Biological Data Analysis

Carl Herrmann
Health Data Science Unit
Medizinische Fakultät

bioinfo.ipmb.uni-heidelberg.de/crg/datascience3fs

Medizinische Fakultät Heidelberg

Data Analysis

Medizinische Fakultät Heidelberg

- New course in the MoBi Bachelor : **Data Analysis**

- Goals

  1. give students the basic **concepts** and **tools** for modern biological data analysis

  2. explain the underlying **mathematical/statistical concepts**

  3. provide practicals **“hands-on”** on the computer

Organisation : 2 phases

Medizinische Fakultät Heidelberg

- **MoBi 3. FS**
  - **Basics of data analysis**
    - 10 weeks
    - C. Herrmann + tutors
    - *Thursday 10-12* : theoretical concepts (BQ SR041)
    - *Friday 10-11.30 and 11.30-13* : pratical sessions (IPMB 5th floor)
  - **Basics of image analysis**
    - 4 weeks
    - K. Rohr + tutors
    - organisation will be explained by K. Rohr later
- **MoBi 4. FS**
  - **Data analysis projects**
    - groups of 4-5 students
    - 5 topics (biomedical imaging, cancer genomics, population genomics,…)
    - 4-5 subjects in each topic
  - **Goals**
    - set up a **project plan** (timeline, milestones, …)
    - use modern project tools (RStudio, markdown, git)
  - **Setup**
    - supervisors + master tutors
    - weekly meeting with tutors / monthly with supervisors

Steps of data analysis

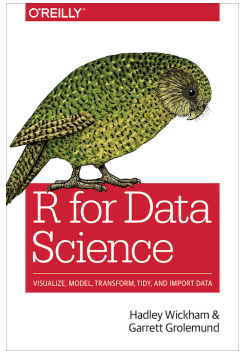
Medizinische Fakultät Heidelberg

- **data cleaning:**
  - bring data into right shape
  - label columns / rows
  - (transform data)
  - remove missing values / weird values
- **data exploration**
  - descriptive statistics
  - graphical representation
- **data reduction**
  - dimensionality reduction
  - find important features
- **data modelling**
  - find relationships between variables
  - build predictive models *("can I explain cholesterol level y age?")*

Data Analysis in R

Medizinische Fakultät Heidelberg

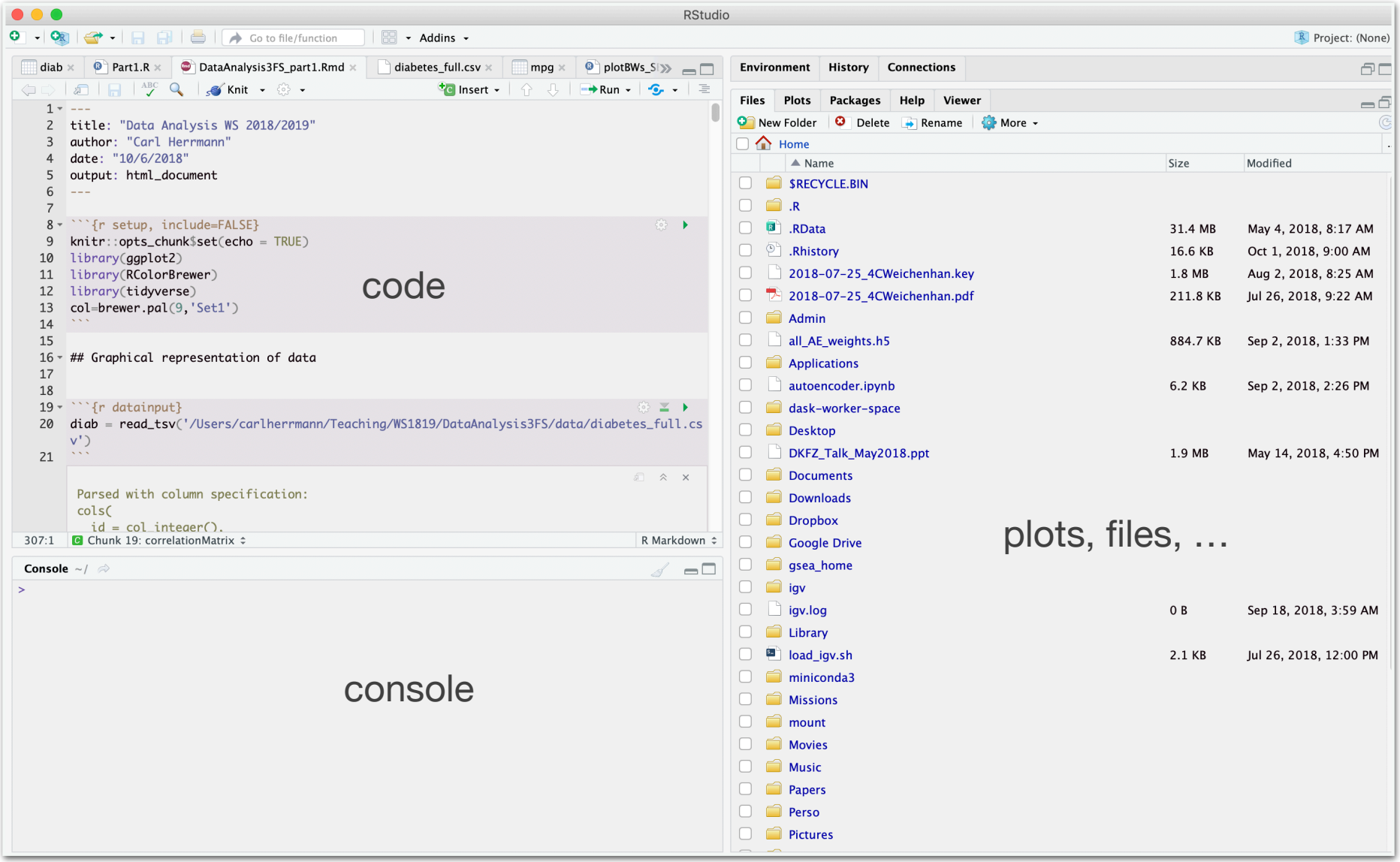
- R is a free software which is heavily used for data analysis in (biological) sciences
- How to learn / use R ?
  - user-friendly free programming interface RStudio
  - easy to install on any notebook
  - many online tutorials to learn R
    - https://www.datacamp.com/courses/free-introduction-to-r
  - online books: ***R for data science***
    http://r4ds.had.co.nz
  - ***Data Analysis for the Life Sciences***
    http://leanpub.com/dataanalysisforthelifesciences



- Hands-on courses will be based on R running under RStudio

RStudio

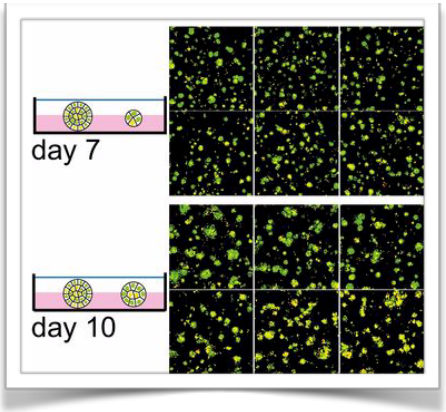
Medizinische Fakultät Heidelberg



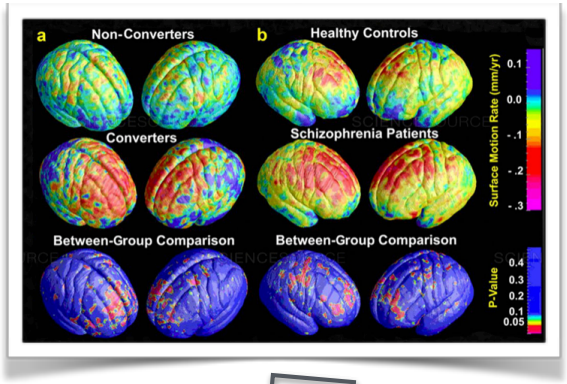
Biological data sources

Medizinische Fakultät Heidelberg

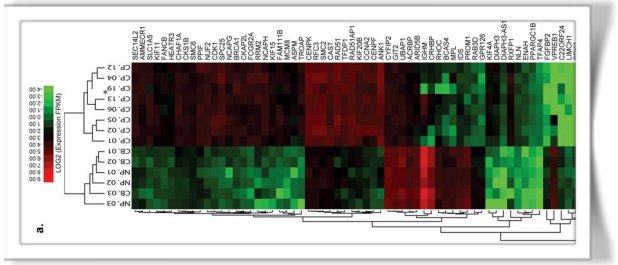
drug screens



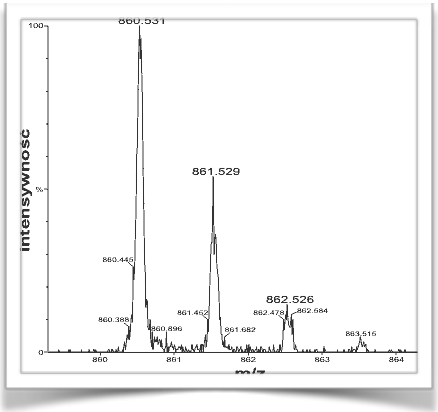
MRI imaging



Genomics



proteomics

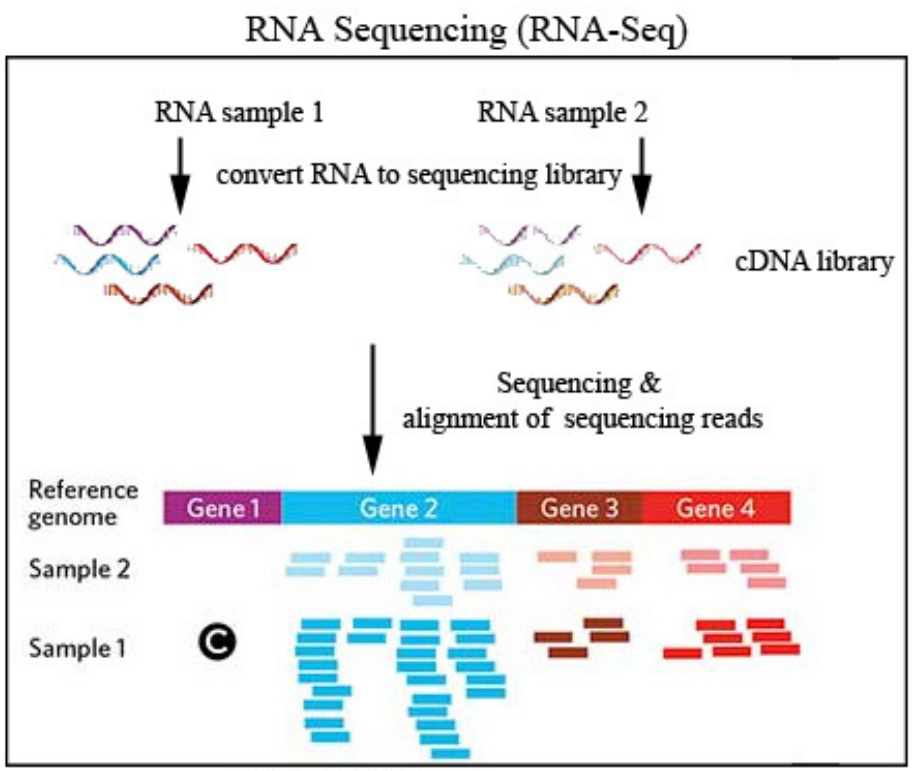


large (numerical) data matrices

Patient data

Example: gene expression

- mRNA molecules are retro-transcribed into cDNA
- cDNA libraries are sequenced into short fragments (~ 100 bp)
- Fragments are aligned to the genome
- Normalized number of fragments per gene is a measure of the gene expression
- examples of normalizations: reads per kilobase per million aligned reads (RPKM)



Example : gene expression

Medizinische Fakultät Heidelberg

Normalized gene expression values

patients

genes

genes	patient_1	patient_2	patient_3	patient_4	patient_5	patient_6
gene_1	4.14391122151671	3.09824631414336	0.613539225627942	4.01373315182408	0.383657955269923	3.48703239461242
gene_2	2.97523610717862	1.20443006241933	4.3447882784591	3.85312401281436	2.99916338009409	1.05441291342982
gene_3	-3.46753471324685	-2.00101678051065	2.58908155510034	3.31290558096149	-0.519502934359013	2.06695160803053
gene_4	-0.0542174928459582	0.886366385234324	1.3073011508442	3.59288980826203	-1.68381158067256	0.825554440360175
gene_5	2.61181313832951	3.4882302199402	4.32813766407381	-0.0505145145819221	1.98050782176216	0.244541604241015
gene_6	-1.45317487435791	4.88710271285716	2.42872910997122	2.21426414198806	0.718811329188614	-0.571744532706626
gene_7	2.77215110740787	3.54253589333533	2.38051482667397	5.01641643779523	0.509570207501988	-0.186408874152574
gene_8	2.2927688092584	0.212519696631346	-2.59010108611397	1.17484649801012	-1.15221575225	3.35570196842051
gene_9	3.48233929969678	4.06841521869038	2.1067009132259	0.246581008396974	2.49873236606263	5.56743249318474
gene_10	1.75899552881198	1.85662970682716	2.72316312314842	4.14423009423598	3.28213973395049	1.28099713785821
gene_11	3.74044882780187	2.31662619850529	3.95751969567219	2.24119784792571	1.92077174598244	2.74344347098465
gene_12	3.38088419504069	2.21860133478236	3.05142668563855	1.14119750261044	0.599592016307555	2.1602185454517
gene_13	0.440133982460589	4.10650451851398	3.33932041189726	4.56590629189043	-0.362420445668377	1.68151413513521
gene_14	4.20025356837806	3.91651458579653	2.78369552661006	0.100636722195449	4.51116601360399	-0.624700579766187
gene_15	-1.16027743407016	1.1845835671856	4.62261382573676	0.99468091768471	2.40311034649016	2.49773441195641
gene_16	1.83405661461064	1.23815661160894	1.65011336969598	-1.80471720459685	3.522635153674	0.751495646081527
gene_17	1.62930097507458	4.41001494421025	1.13494604699906	3.73334451573132	1.62900737559927	1.36701944902375
gene_18	1.62359008551869	3.25682343996691	0.37451839172457	3.78150039472285	3.85129869584565	2.57824911102574
gene_19	3.68845851101034	5.11443872961371	0.46300831913987	0.553266571878013	2.06750452854509	0.792616197217081
gene_20	1.02923146568405	2.97238032056755	0.771149484664369	3.36761511762975	1.82242868241571	-0.444194242723608

Example : gene expression

Medizinische Fakultät Heidelberg

Gene expression values

genes

patients

genes	gene_1	gene_2	gene_3	gene_4
patient_1	4.14391122151671	2.97523610717862	-3.46753471324685	-0.0542174928459582
patient_2	3.09824631414336	1.20443006241933	-2.00101678051065	0.886366385234324
patient_3	0.613539225627942	4.3447882784591	2.58908155510034	1.3073011508442
patient_4	4.01373315182408	3.85312401281436	3.31290558096149	3.59288980826203
patient_5	0.383657955269923	2.99916338009409	-0.519502934359013	-1.68381158067256
patient_6	3.48703239461242	1.05441291342982	2.06695160803053	0.825554440360175
patient_7	0.515634555377573	0.96825404034163	-4.24026922022054	-2.15459287415868
patient_8	0.833008182831021	1.66882594396195	3.54523865807237	3.91542374367522
patient_9	3.27748316735109	3.44774731855558	3.87916459892164	2.37301081167968
patient_10	2.4007858903261	-1.25740300340426	2.39681703044223	2.47361107715352
patient_11	2.52460225221238	0.705443998616781	0.305480184099154	4.1439933583416
patient_12	1.14585639287521	0.876609151235227	1.06556812351244	0.857585380566236
patient_13	0.912240962281692	1.43642915421264	2.21361502482881	0.191184996769756

Tables are usually oriented such that # rows > # columns

Example : clinical data

Medizinische Fakultät Heidelberg

Cohort of diabetes patients with multiple clinical parameters

id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1000	203	82	56	3.60	4.31	Buckingham	46	female	62	121	medium	118	59	NA	NA	29	38	720
1001	165	97	24	6.90	4.44	Buckingham	29	female	64	218	large	112	68	NA	NA	46	48	360
1002	228	92	37	6.20	4.64	Buckingham	58	female	61	256	large	190	92	185	92	49	57	180
1003	78	93	12	6.50	4.63	Buckingham	67	male	67	119	large	110	50	NA	NA	33	38	480
1005	249	90	28	8.90	7.72	Buckingham	64	male	68	183	medium	138	80	NA	NA	44	41	300
1008	248	94	69	3.60	4.81	Buckingham	34	male	71	190	large	132	86	NA	NA	36	42	195
1011	195	92	41	4.80	4.84	Buckingham	30	male	69	191	medium	161	112	161	112	46	49	720
1015	227	75	44	5.20	3.94	Buckingham	37	male	59	170	medium	NA	NA	NA	NA	34	39	1020
1016	177	87	49	3.60	4.84	Buckingham	45	male	69	166	large	160	80	128	86	34	40	300
1022	263	89	40	6.60	5.78	Buckingham	55	female	63	202	small	108	72	NA	NA	45	50	240
1024	242	82	54	4.50	4.77	Louisa	60	female	65	156	medium	130	90	130	90	39	45	300
1029	215	128	34	6.30	4.97	Louisa	38	female	58	195	medium	102	68	NA	NA	42	50	90
1030	238	75	36	6.60	4.47	Louisa	27	female	60	170	medium	130	80	NA	NA	35	41	720
1031	183	79	46	4.00	4.59	Louisa	40	female	59	165	medium	NA	NA	NA	NA	37	43	60
1035	191	76	30	6.40	4.67	Louisa	36	male	69	183	medium	100	66	NA	NA	36	40	225
1036	213	83	47	4.50	3.41	Louisa	33	female	65	157	medium	130	90	120	96	37	41	240
1037	255	78	38	6.70	4.33	Louisa	50	female	65	183	medium	130	100	NA	NA	37	43	180

Different data types

Medizinische Fakultät Heidelberg

Variable	Explanation	Unit	Type
chol	total cholesterol		numerical
stab.glu	Stabilized Glucose		numerical
hdl	High Density Lipoprotein		numerical
ratio	Cholesterol/HDL Ratio		numerical
glyhb	Glycosolated Hemoglobin		numerical
location			**categorical**
age			numerical
gender			**categorical**
height		inches	numerical
weight		pounds	numerical
frame			**categorical**
bp.1s	systolic blood pressure		numerical
bp.1d	diastolic blood pressure		numerical
bp.2s	systolic blood pressure		numerical
bp.2d	diastolic blood pressure		numerical
waist		inches	numerical
hip		inches	numerical
time.ppn	Time since last meal	minutes	numerical

