louisa	38
1030	238	75	36	6.60	4.47	Louisa	27
1031	183	79	46	4.00	4.59	Louisa	40
1035	191	76	30	6.40	4.67	Louisa	36
1036	213	83	47	4.50	3.41	Louisa	33
1037	255	78	38	6.70	4.33	Louisa	50

each dot is a patient

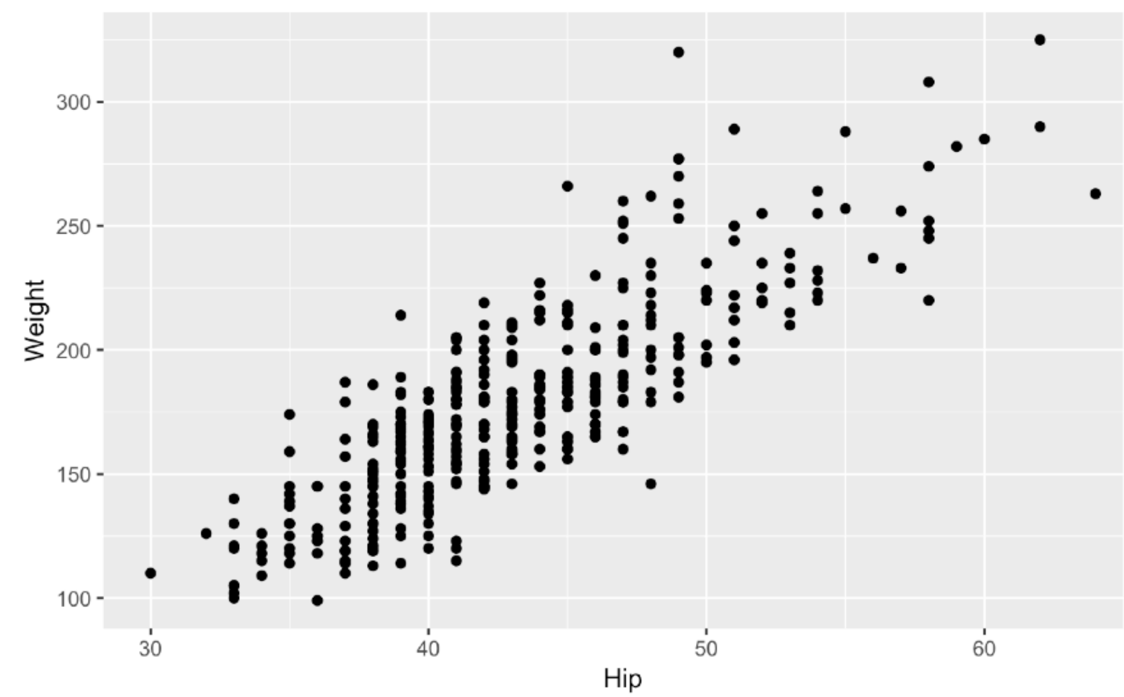
outlier



Numerical data
comparing variable with scatter plots