Different data types

Medizinische Fakultät Heidelberg

data

numerical

discrete
- integer numbers
- ex: counts

continuous
- float numbers

interval
- can be measured
- intervals are identical
- no absolute zero!
- example: temperature

ratio
- can be measured
- absolute zero
- example: speed / mass

categorical

nominal
- categories
- cannot be ranked
- example: names / IDS

ordinal
- categories
- have a natural order
- ex: german grades

Different data types

Medizinische Fakultät Heidelberg

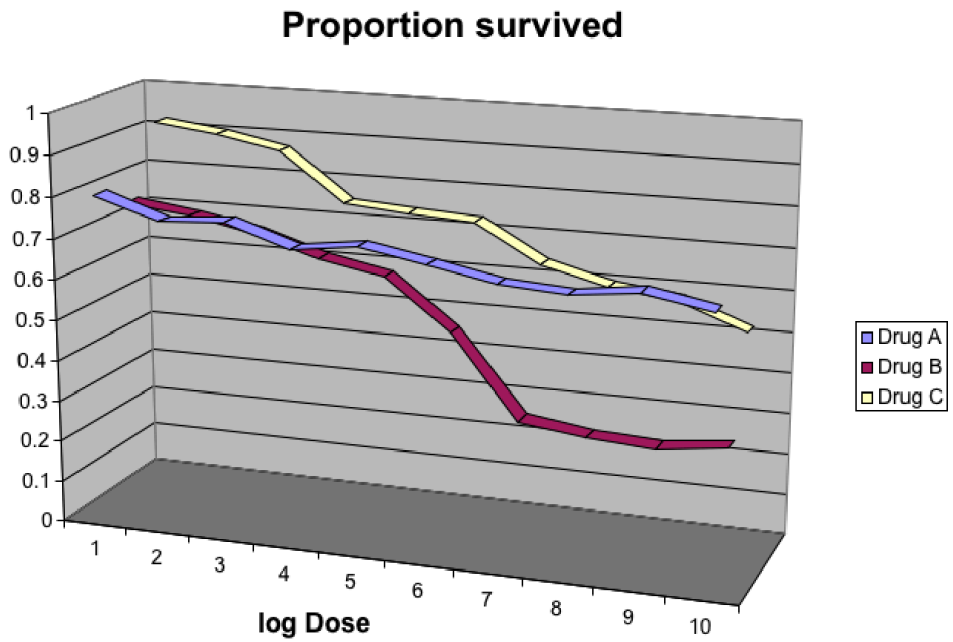
Variable	Explanation	Unit	Type	Subtype
chol	total cholesterol		numerical	ratio
stab.glu	Stabilized Glucose		numerical	ratio
hdl	High Density Lipoprotein		numerical	ratio
ratio	Cholesterol/HDL Ratio		numerical	ratio
glyhb	Glycosolated Hemoglobin		numerical	ratio
location			**categorical**	**nominal**
age			numerical	ratio
gender			**categorical**	**nominal**
height		inches	numerical	ratio
weight		pounds	numerical	ratio
frame			**categorical**	**ordinal (small/medium/large)**
bp.1s	systolic blood pressure		numerical	ratio
bp.1d	diastolic blood pressure		numerical	ratio
bp.2s	systolic blood pressure		numerical	ratio
bp.2d	diastolic blood pressure		numerical	ratio
waist		inches	numerical	ratio
hip		inches	numerical	ratio
time.ppn	Time since last meal	minutes	numerical	ratio

Medizinische Fakultät Heidelberg

1. Visualizing data with plots

Graphical representation

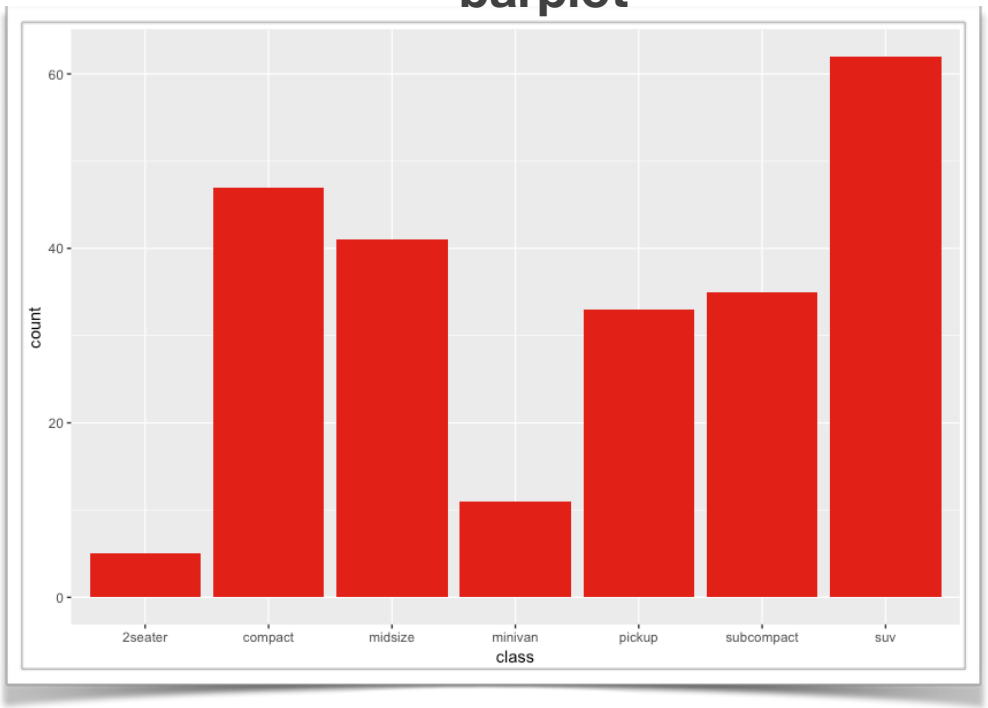
- Appropriate graphical representation depends on the type of data
  - categorical
  - counts
  - continuous data
- Aim of good data graphics: **display data accurately and clearly** (Karl Broman http://kbroman.org)
- **Bad practice:**
  - as little information as possible
  - make things obscure through inappropriate graphics
  - pseudo 3D
  - poor scales



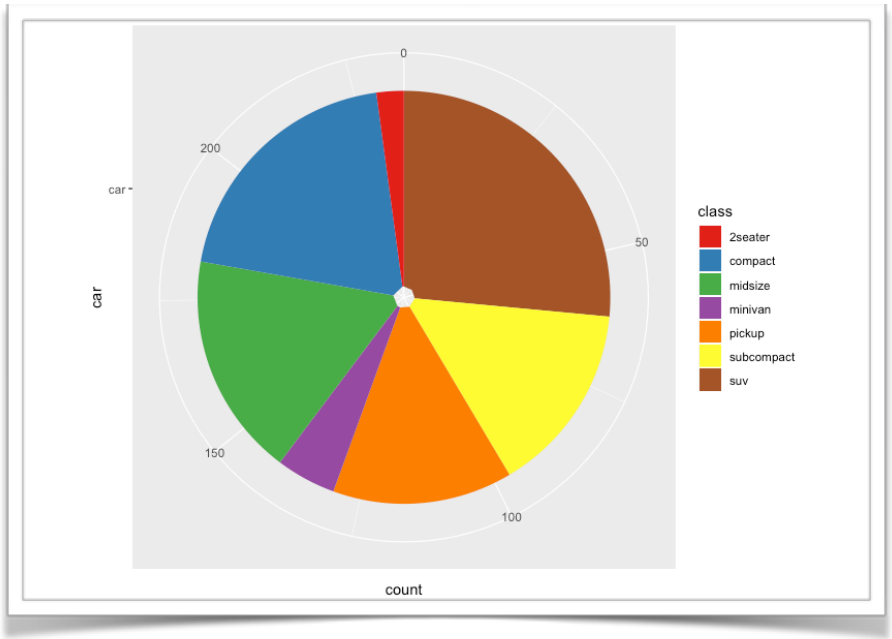
Categorical Data
Barplots

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

barplot

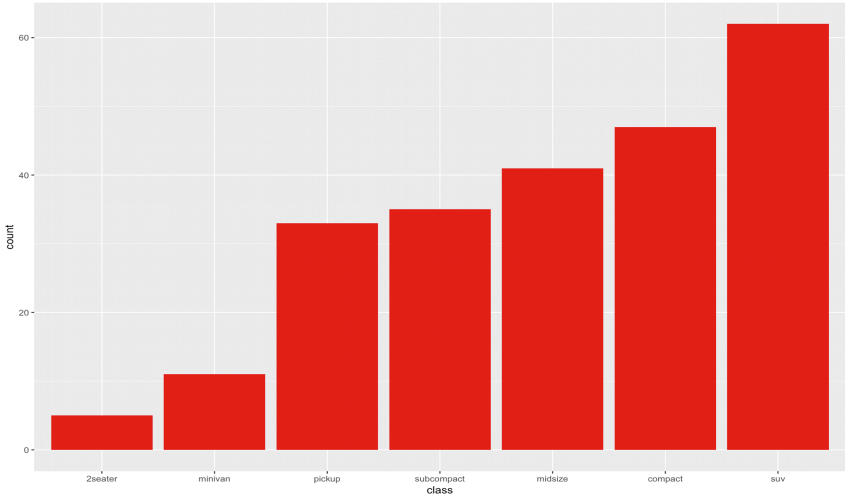


piechart

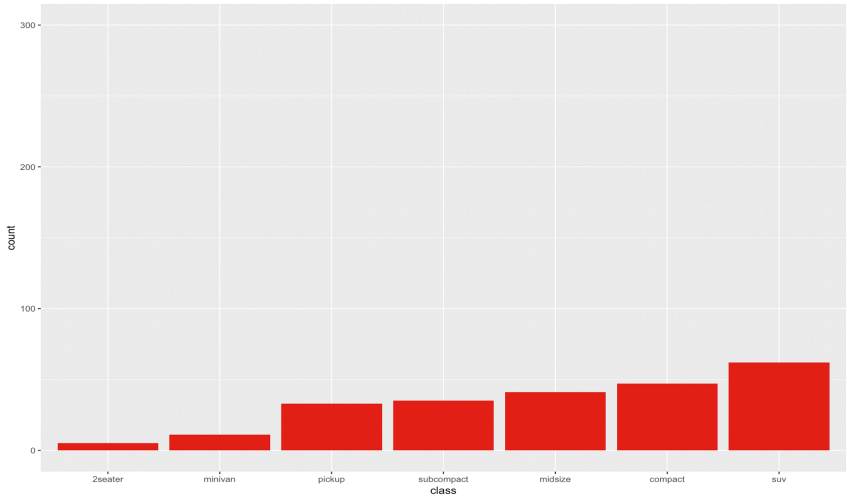


Avoid piechart : areas are more difficult to judge than length!

Categorical Data
Barplots



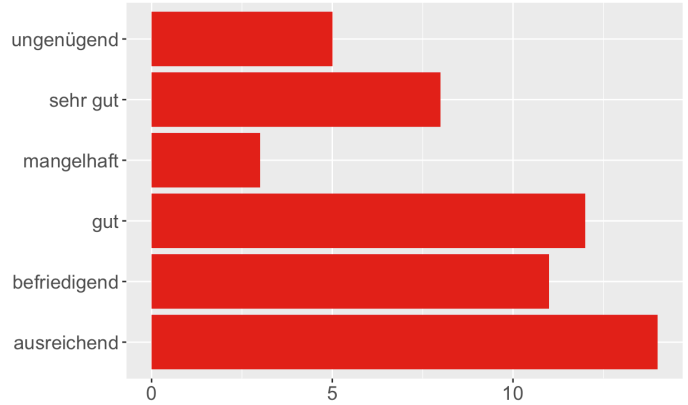
- Consider ordering the data by counts (no natural order of categories for nominal data)



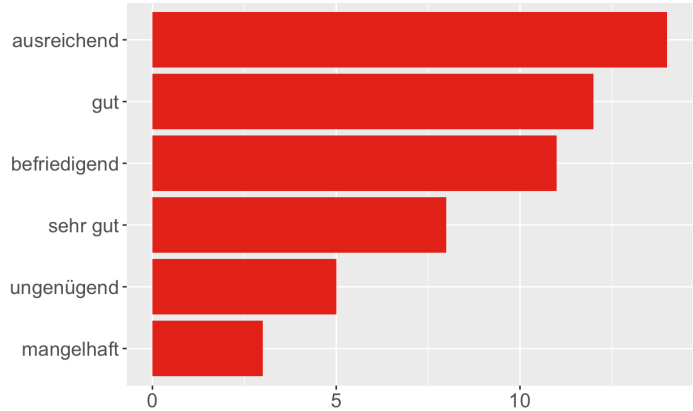
- Beware of selecting the proper scales for plotting!

Categorical Data

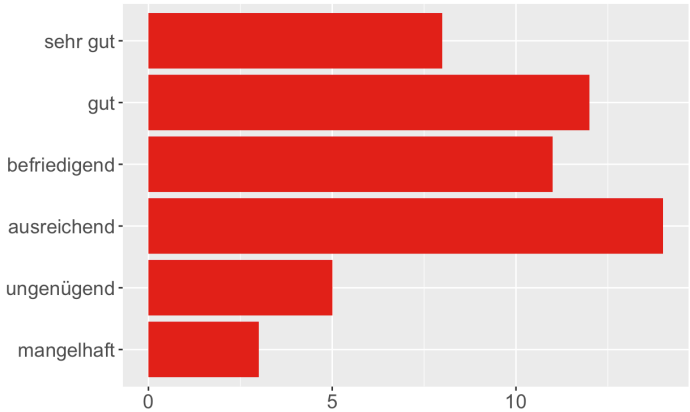
Barplots



- Random order



- Order by increasing / decreasing counts



- Natural order of ordinal data

Mode

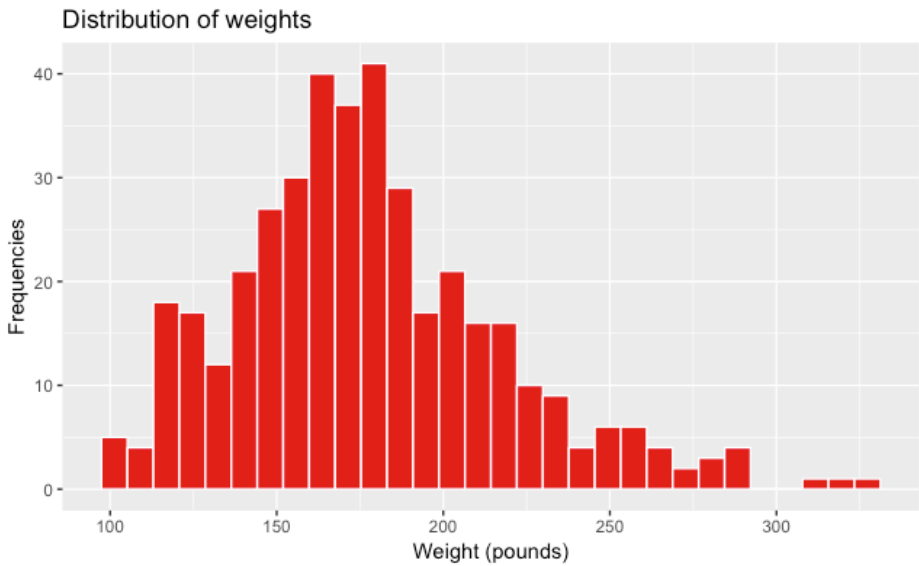
Numerical variables

Medizinische Fakultät Heidelberg

- Numerical data are instances of underlying **random variable**
- Random variables *X* have
  - **Density** distributions *p(X)*
  - **Expectation values** *E(X)*
  - **Variances** *Var(X)*