Weak relationship
between Cholesterol and age

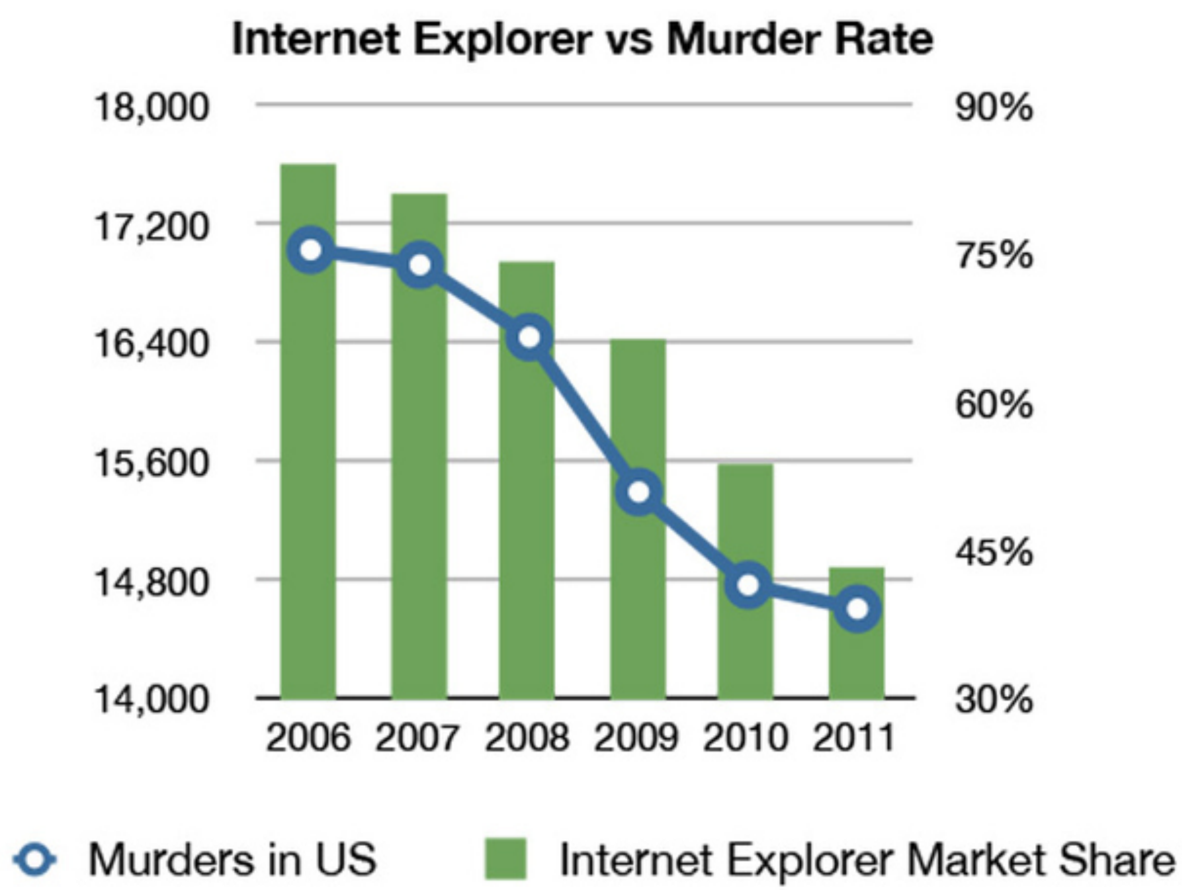


Strong relationship
between Hip and Weight

- we will later **quantify** this relationship in terms of **covariance / correlation** and determine how **significant** this relationship is!