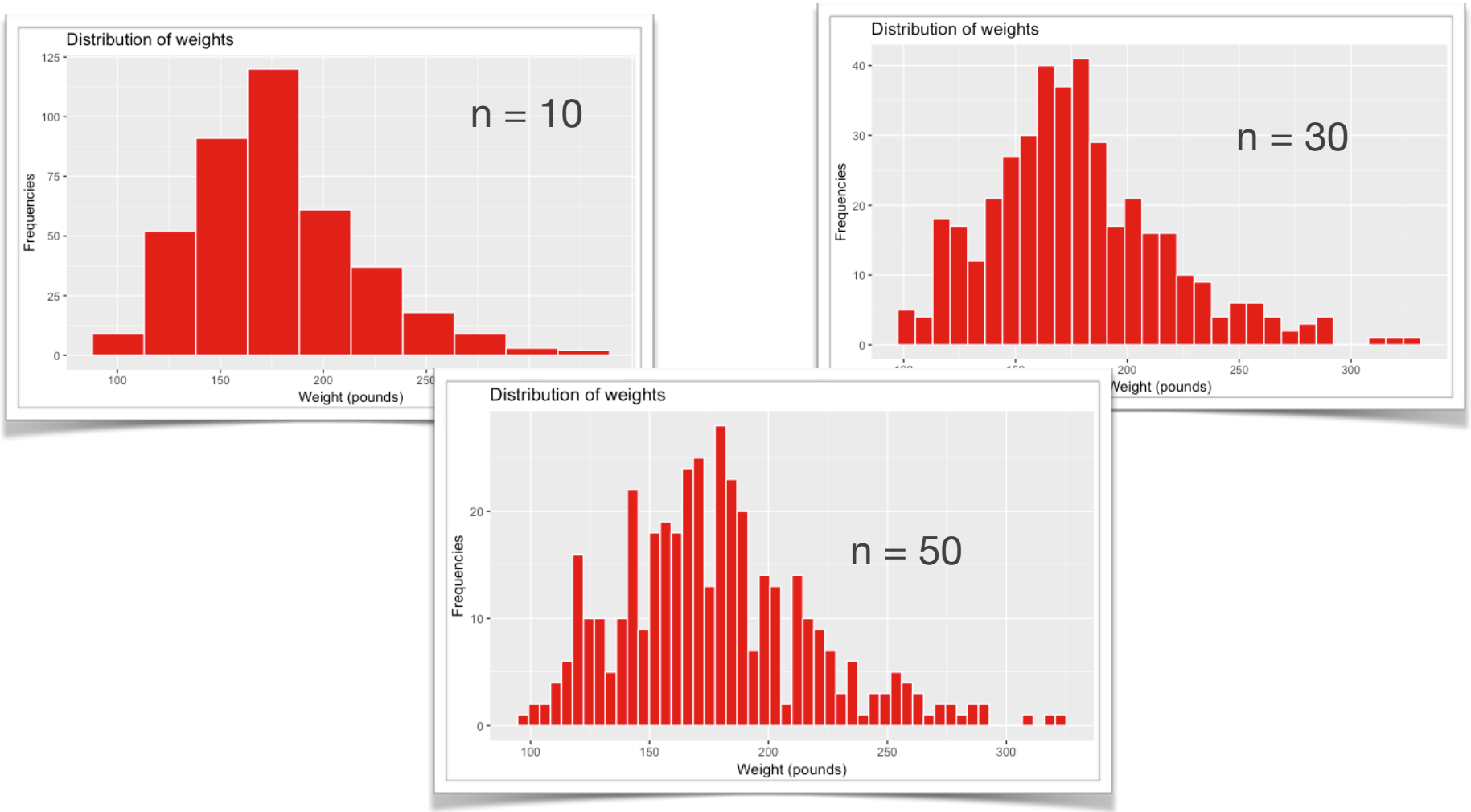
Numerical data
Histograms

- 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**



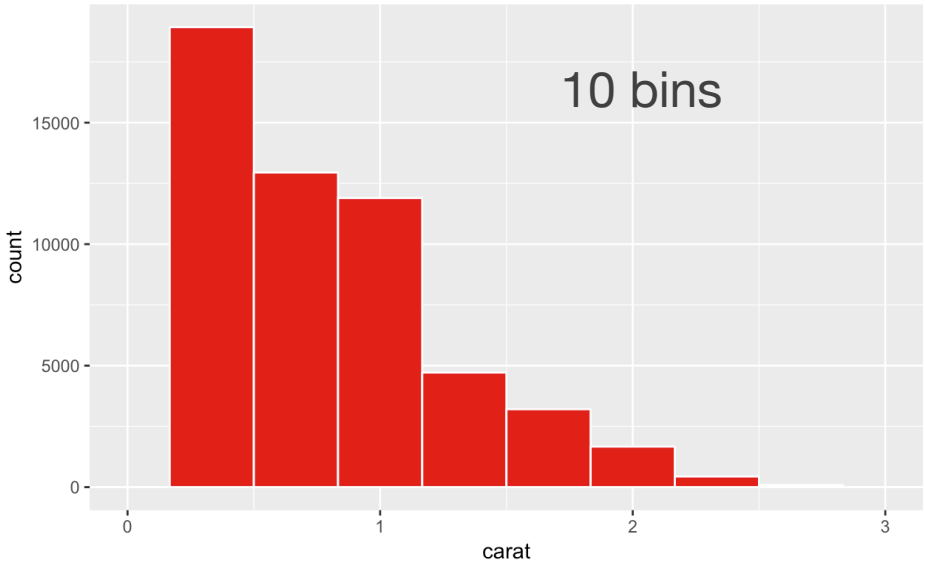
Numerical data
Histograms

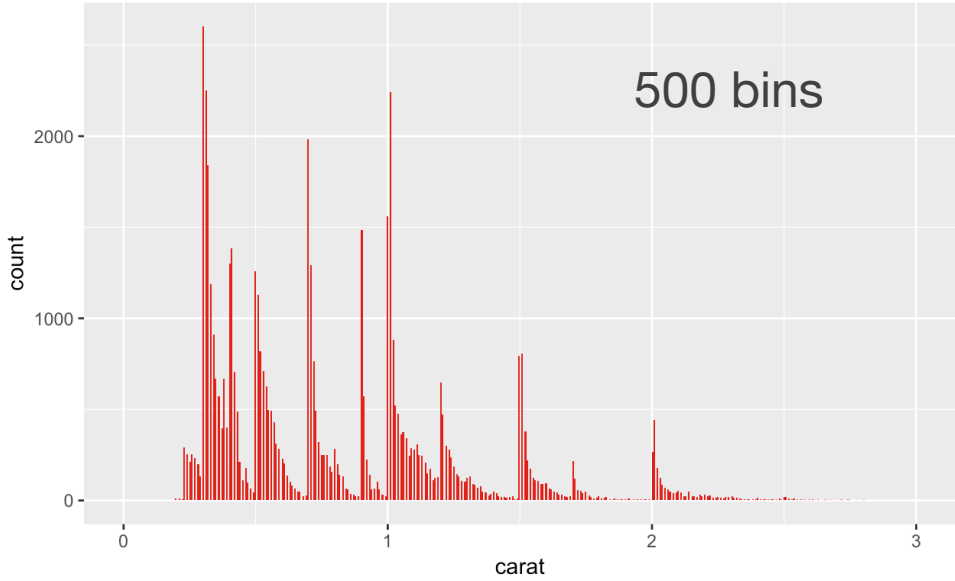
- Right choice of interval depends on the data type
- **Number of bins has a strong impact on appearance of plot!**



Numerical data
Histograms

Distribution of carat values for diamonds

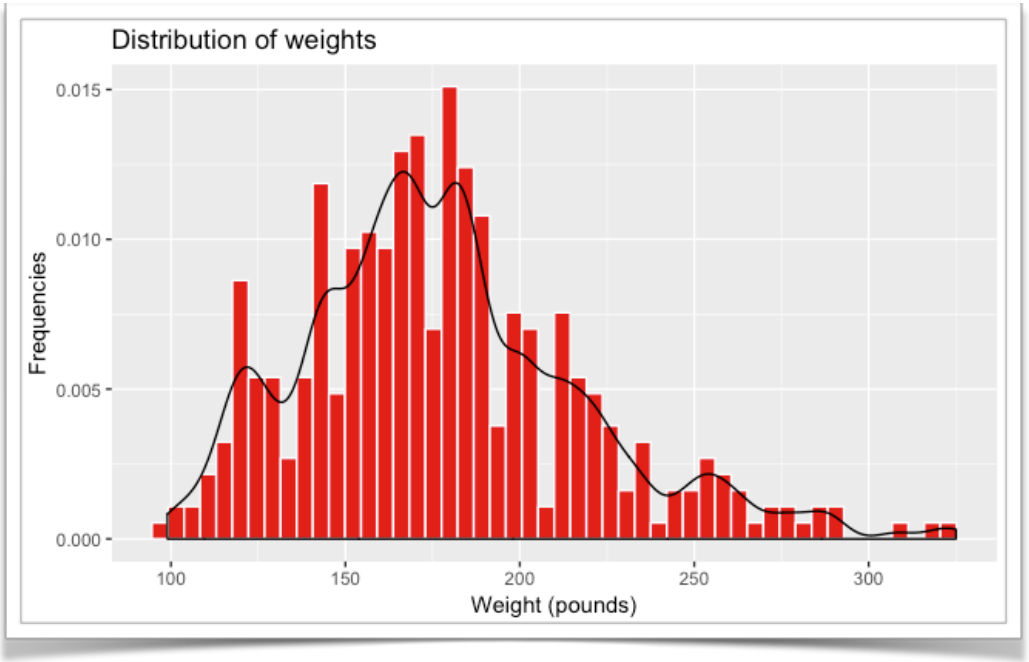




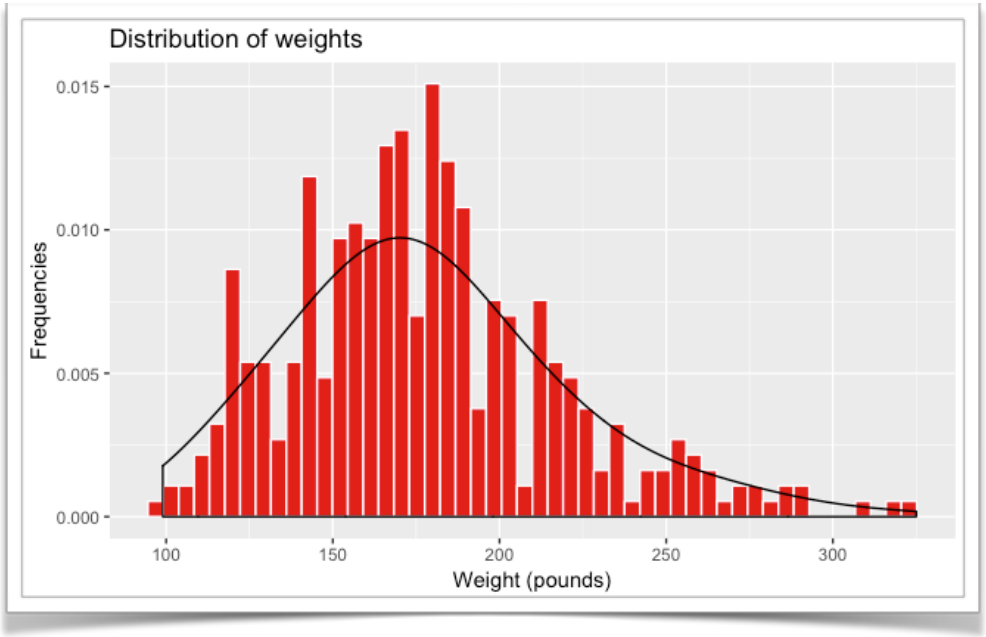
- pattern becomes visible at high resolution
- peaks around integer values (why?)
- tail on the right of integer values (why?)

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)



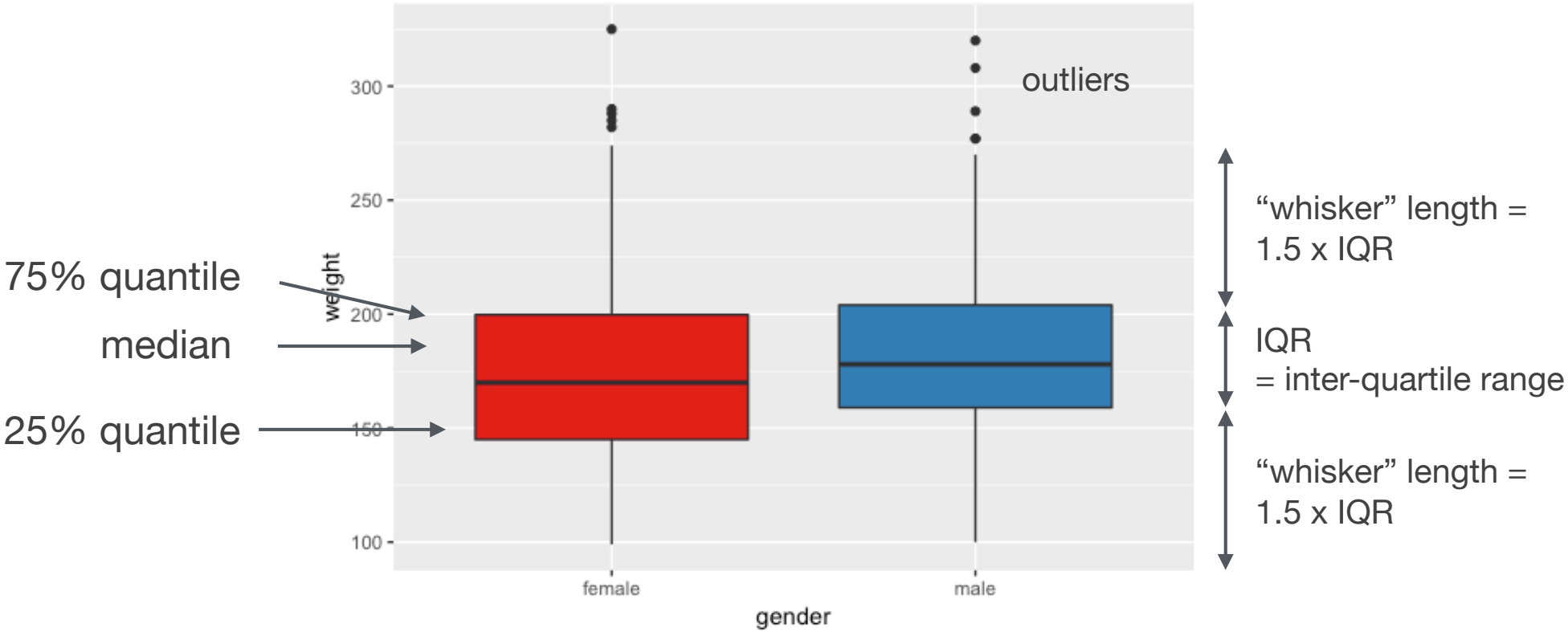
bandwith = 5



bandwith = 20

Numerical values
boxplots

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

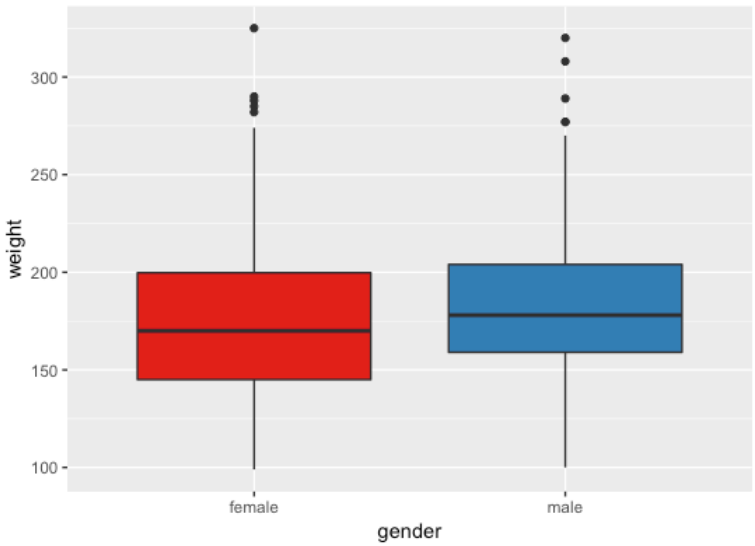


Whisker does not go not beyond maximun or minimum value !
(Hence both whisker can have different length < 1.5 x IQR)

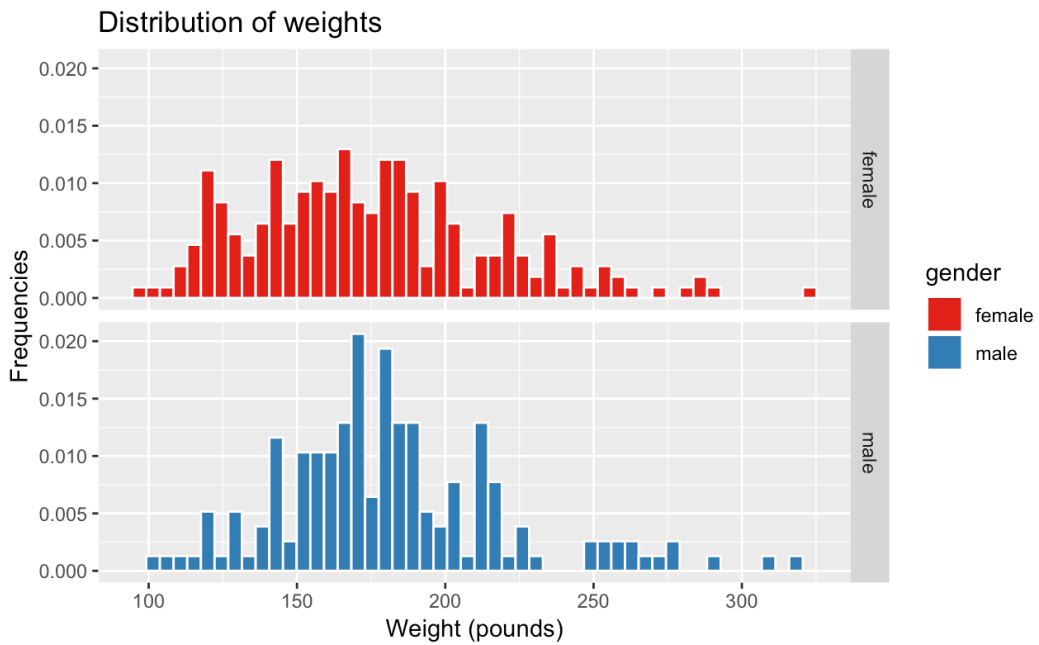
Numerical values
boxplots

Medizinische Fakultät Heidelberg

Boxplot

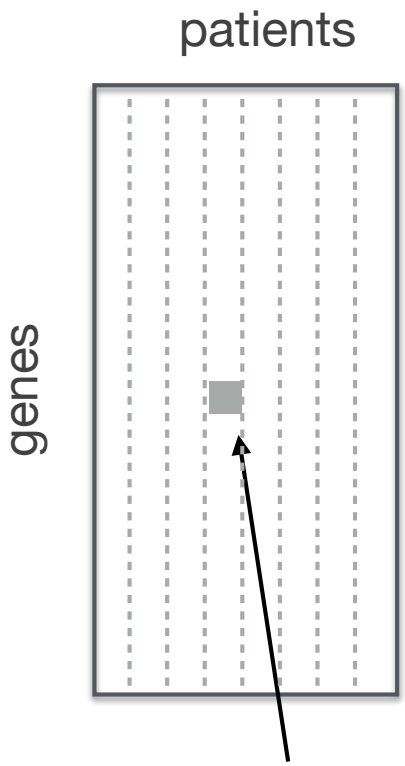


Histogram

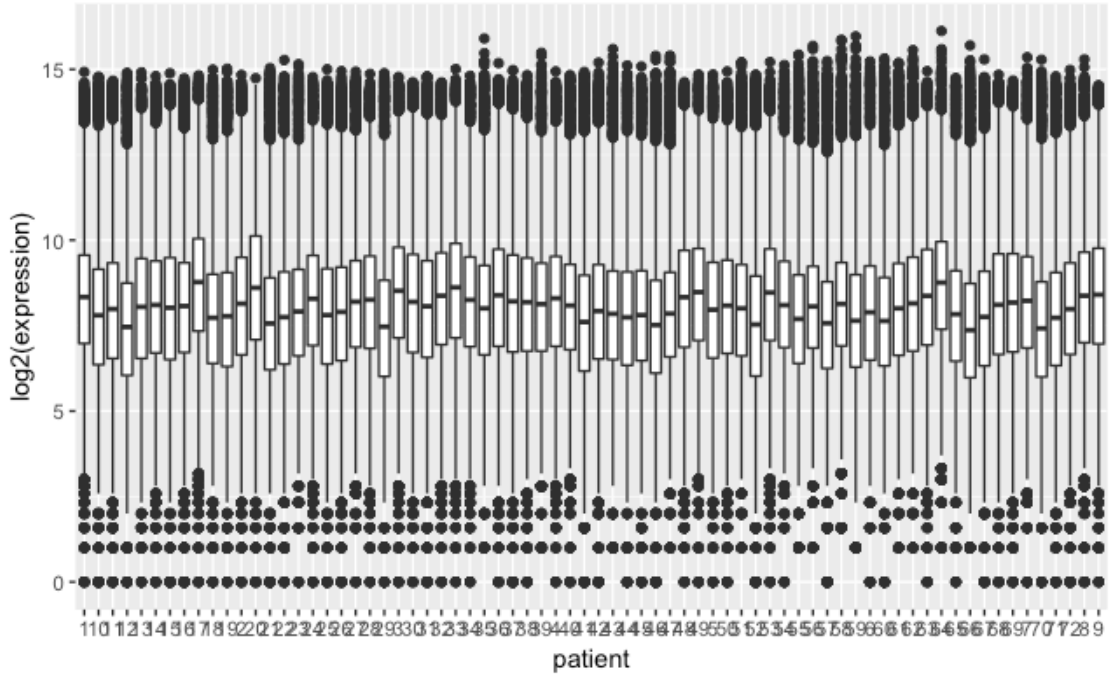


Boxplots summarize the properties of the distribution
Usefull to compare many distributions side-by-side

Numerical values
boxplots



Expression of gene *i* in patient *j*



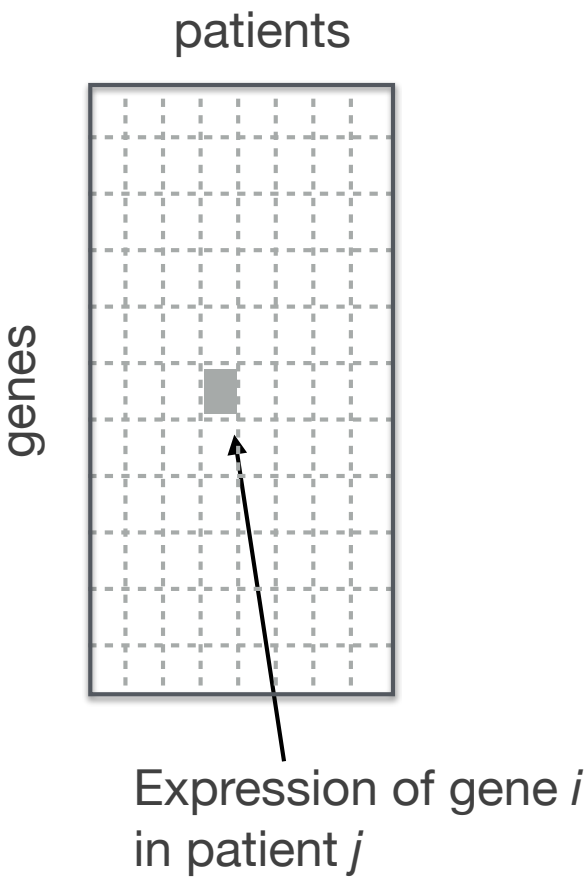
Question: do some patients have a different median gene expression?

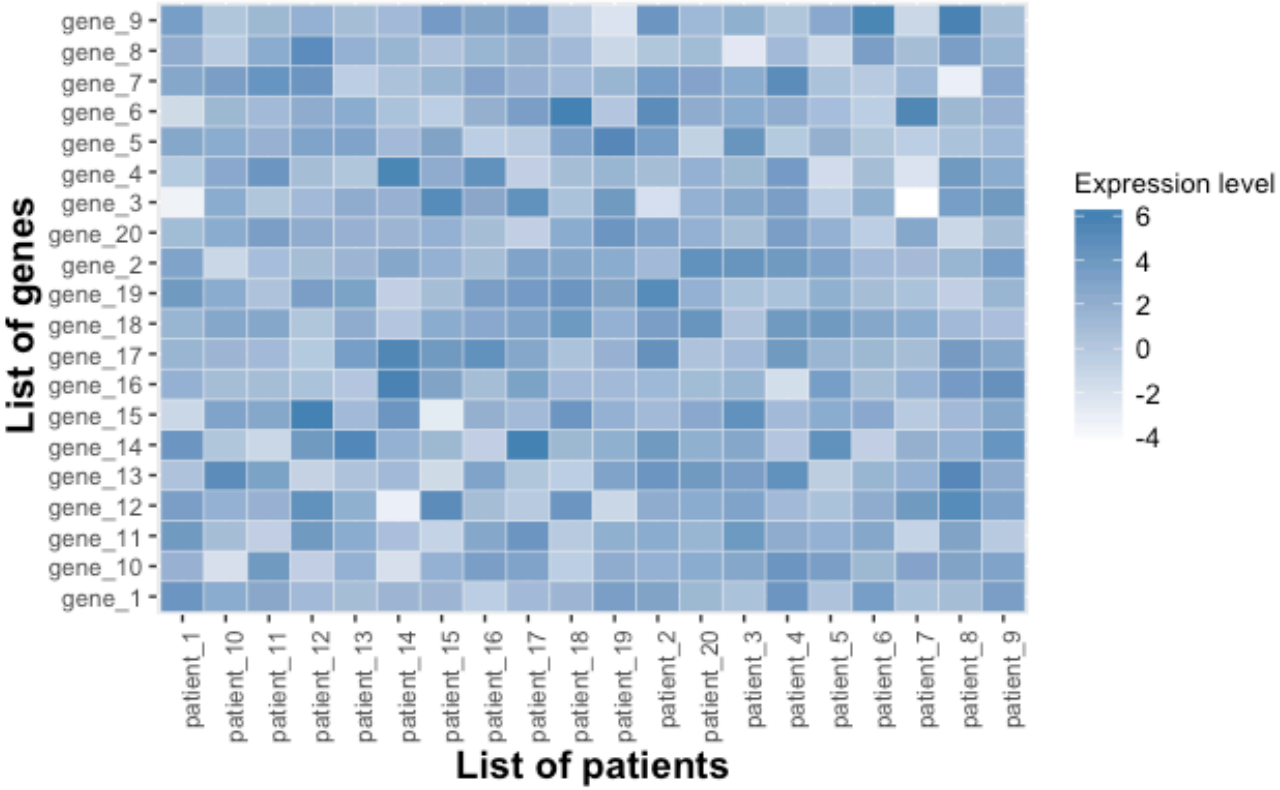
→ *values for individual genes are lost in this type of plot!*

Numerical Data :
heatmaps

Medizinische Fakultät Heidelberg

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

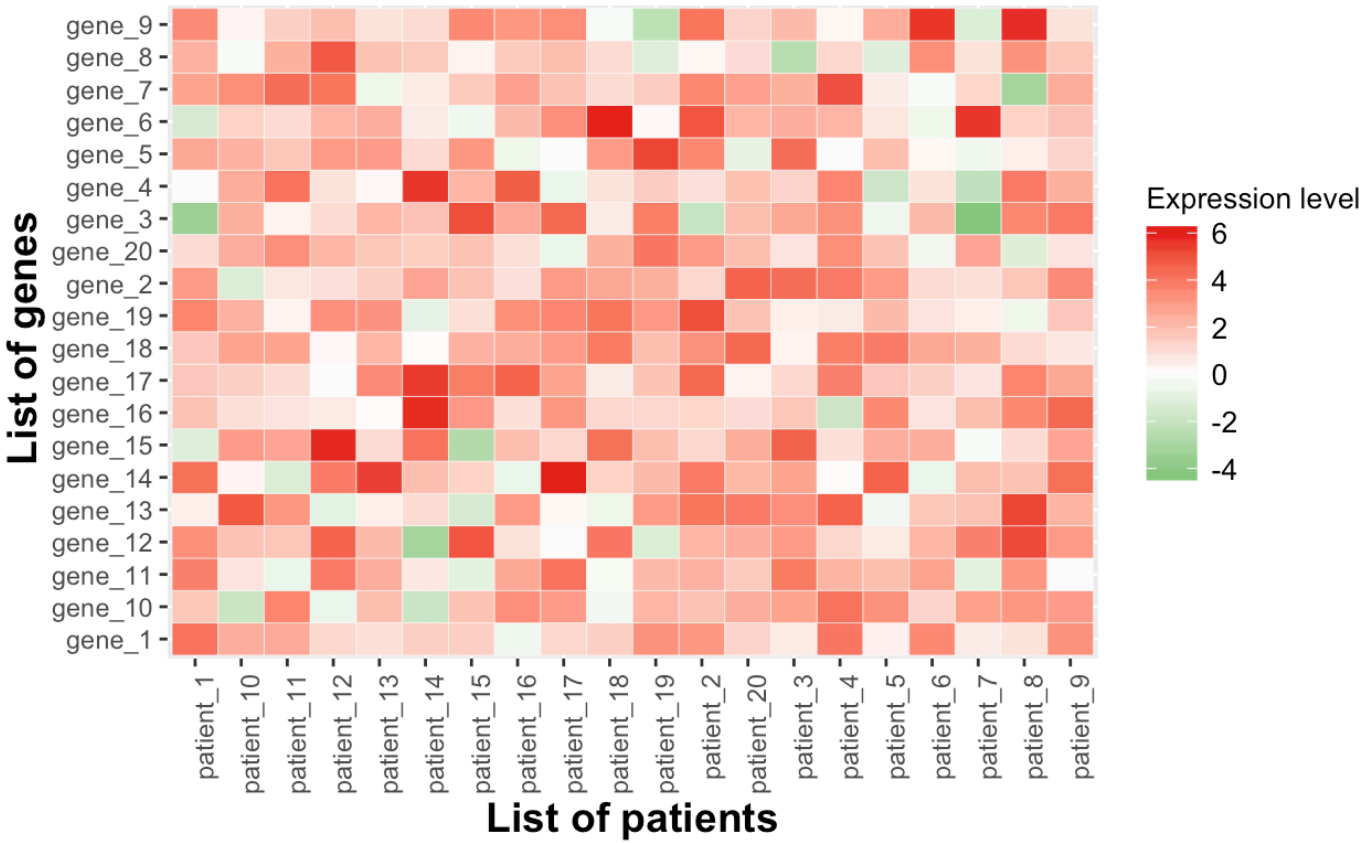




Question: do some genes in some patients have a different gene expression?

Numerical Data :
heatmaps

use symmetrical color scales for symmetrical ranges



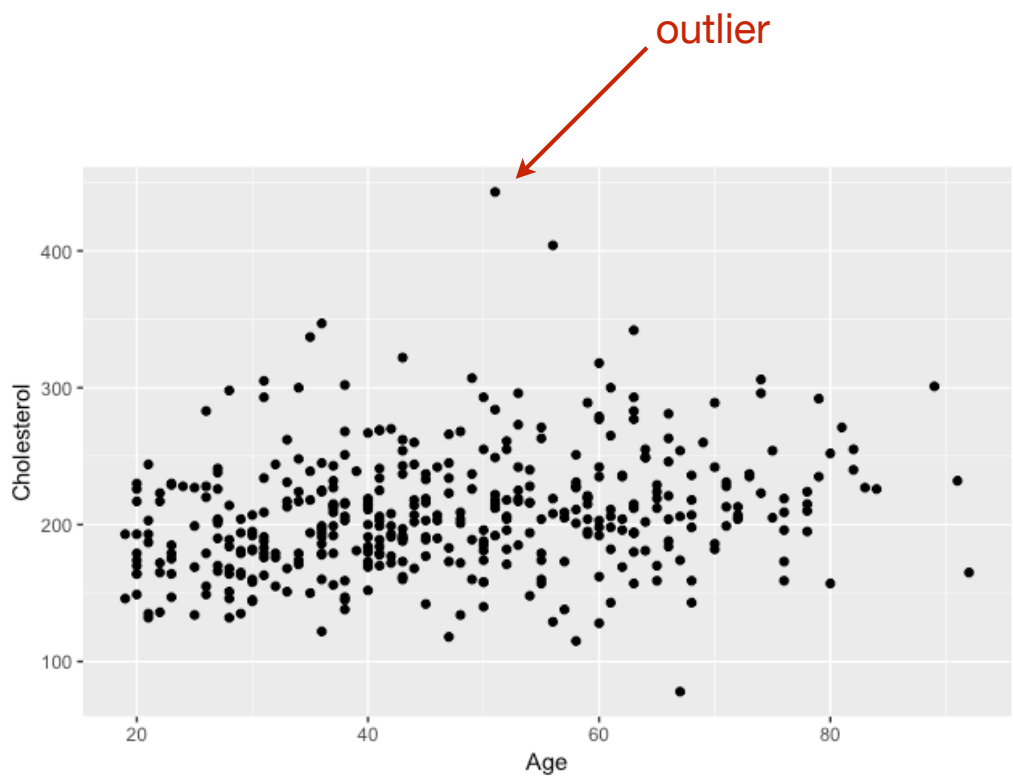
Numerical data

comparing variable with scatter plots

any relation between age and cholesterol?

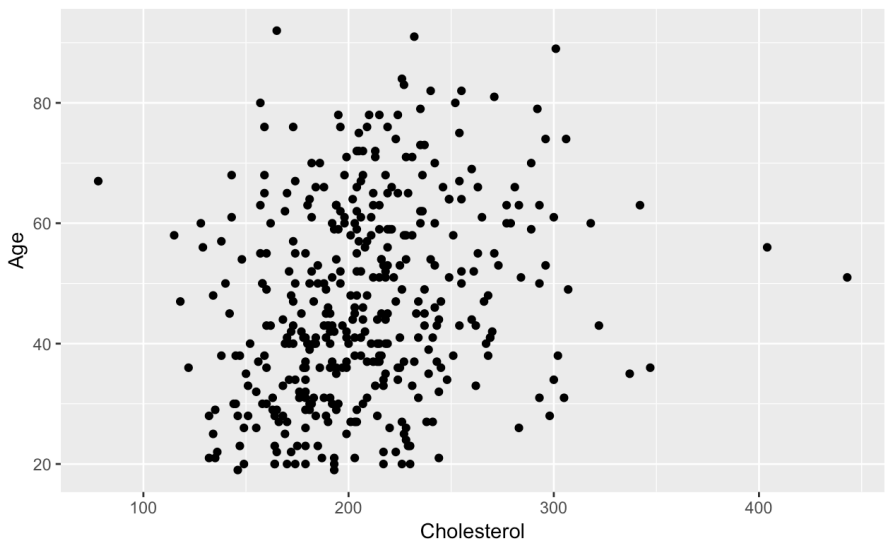
id	chol	stab.glu	hdl	ratio	glyhb	location	age
1000	203	82	56	3.60	4.31	Buckingham	46
1001	165	97	24	6.90	4.44	Buckingham	29
1002	228	92	37	6.20	4.64	Buckingham	58
1003	78	93	12	6.50	4.63	Buckingham	67
1005	249	90	28	8.90	7.72	Buckingham	64
1008	248	94	69	3.60	4.81	Buckingham	34
1011	195	92	41	4.80	4.84	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

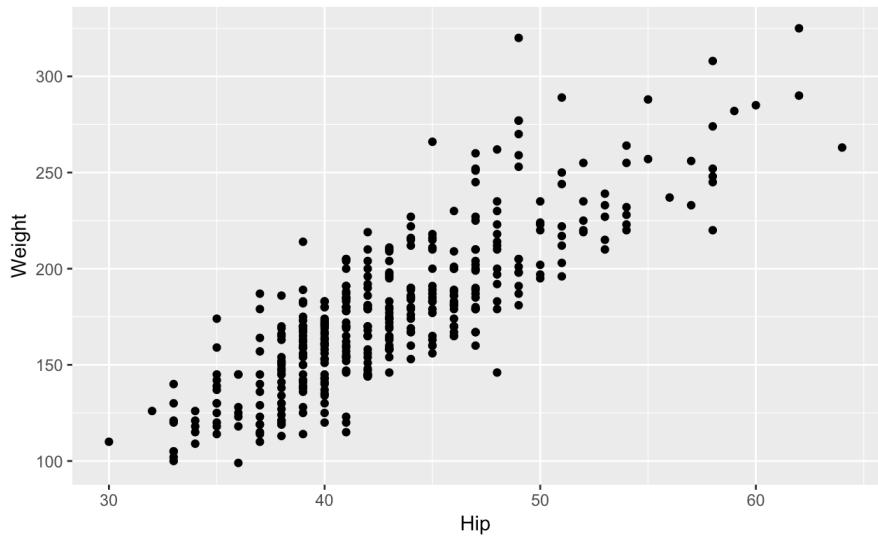


Numerical data

comparing variable with scatter plots



Weak relationship
between Cholesterol and age



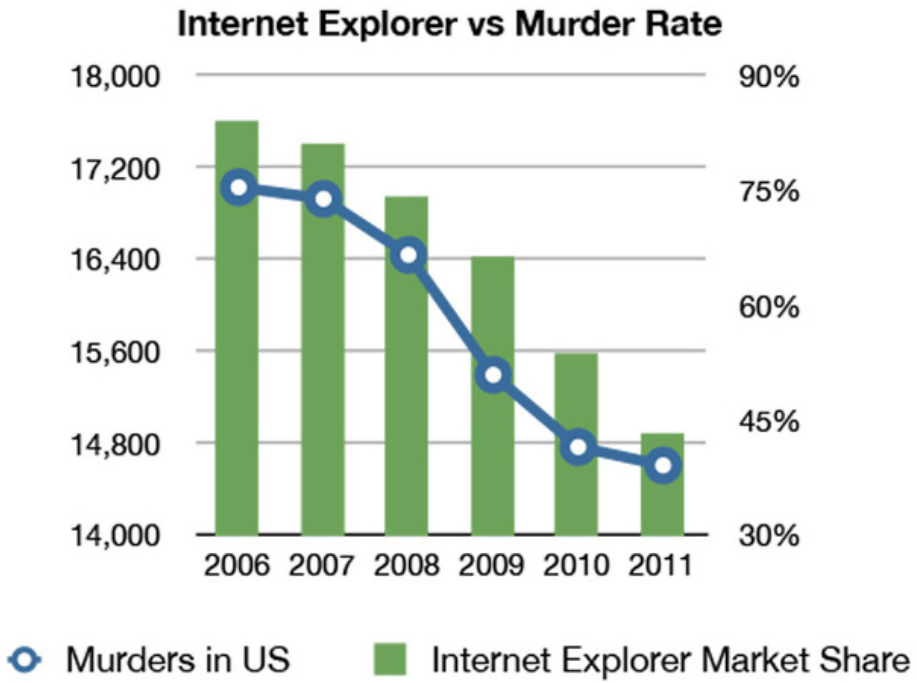
Strong relationship
between Hip and Weight

- 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

Medizinische Fakultät Heidelberg

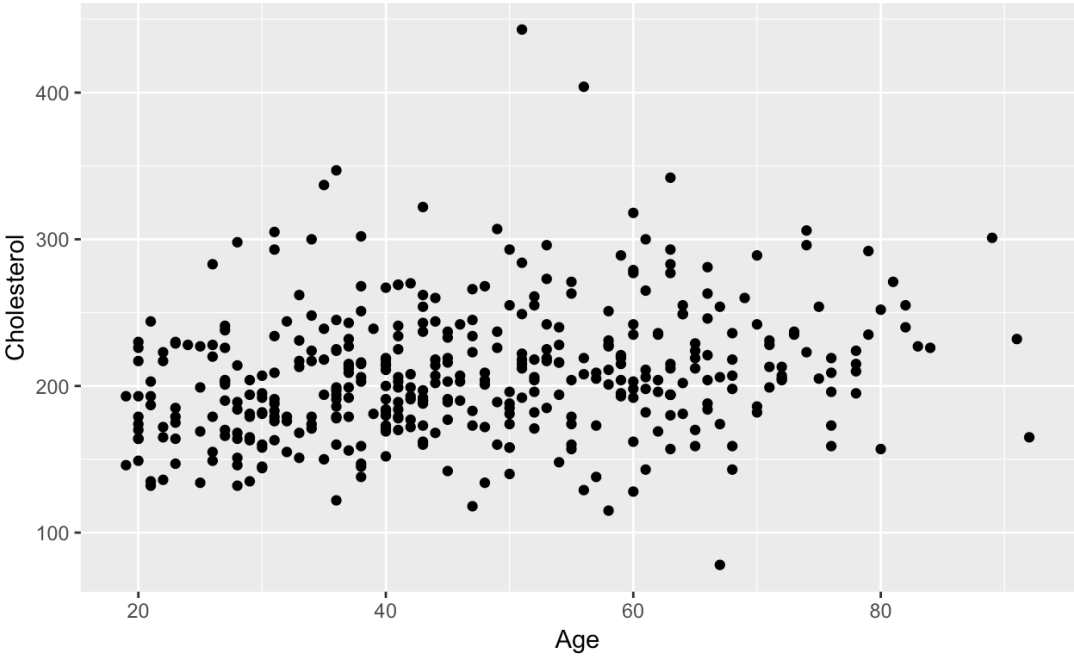
- Do not over-interpret scatter plots!
- Existence of relation between variables does not mean that there is a causal relationship between them!
- Correlation is NOT causality!!