Numerical data
comparing variable with scatter plots

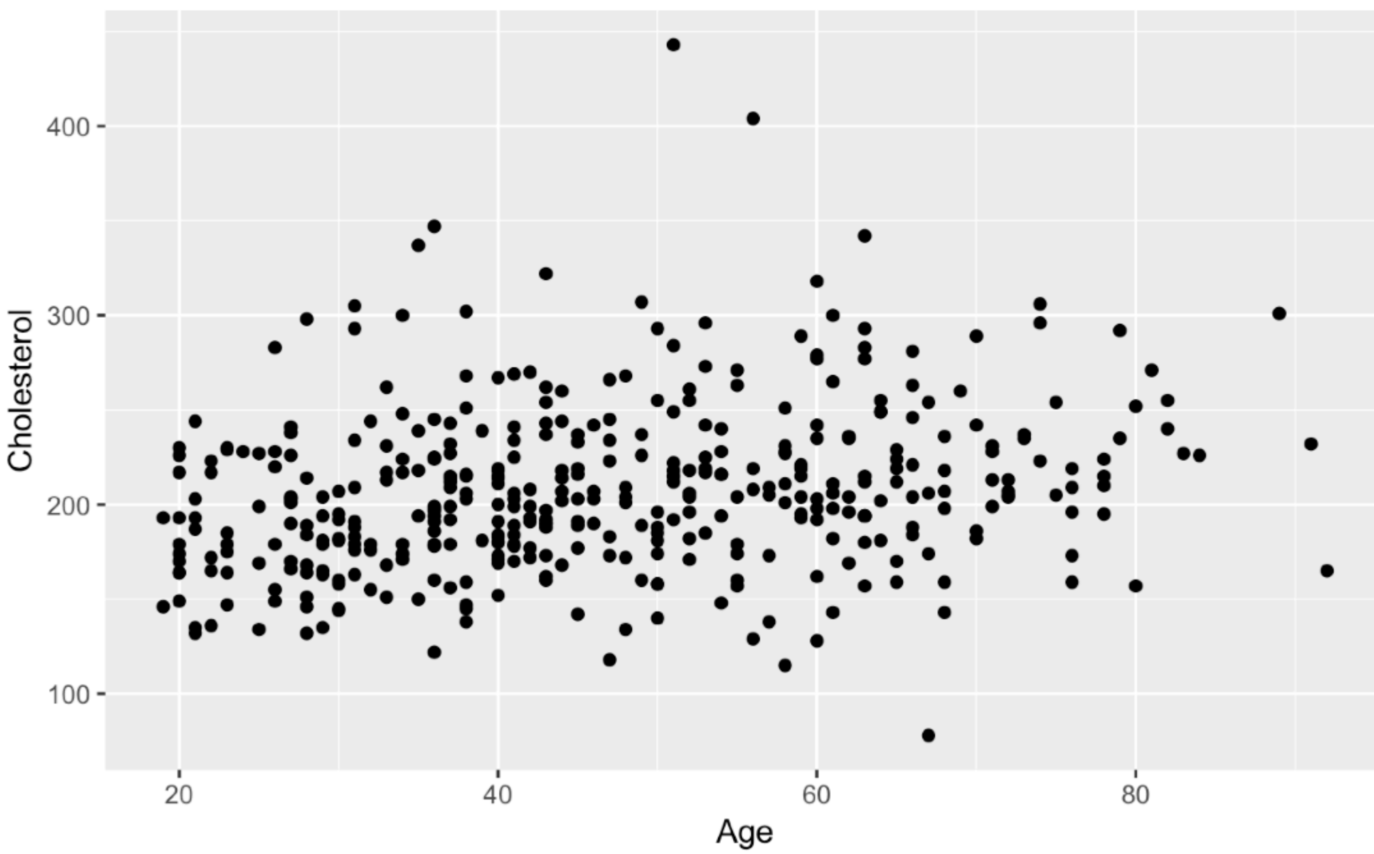
- Do not over-interpret scatter plots!
- Existence of relation between variables does not mean that there is a causal relationship between them!
- Correlation is NOT causality!!

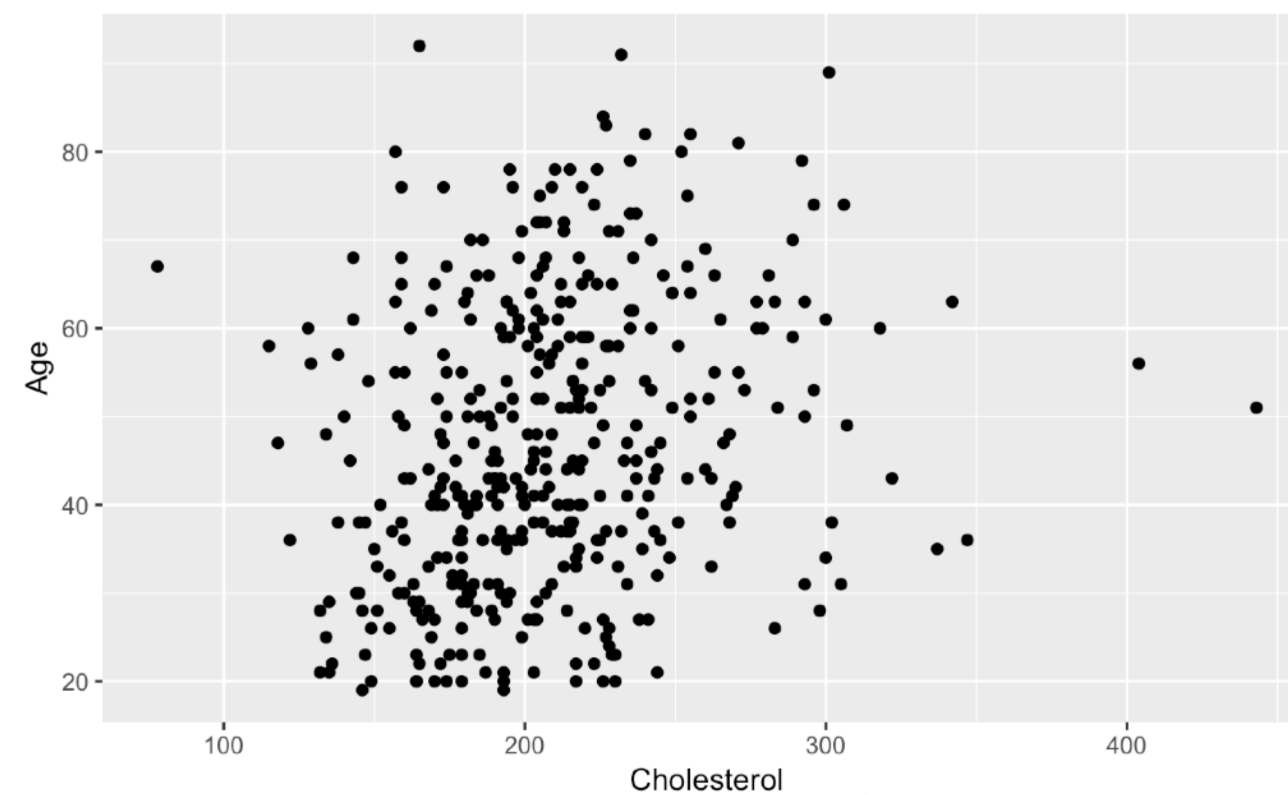


http://www.tylervigen.com/spurious-correlations

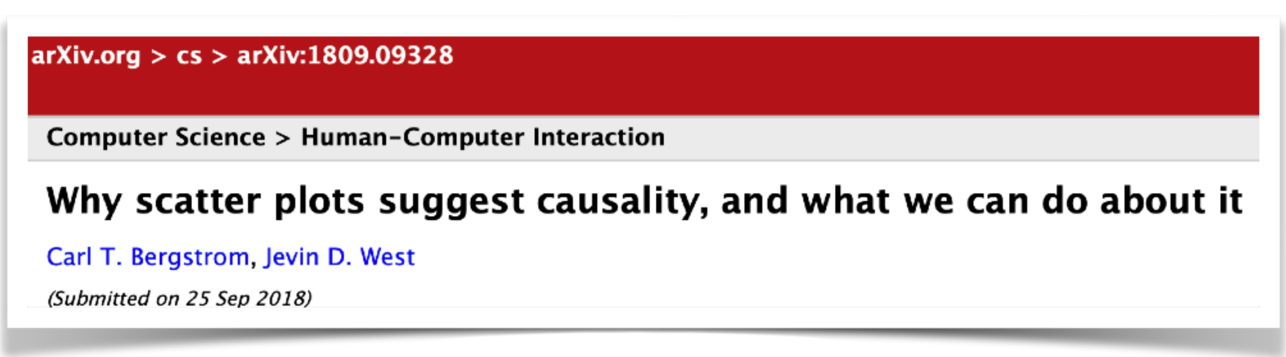
Numerical data

comparing variable with scatter plots





- In scatter plots, the x and y axis are exchangeable
- To avoid interpreting x→ y, **diamond plots**