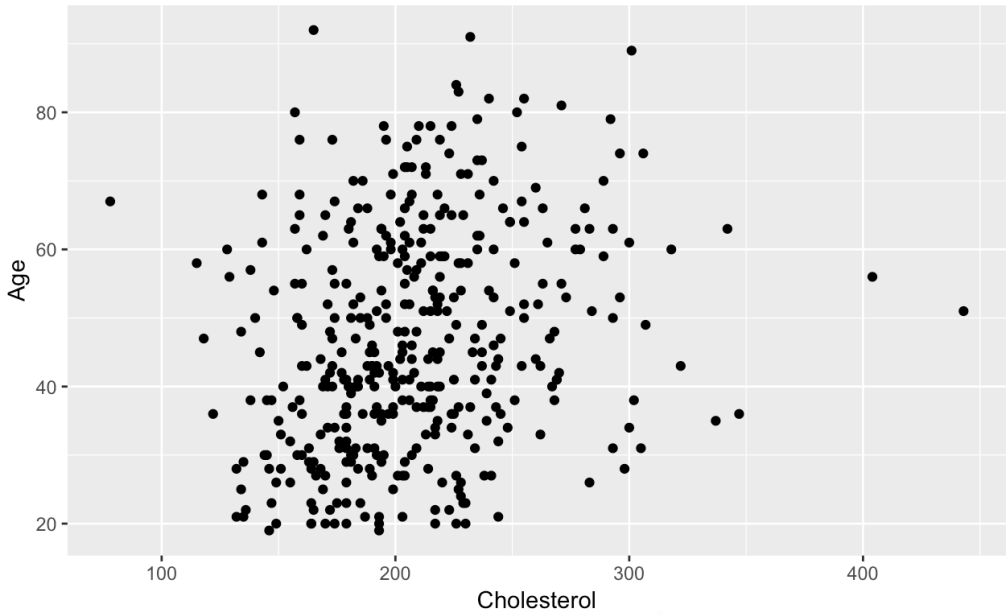


http://www.tylervigen.com/spurious-correlations

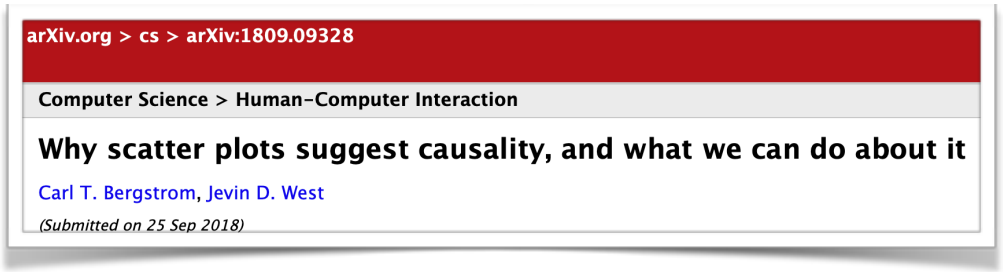
Numerical data

comparing variable with scatter plots





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

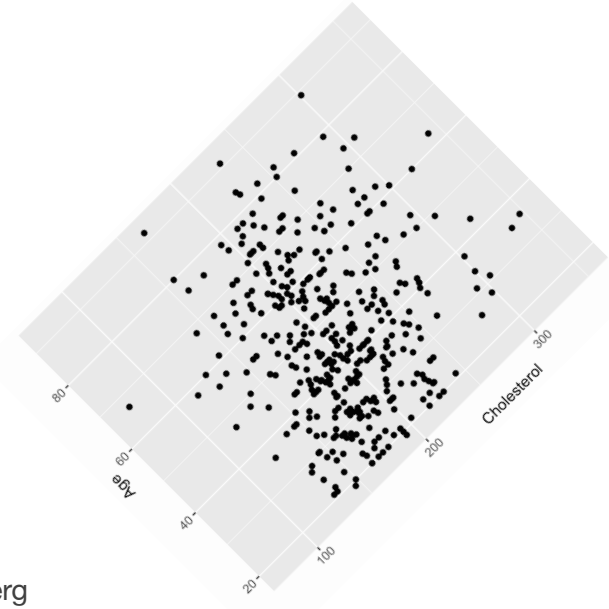

arXiv.org > cs > arXiv:1809.09328

Computer Science > Human-Computer Interaction

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

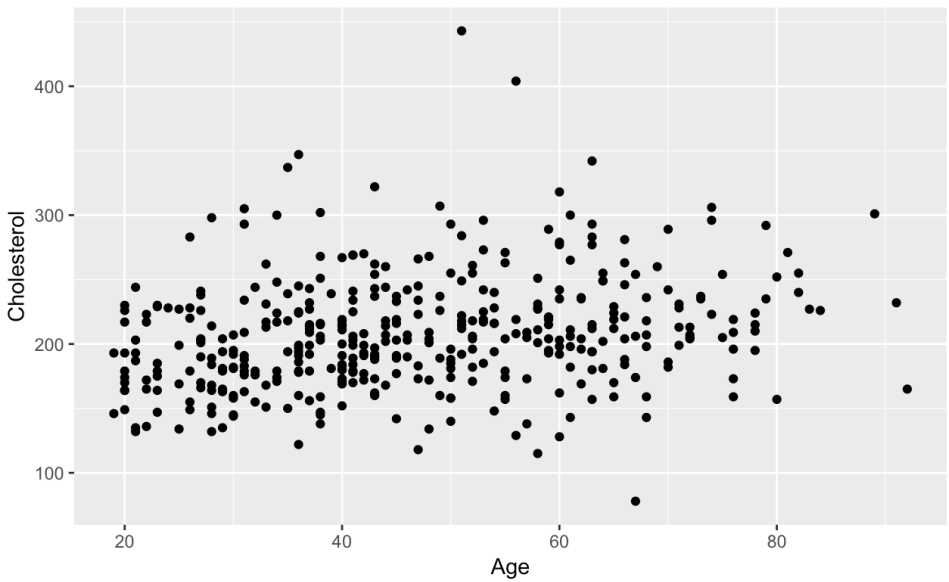
Carl T. Bergstrom, Jevin D. West

(Submitted on 25 Sep 2018)

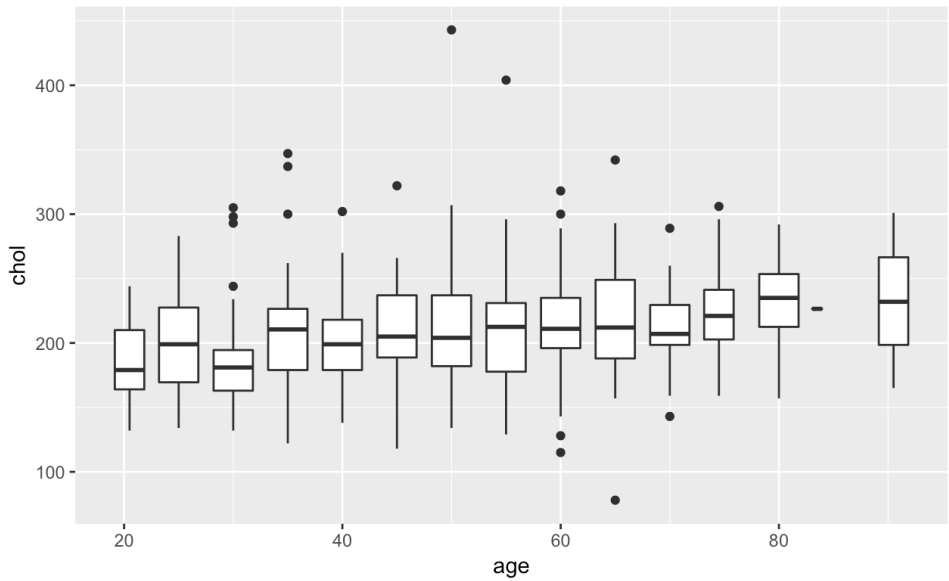


Numerical data

comparing variable with scatter plots



Age = continuous variable

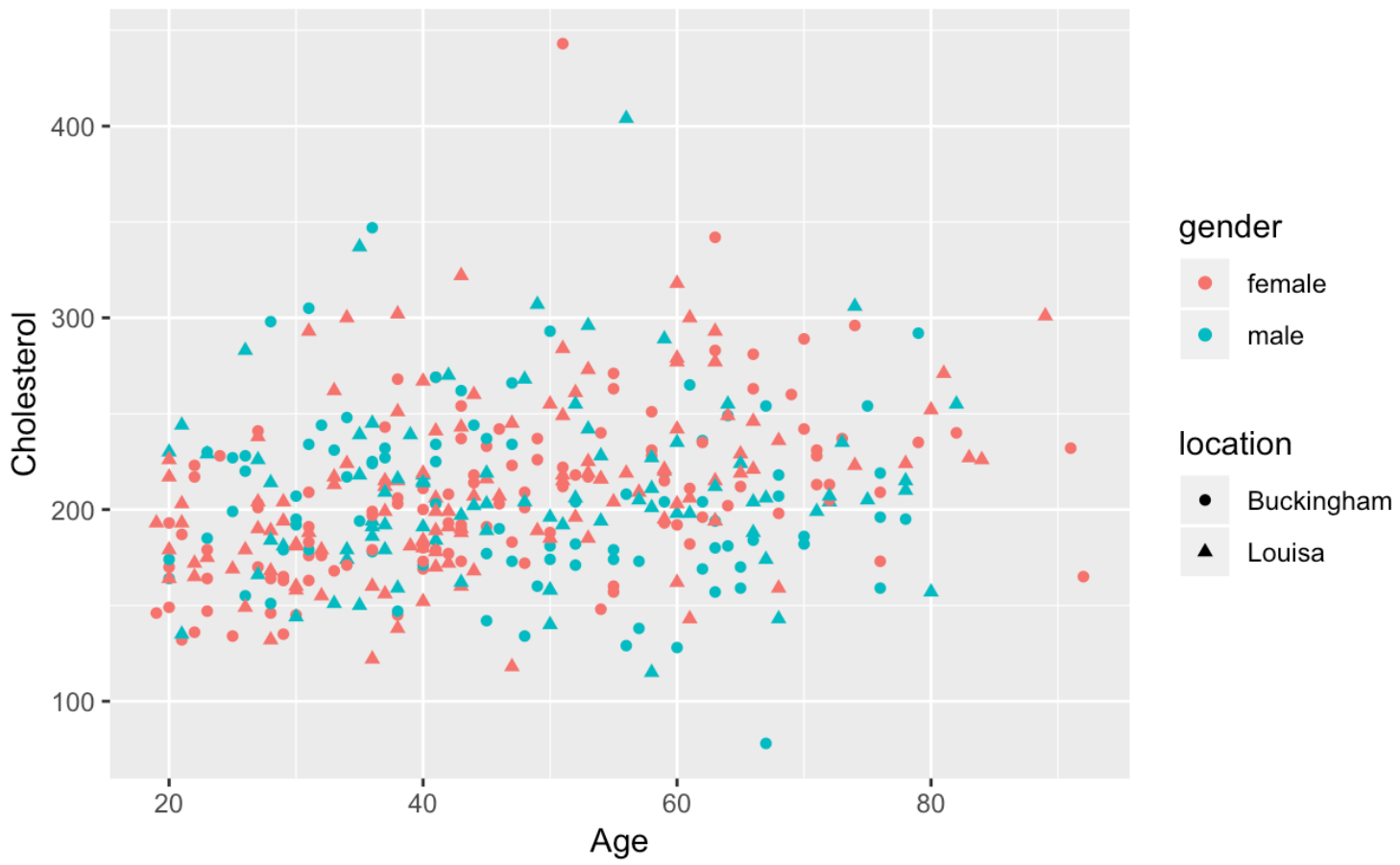


Age = ordinal variable (bins)

- 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

Medizinische Fakultät Heidelberg

2. describing data

Descriptive statistics

Medizinische Fakultät Heidelberg

- Datasets represent samples taken from a larger population
- Example
  - diabetes dataset: sample representing a group of diabetes patients
  - gene expression dataset: samples (e.g. 10 mice before and after treatment) represent a group of mice extracted from a larger population
  - opinion polls: opinion of 1000 people extracted from a more general population

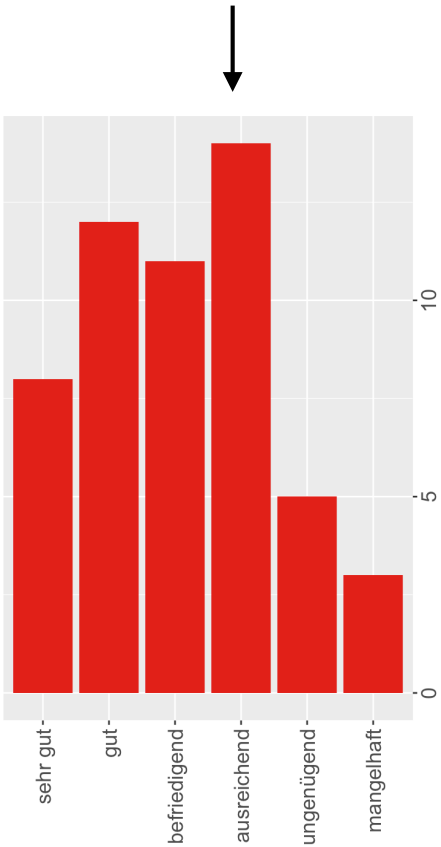
$$\{x_1, x_2, \ldots, x_N\}$$

Descriptive statistics provides a description of the sample dataset, without making any extrapolation to the more general population!

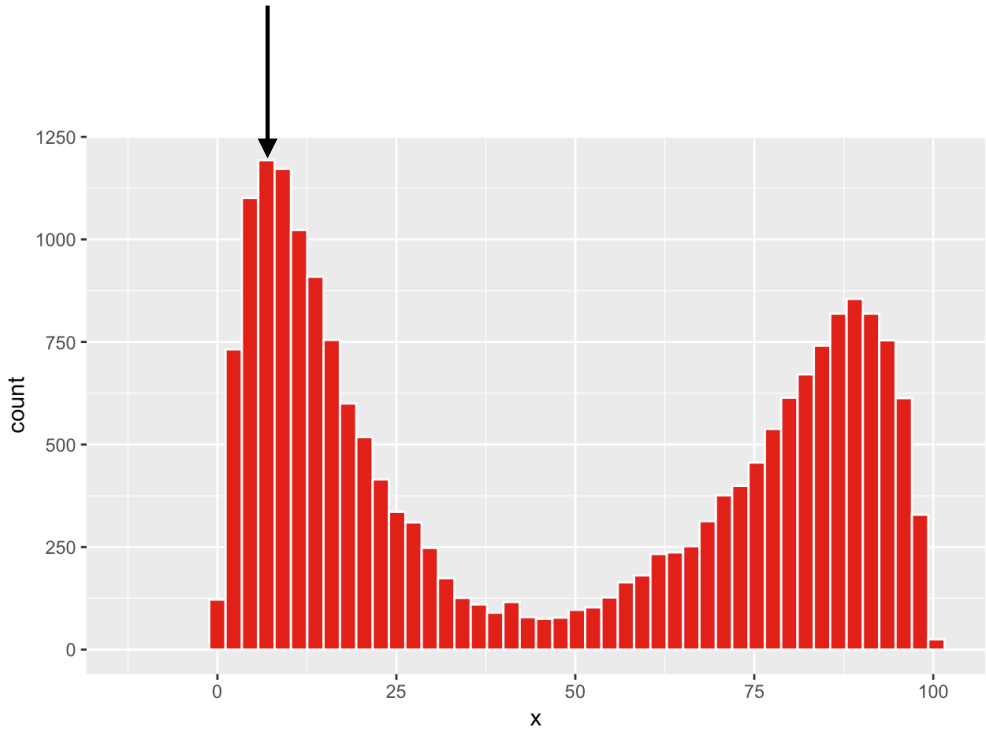
Computing the average

- Categorical data: **mode** = category with the highest count

Mode



Mode



Such a distribution is called a ***"bimodal distribution"***

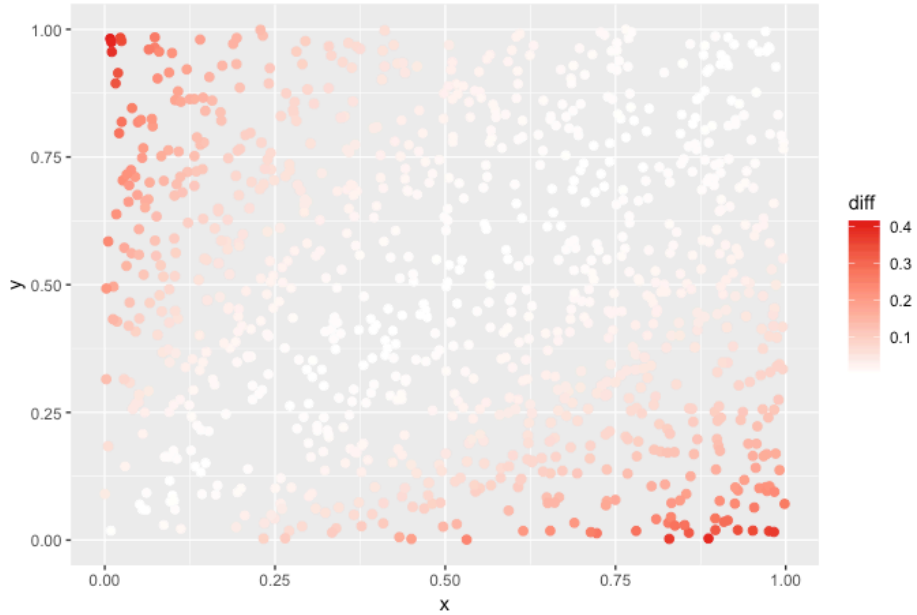
Computing the average

- Numerical data: **mean** of the data

$$\text{arithmetic mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

- arithmetic mean is always greater/equal the geometric mean

difference arithm. - geom.
for 2 values (x,y)



Computing the average

- Numerical data: **median**
- median = numerical value which splits the dataset into two equal parts
  - 50% of the values are larger, 50% of the values are smaller that the median value
- **Odd** number of values : sort and take the middle one

8th value

-2.52 -2.14 -1.67 -0.47 -0.43 -0.33 -0.10 **-0.06** 0.26 0.28 0.31 0.34 0.41 0.47 0.63

15 values

- **Even** number of values: sort and mean of the middle 2 values

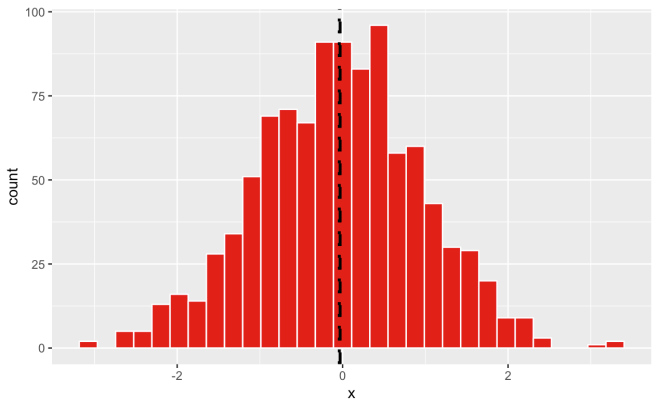
mean = 0.435

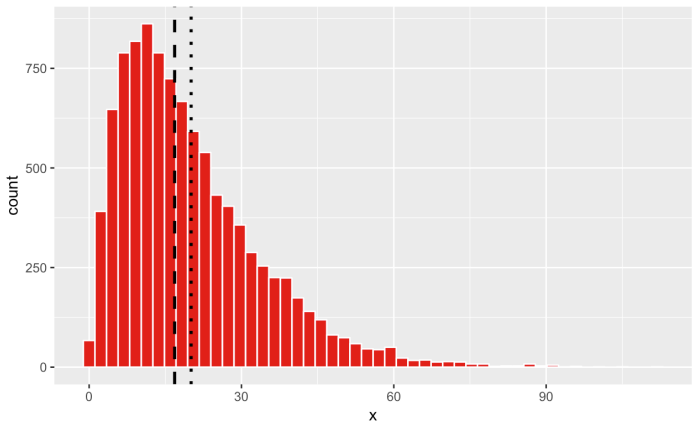
-1.45 -1.15 -0.88 -0.64 0.24 0.34 0.42 0.45 0.46 0.67 0.84 1.59 1.73 2.19

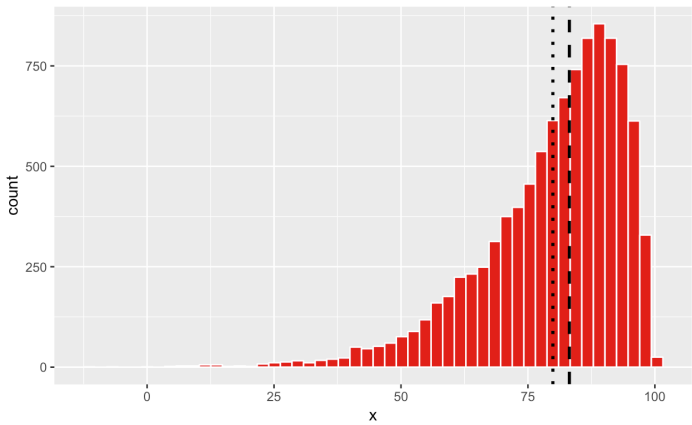
14 values

Mean / Median

Which line is the median, which is the mean ?





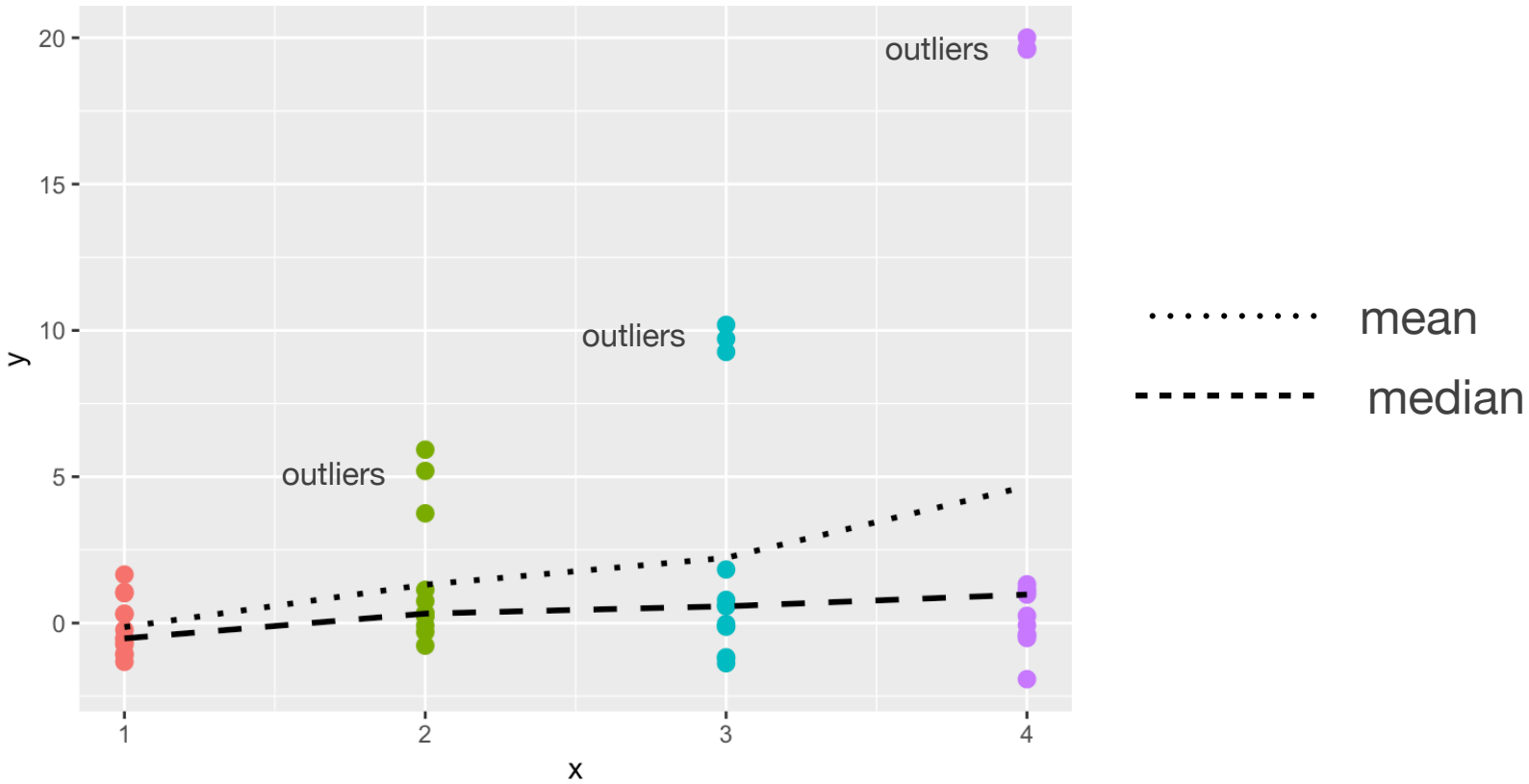


symmetrical distribution
median ~ mean

right-skewed distribution
mean > median

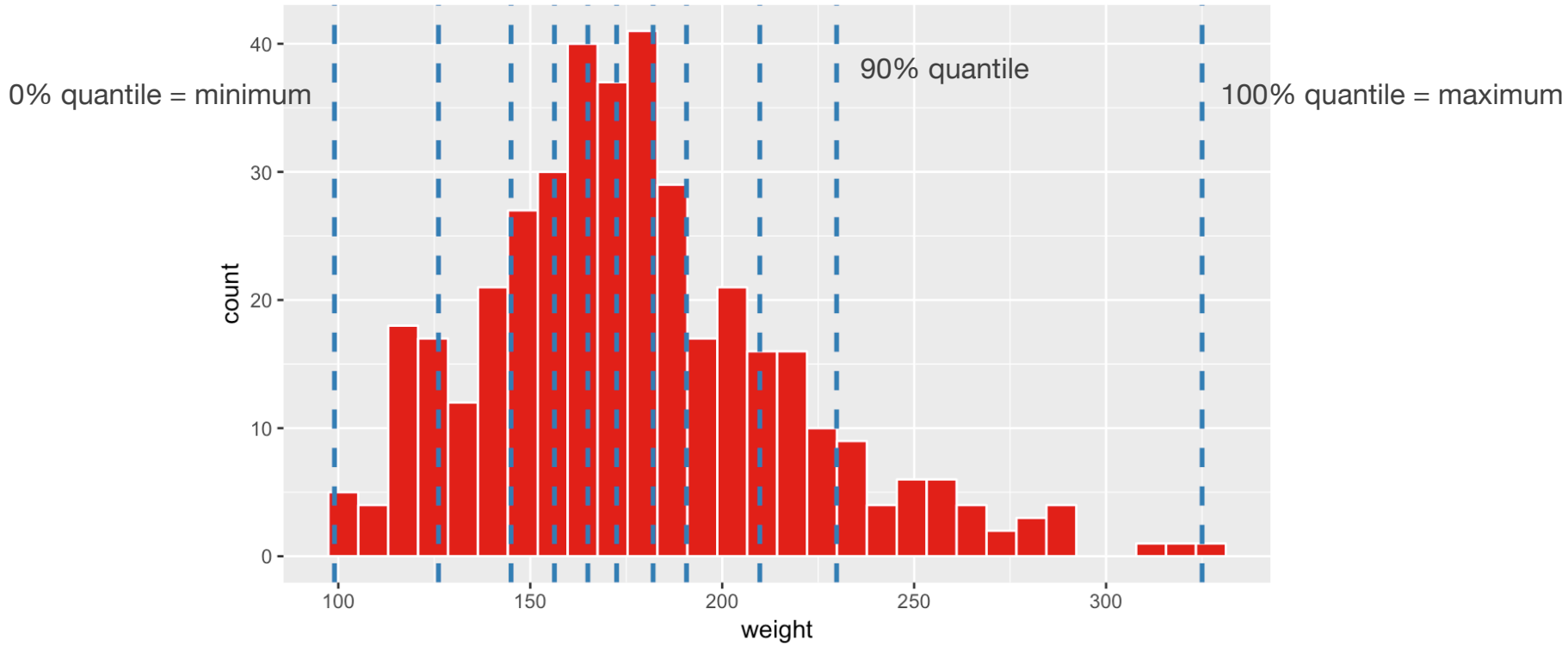
left-skewed distribution
mean < median

Mean vs. median



- 4 groups, with 3 outliers with increasing effect
- *median is more robust to outliers than mean*

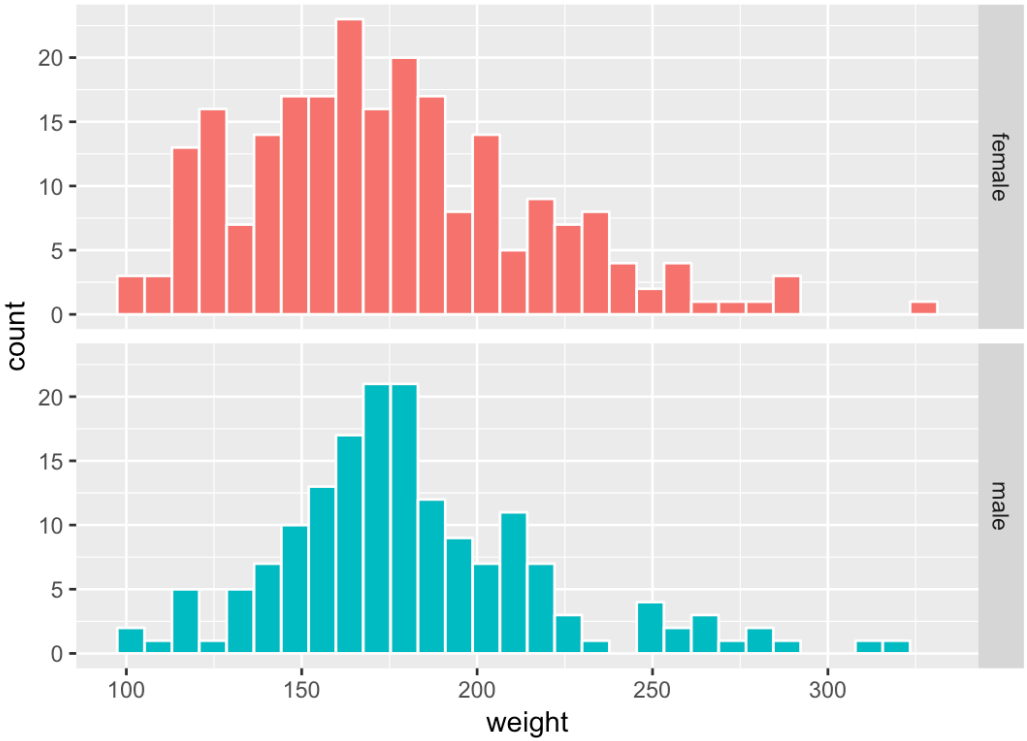
Quantiles



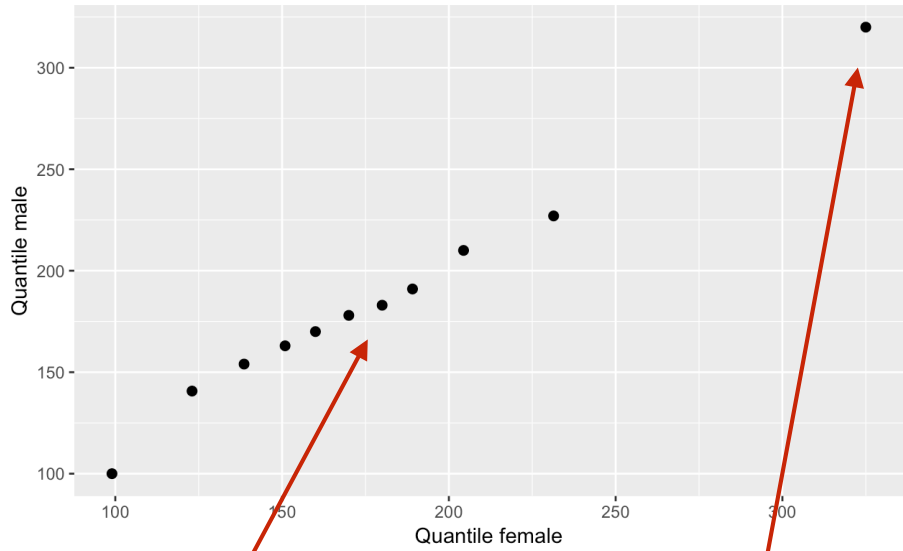
- **p-quantile** = value such that p% of the values in the data set are smaller
- distribution of quantile value given an indication about the **shape** of the distribution

Quantile-quantile plots

weight distribution men / women



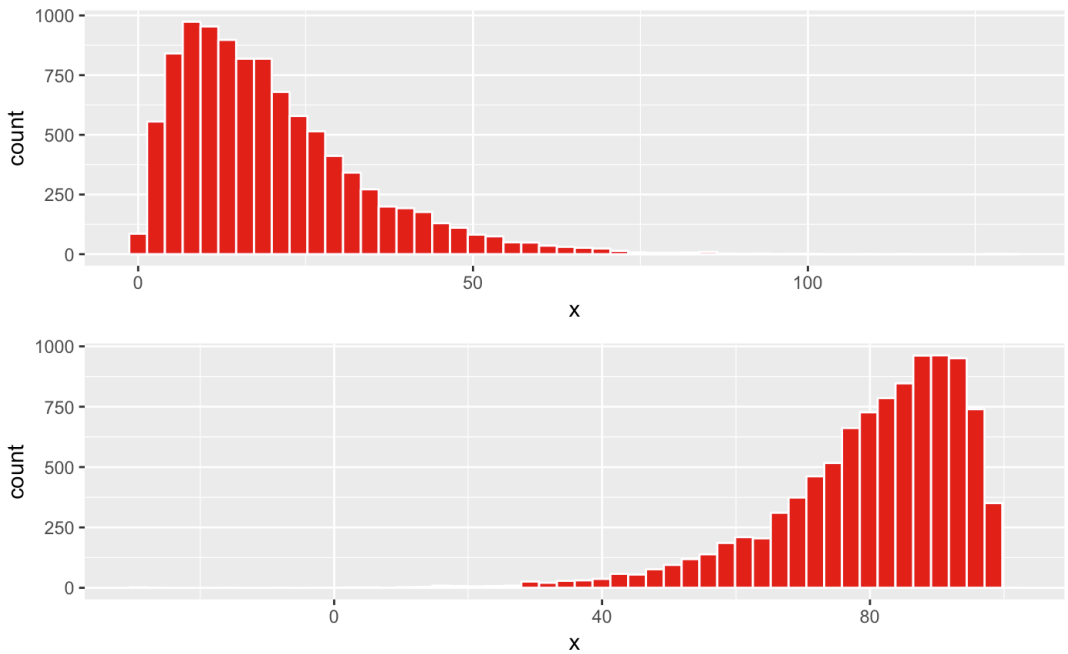
quantile-quantile plot

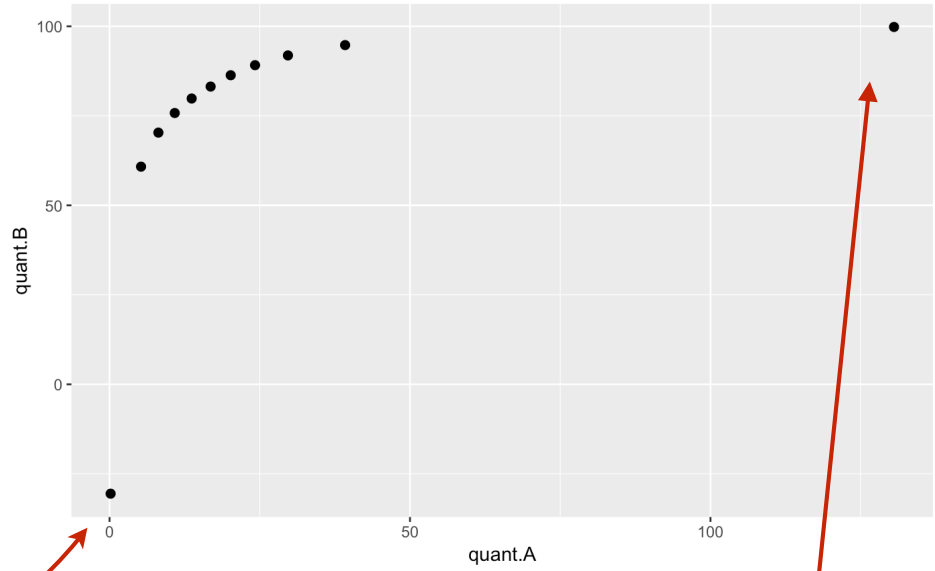


	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
quant.female	99	123.0	138.6	150.9	160	170	180	189.1	204.4	231.4	325
quant.male	100	140.7	154.0	163.0	170	178	183	191.0	210.0	227.0	320

if QQ-plot is a straight line, then both distributions have a similar shape (up to translations / dilatation)

Quantile-quantile plots





	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
quant.A	0.18	5.24	8.13	10.85	13.66	16.82	20.16	24.22	29.69	39.20	130.55
quant.B	-30.55	60.80	70.31	75.78	79.84	83.18	86.34	89.15	91.87	94.76	99.82

Dealing with weird values

Medizinische Fakultät Heidelberg

- Datasets originate from measurements, and might contain **unexpected values**
  - missing values
  - outliers
- Examples:
  - cholesterol measurement for one patient was not made and is missing
    → *how is this indicated in the dataset? 0 or "NA" (not available)*
  - machine was wrongly calibrated for one experiment and returned an unusually high value
    → *technical artifact or biologically relevant phenomenon?*
- Outliers might have a **technical** origin (artifacts) or a **biological** origin (interesting!)