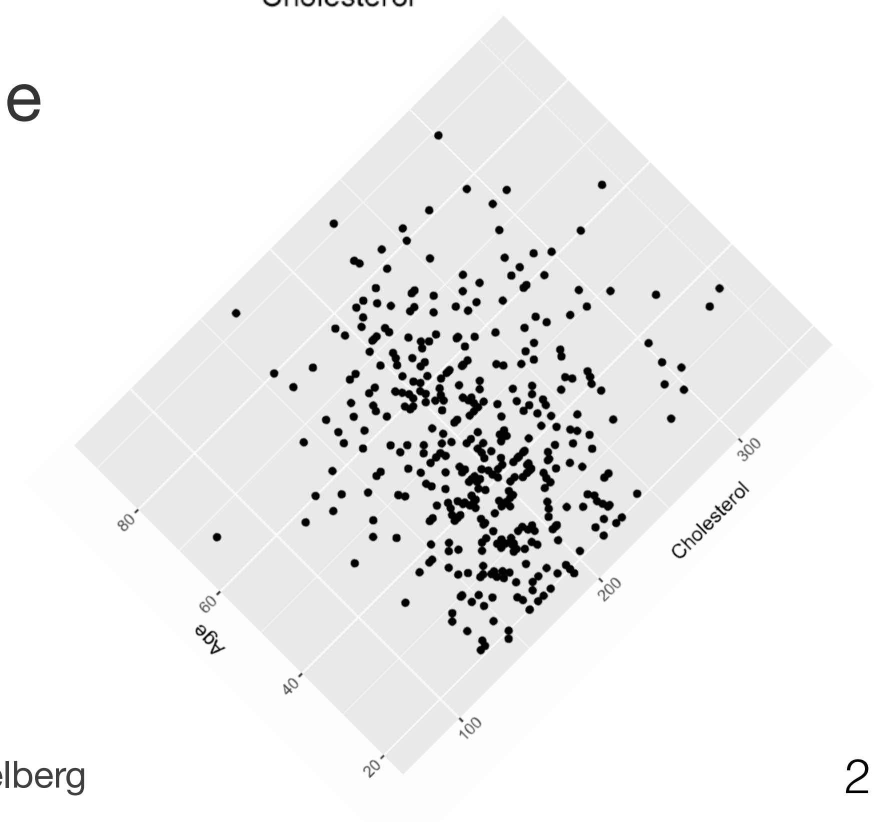
arXiv.org > cs > arXiv:1809.09328

Computer Science > Human-Computer Interaction

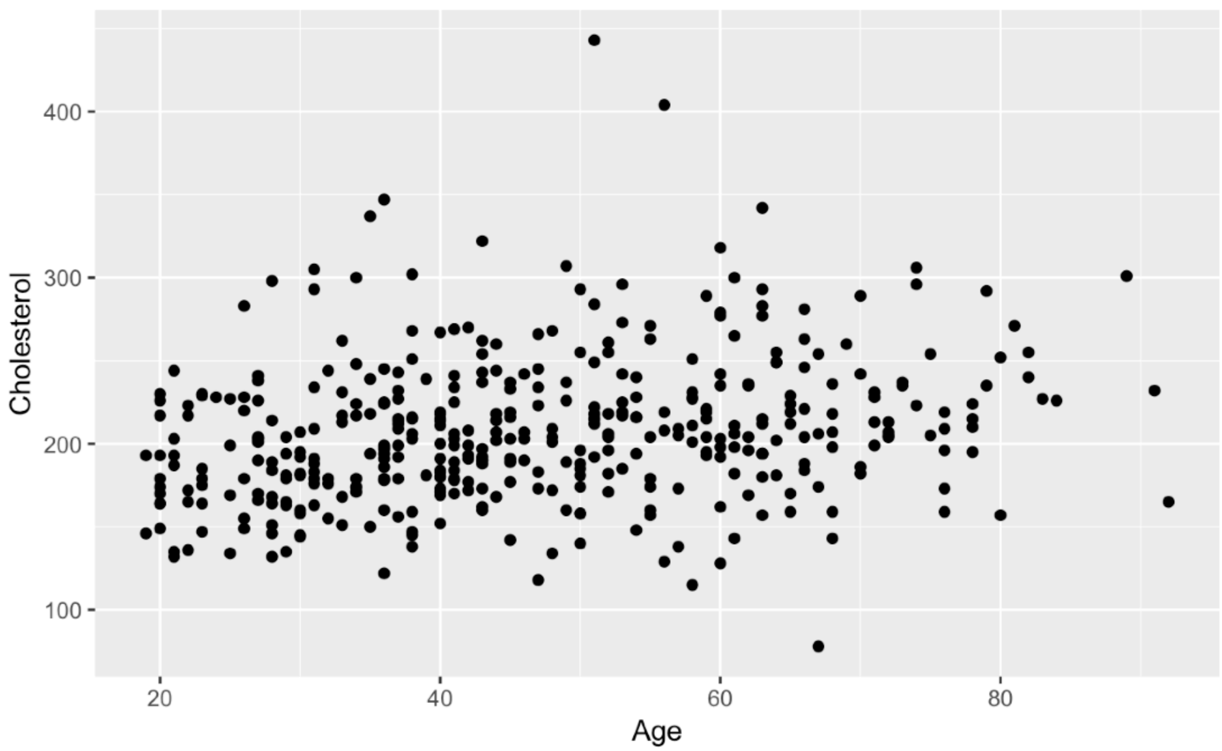
Why scatter plots suggest causality, and what we can do about it