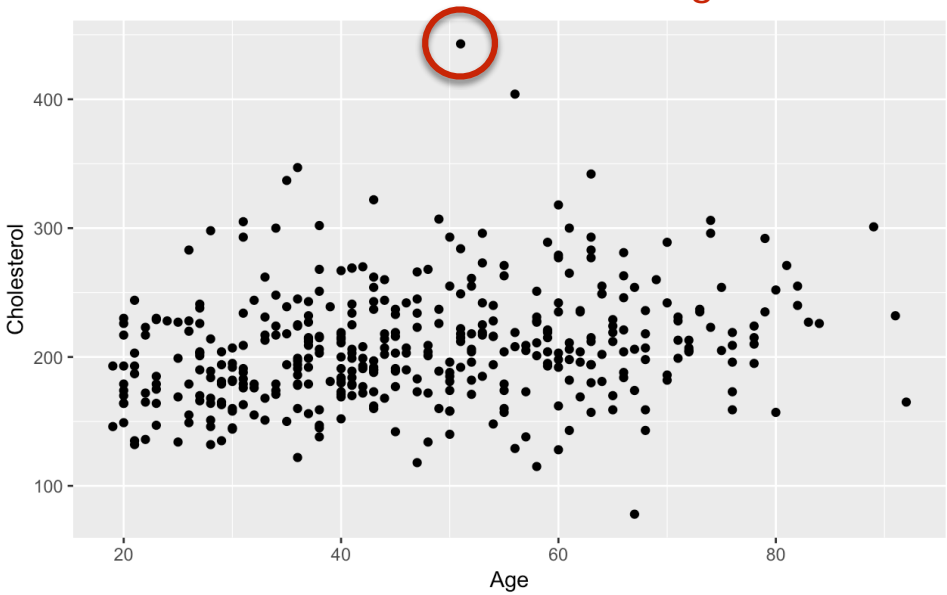
Dealing with weird values

Medizinische Fakultät Heidelberg

missing NA values

id	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1000	62	121	medium	118	59	NA	NA	29	38	720
1001	64	218	large	112	68	NA	NA	46	48	360
1002	61	256	large	190	92	185	92	49	57	180
1003	67	119	large	110	50	NA	NA	33	38	480
1005	68	183	medium	138	80	NA	NA	44	41	300
1008	71	190	large	132	86	NA	NA	36	42	195
1011	69	191	medium	161	112	161	112	46	49	720

technical or biological outlier?



- Solutions:
  - remove rows / columns containing NA values
  - replace missing values with “likely” values (e.g. mean of present values) → **imputation**

- Solutions:
  - if artifact, use measures that are less sensitive to outliers
  - remove outlier values

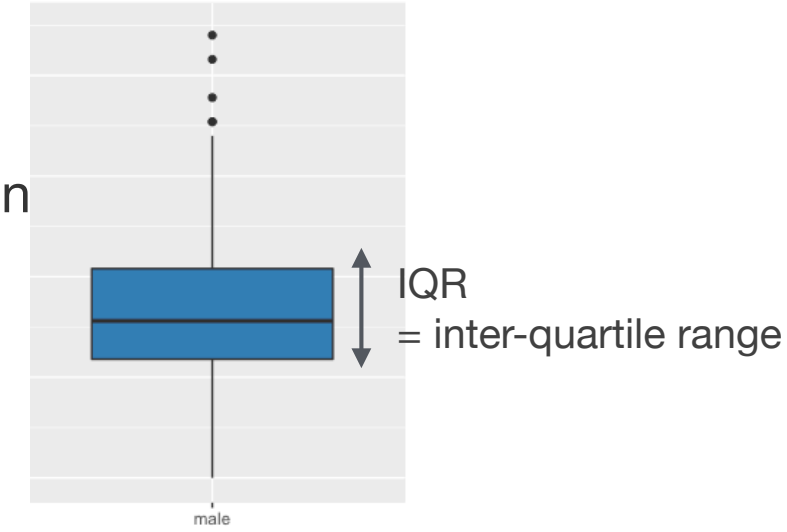
Measuring data spread

- How much spread is there in the dataset?

- **Sample variance** : $$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- **Standard deviation** : $$s_x = \sqrt{Var(x)}$$

- **Interquartile Range (IQR)** : difference between the 75% and 25% percentile

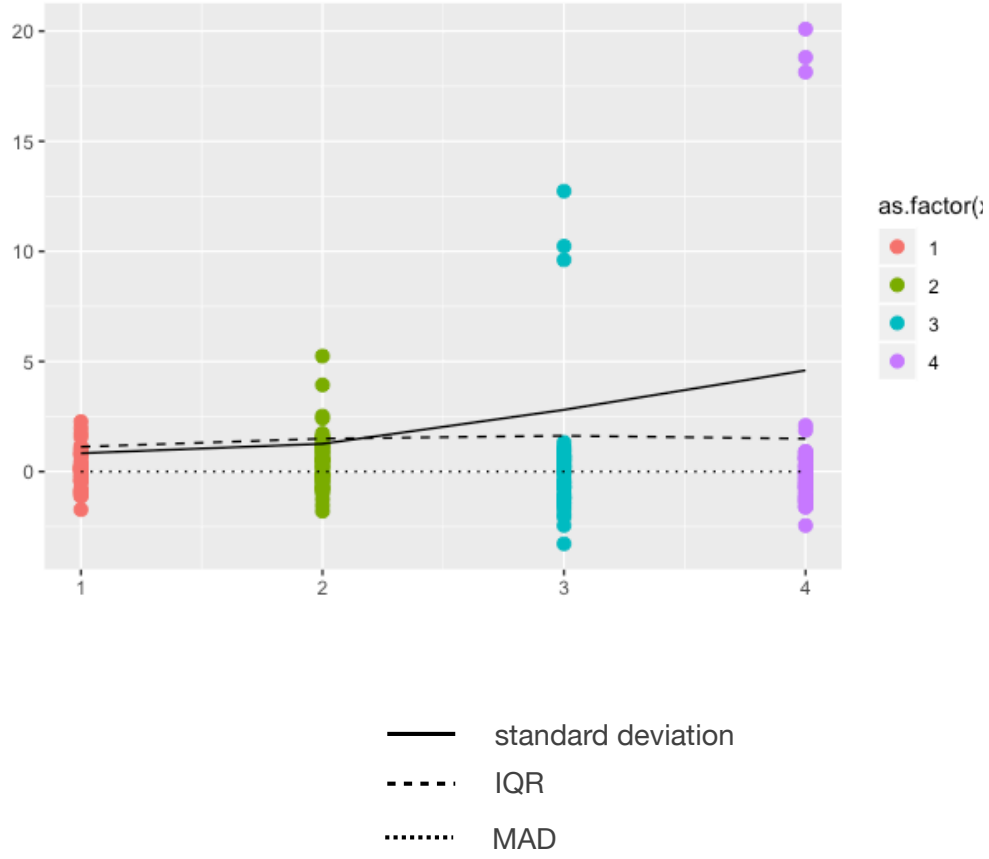


Measuring data spread

- **Median average deviation (MAD)**
  median deviation of the data points from the median…

- How to compute?
  1. compute the median *med(x)* of the dataset
  2. compute the deviation of all points from the median $d_i = x_i - med(x)$
  3. compute the median of the d_i

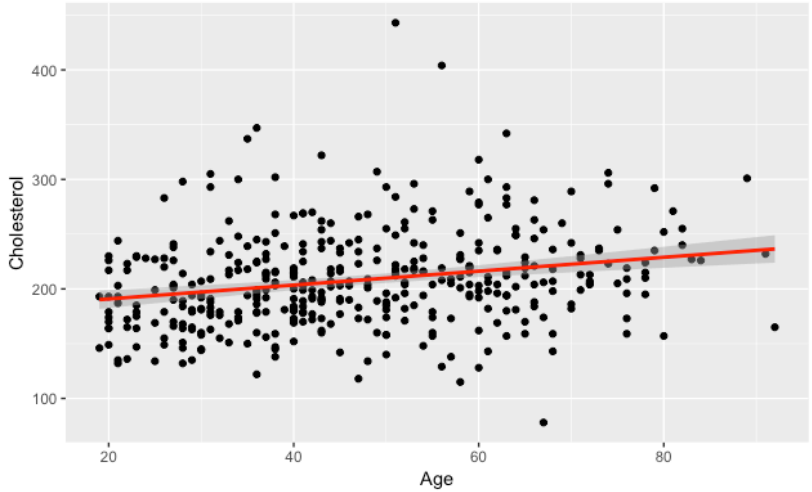
- IQR and MAD are less sensitive to outliers than standard deviation !

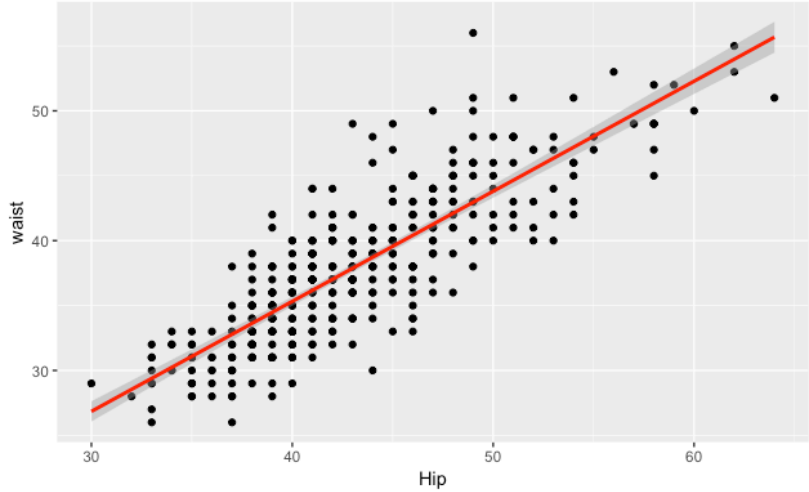


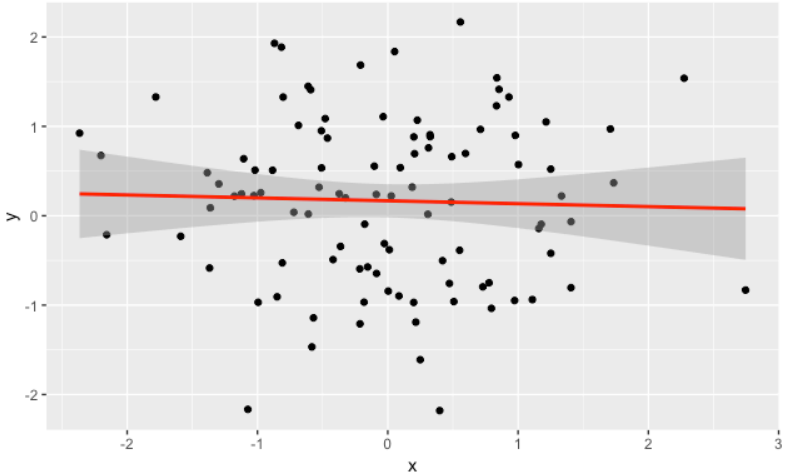
Relation between numerical variables

Medizinische Fakultät Heidelberg

- How easy is it to draw a line through a scatter plot?







Relation between numerical variables

Medizinische Fakultät Heidelberg

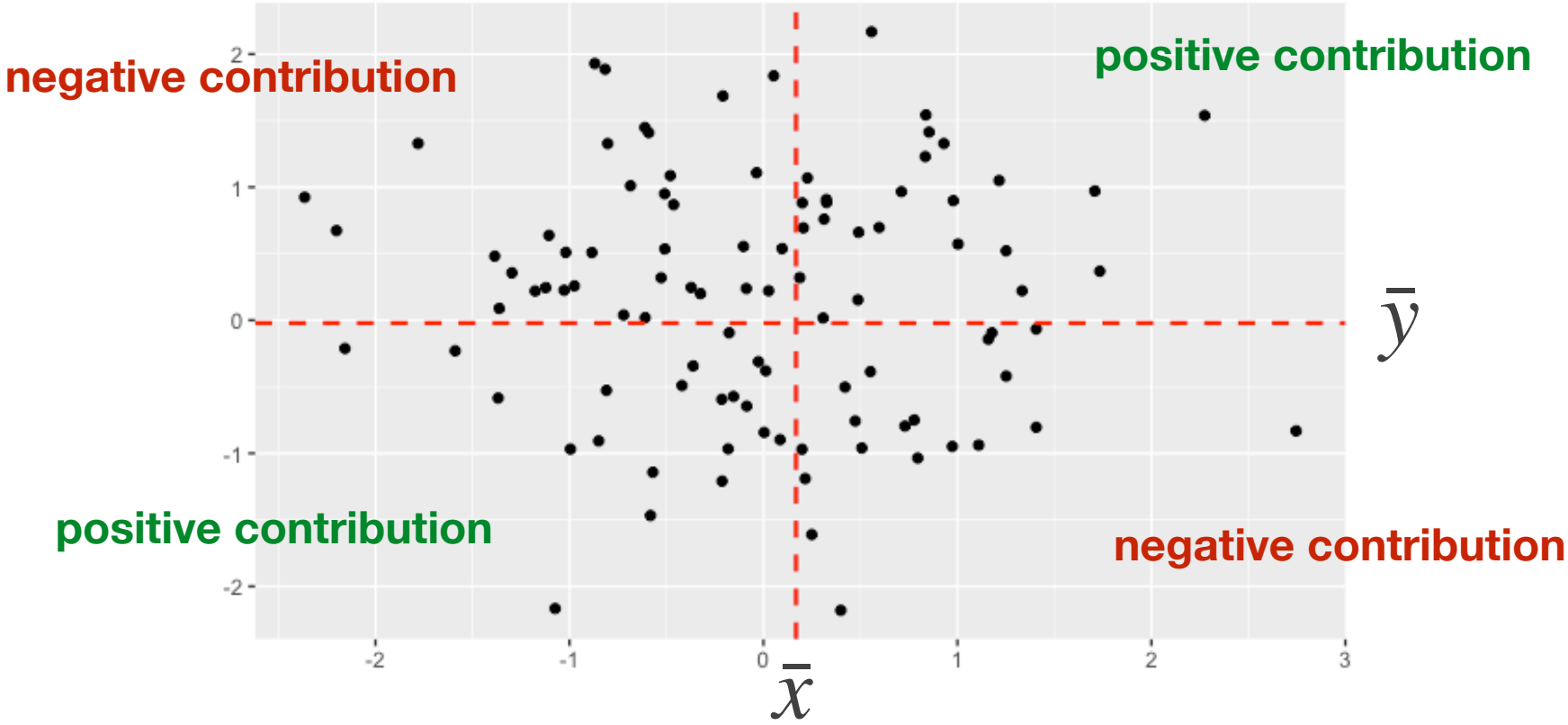
- Variance: $$Var(x) = (s_x)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$ dimension: $[x]^2$

- Covariance : $$Cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$ dimension: [x][y]

- Pearson Correlation : $$Corr(x,y) = r = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}$$ dimension: none

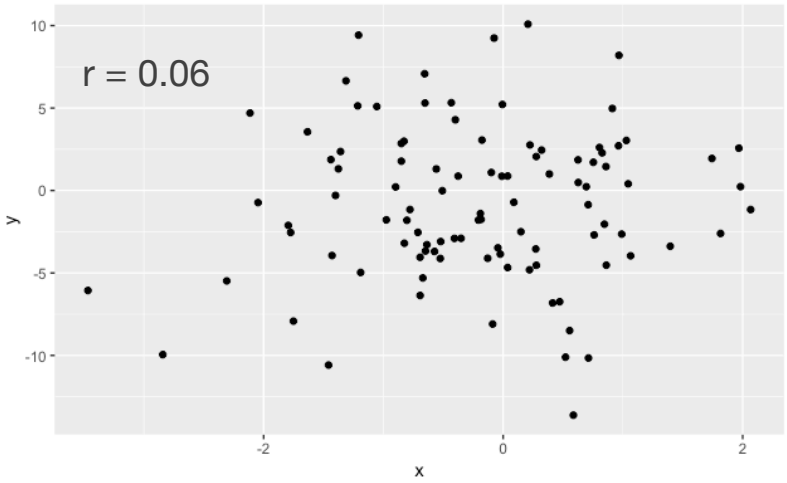
- Properties:
  - correlation is scale invariant, covariance is not!
  - cor(x,x) = 1
  - -1 =< cor(x,y) =< +1