Carl T. Bergstrom, Jevin D. West

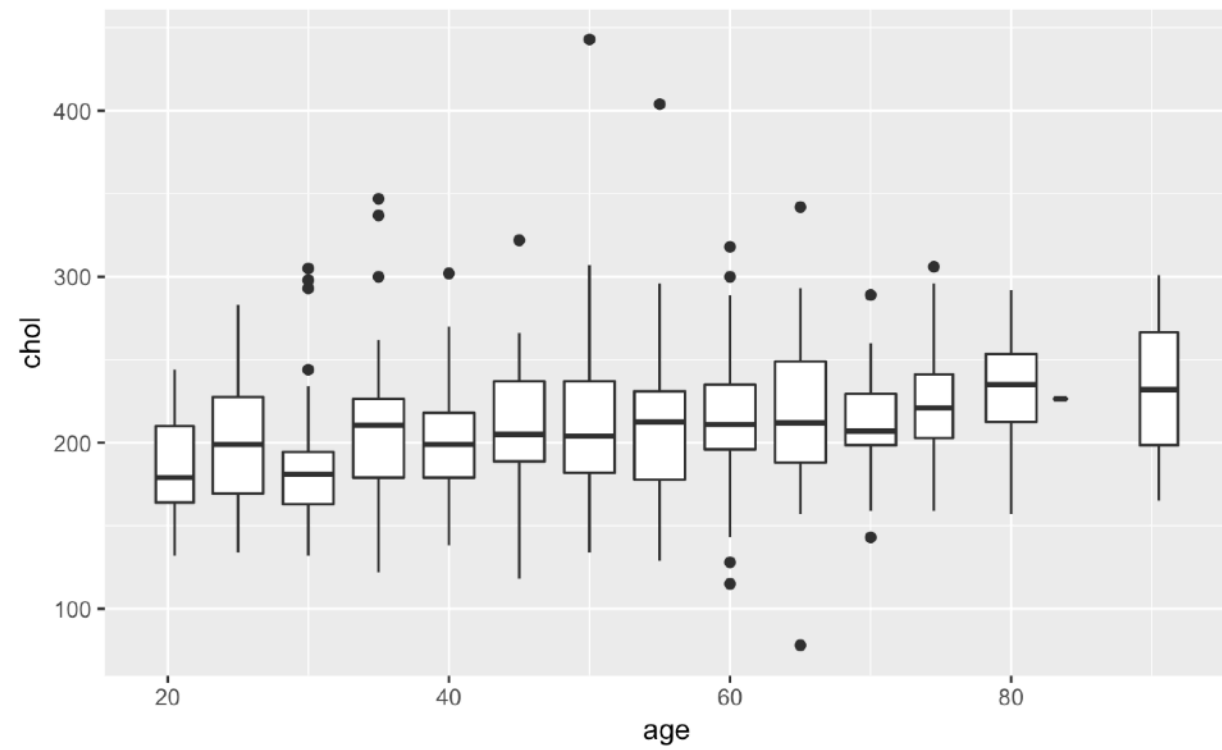
(Submitted on 25 Sep 2018)