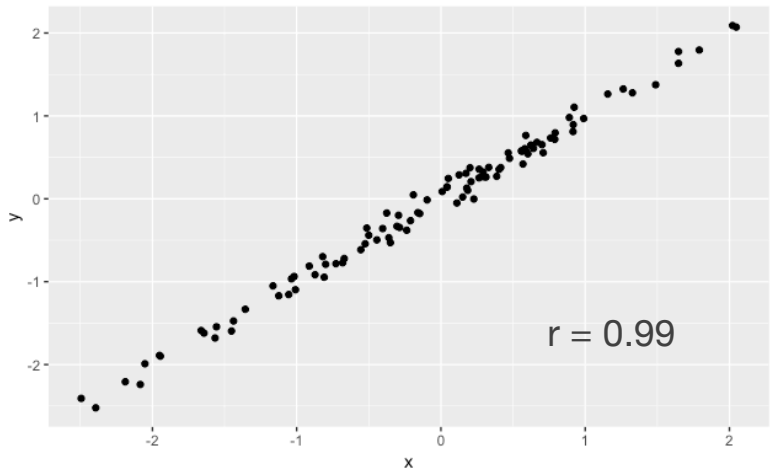
Relation between numerical variables

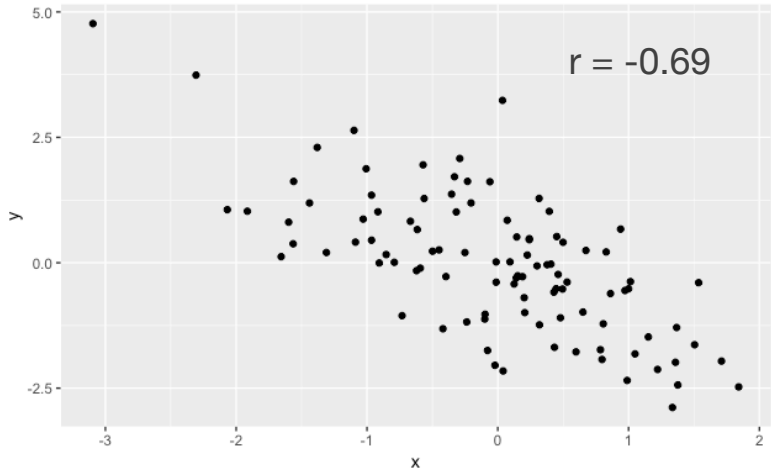


$$Corr(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

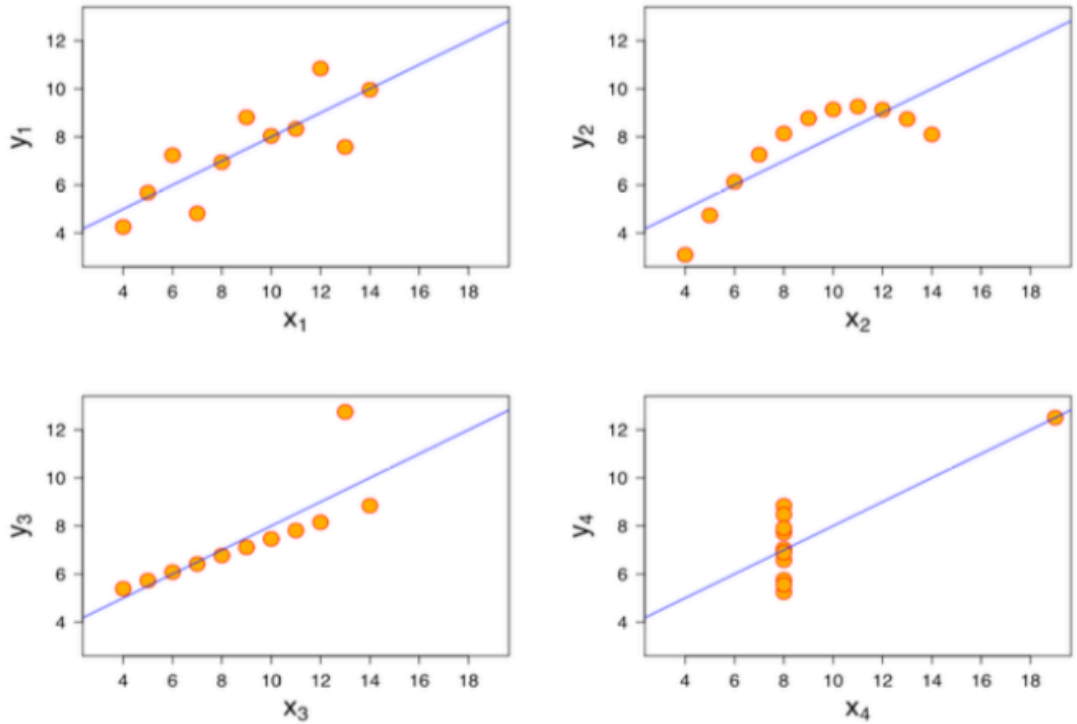
Relation between numerical variables







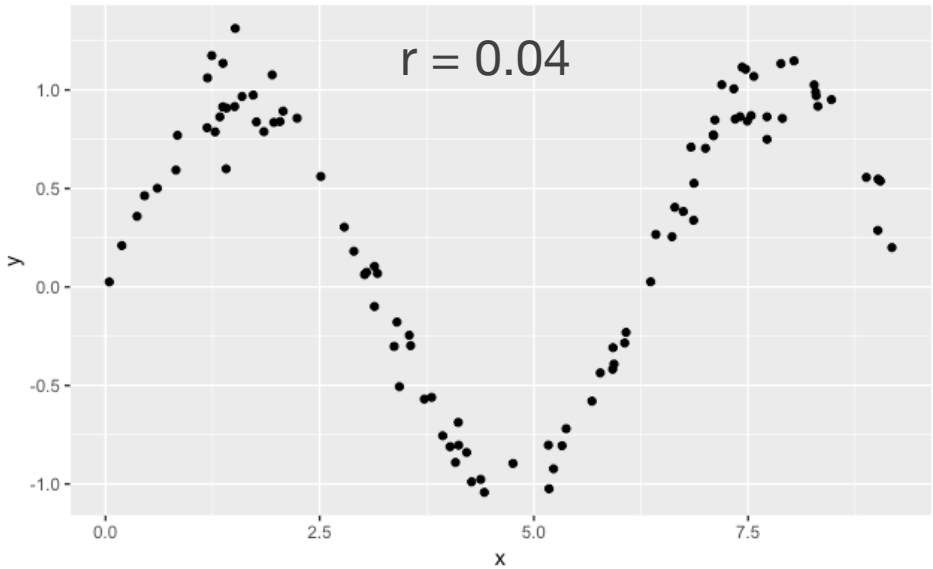
Relation between numerical variables



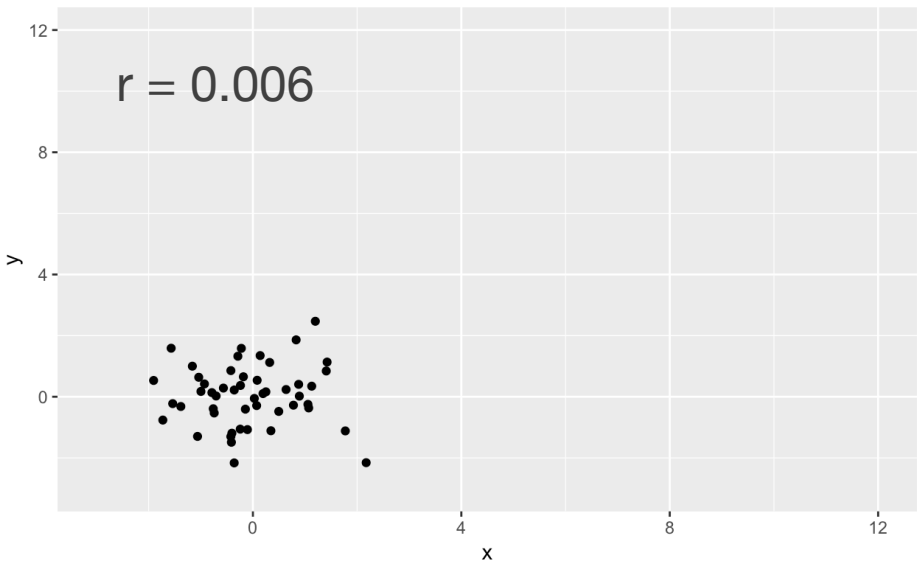
- Anscombe quartet is a dataset to highlight possible caveats of correlation measures
- Correlation is equal in all 4 cases (r = 0.82)

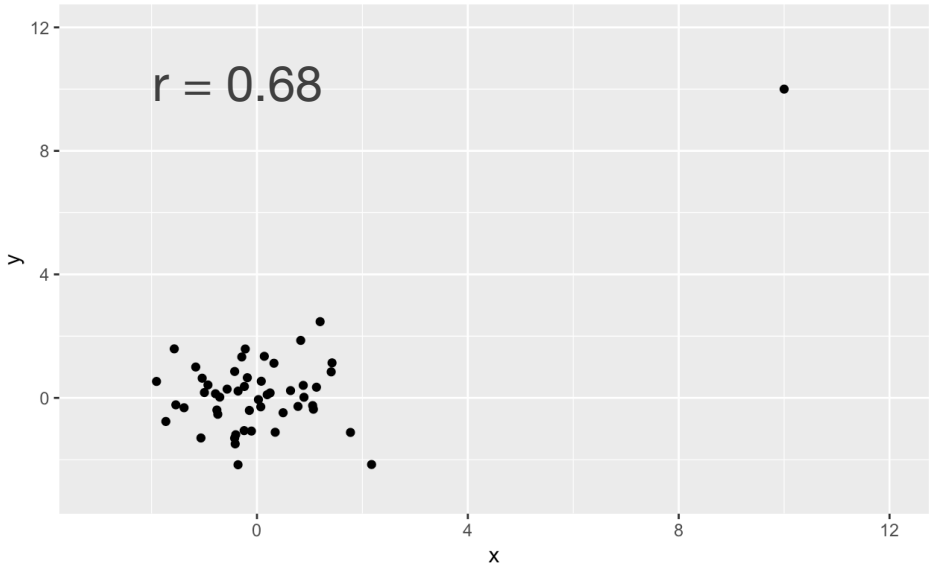
Relation between numerical variables

- Beware that correlation estimates a **linear relationship**!
- Weak correlation **does NOT mean** that there is no relation between both variables!



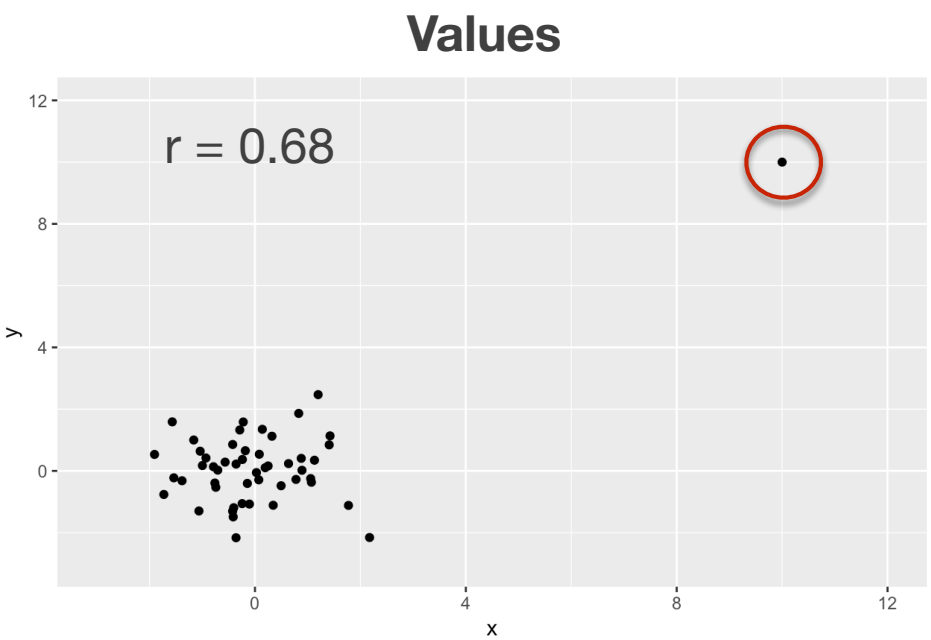
Correlation and outliers

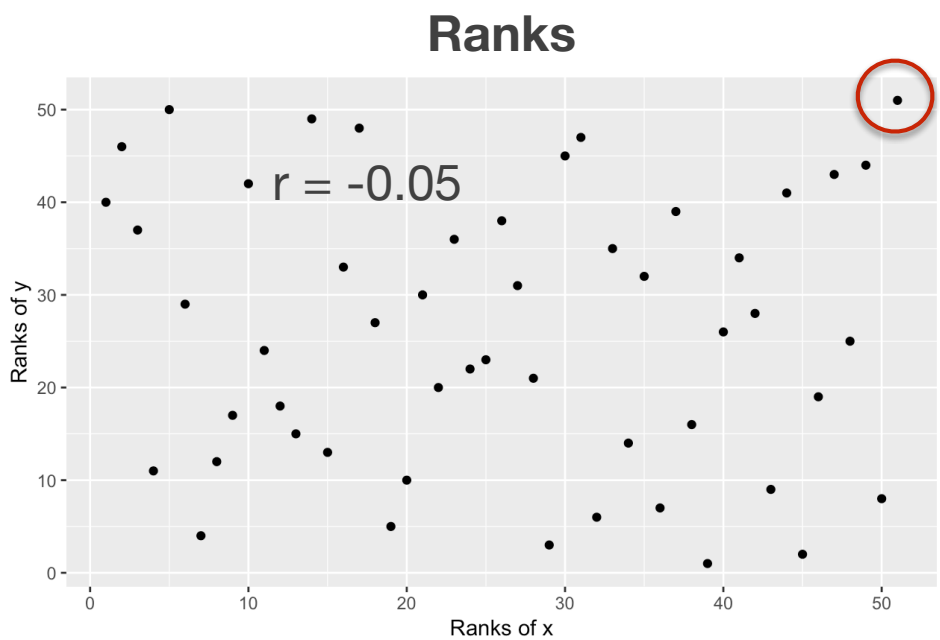




- Pearson correlation is strongly influenced by outliers!
- we need to find robust measure which is insensitive to outliers.

Correlation and outliers





- Transform values to ranks

  x = {2.3 , -1.7 , 1.1 , -0.6 , 12.3} → R(x) = {4 , 1 , 3 , 2 , 5}

  y = {-2.7 , 0.8 , 3.1 , -1.7 , 14.3} → R(y) = {1 , 3 , 4 , 2 , 5}

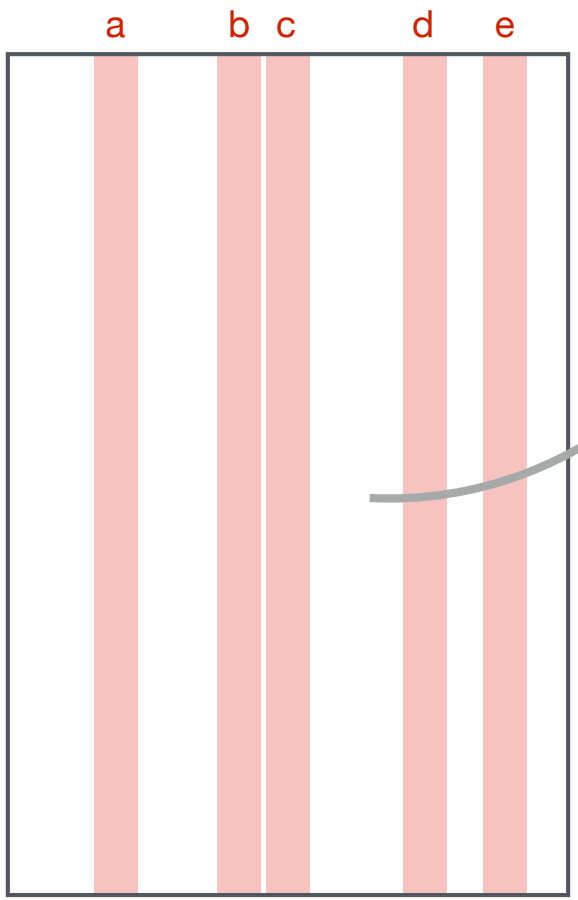
- Compute the correlation **on the ranks :** cor(R(x) , R(y))

  = **Spearman correlation**

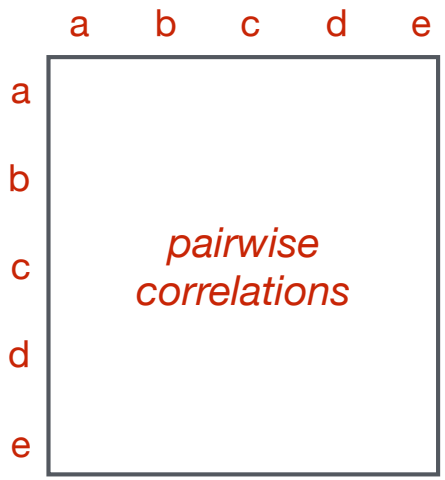
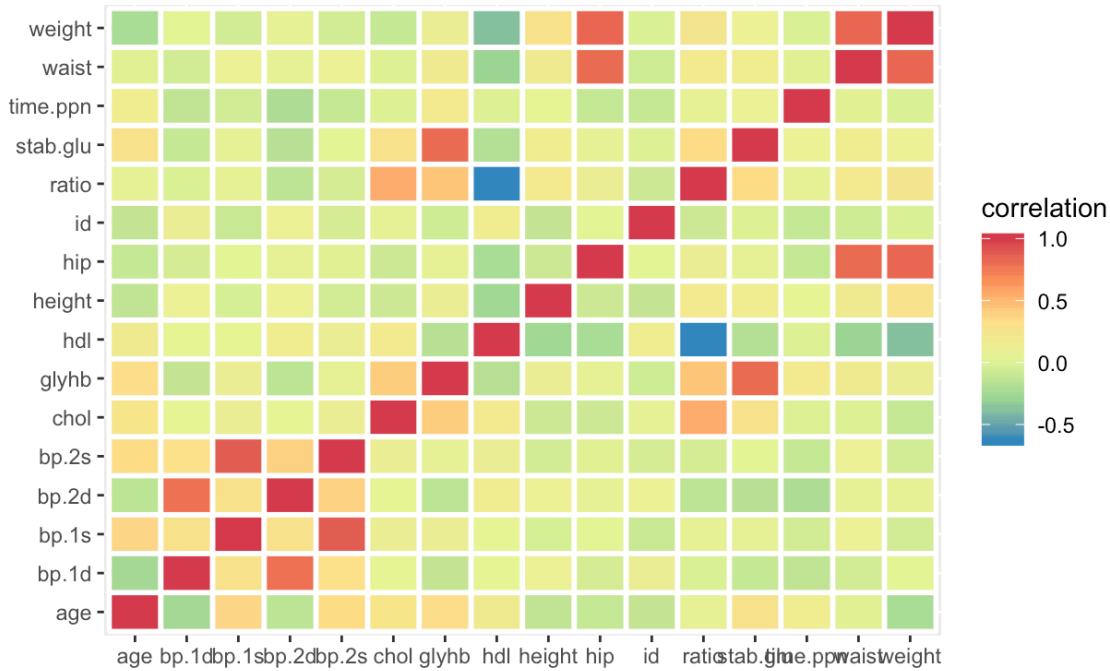
Correlating multiple variables

Medizinische Fakultät Heidelberg

Diabetes dataset



numeric columns



pairwise correlations