Numerical data
comparing variable with scatter plots



Age = continuous variable

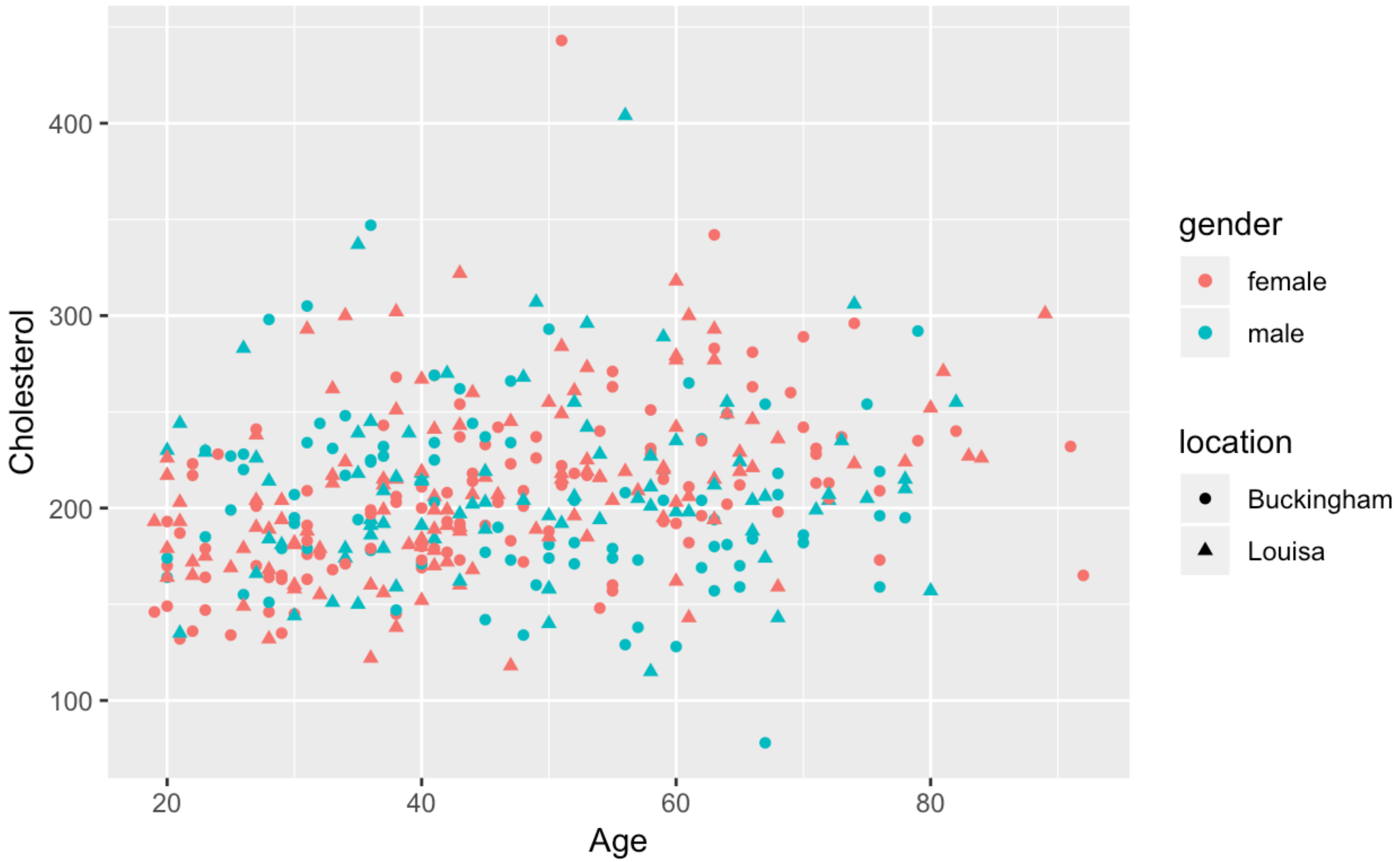


Age = ordinal variable (bins)

- A **continuous numerical** variable can always be transformed into an **ordinal categorical** variable through binning

Numerical data

comparing variable with scatter plots



- Additional categorical / numerical variables can be added using color, shape, size of dot ,…

Summary on visualization

Medizinische Fakultät Heidelberg

Single variable plot
plot counts

	type of plot
continuous variable	histogram
categorical variable	barplot

Two variable plot
plot relationship

	continuous variable	categorial data
continuous variable	scatter plot	boxplot
categorical variable		heatmap