Introduction to R for data analysis

- cleaning data -

Carl Herrmann & Carlos Ramirez

IRTG Course - December 2021

Dealing with weird values

- 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

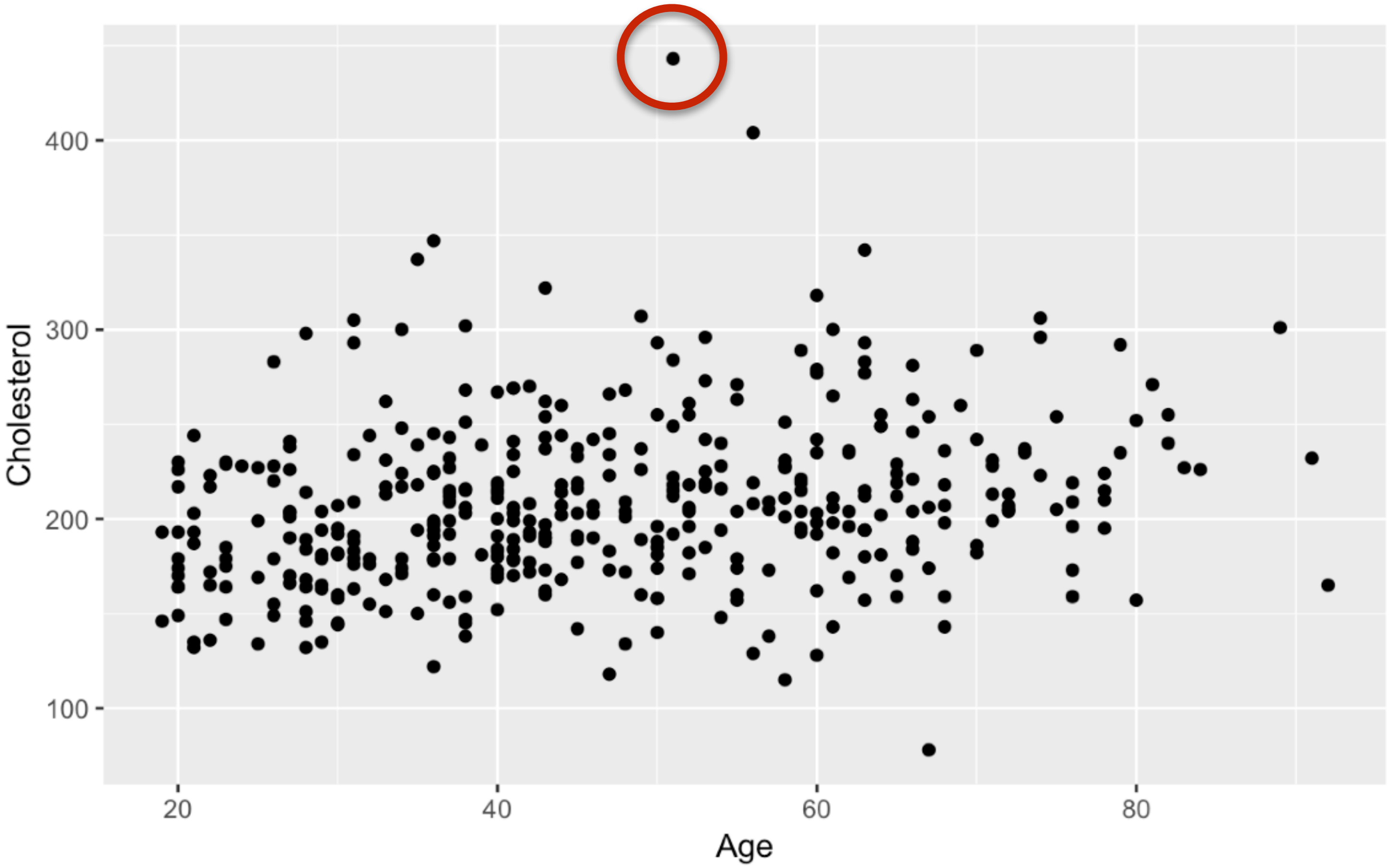
id	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1000	62	121	medium	118	59	**NA**	**NA**	29	38	720
1001	64	218	large	112	68	**NA**	**NA**	46	48	360
1002	61	256	large	190	92	185	92	49	57	180
1003	67	119	large	110	50	**NA**	**NA**	33	38	480
1005	68	183	medium	138	80	**NA**	**NA**	44	41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

missing NA values

- Solutions:
  - remove rows / columns containing NA values
  - replace missing values with “likely” values (e.g. mean of other available values)
    → **imputation**
- Possible imputation strategies
  - replace missing value with **mean** of existing values
  - draw **random value** from theoretical distribution
  - **predict** missing values from other variables

Dealing with weird values

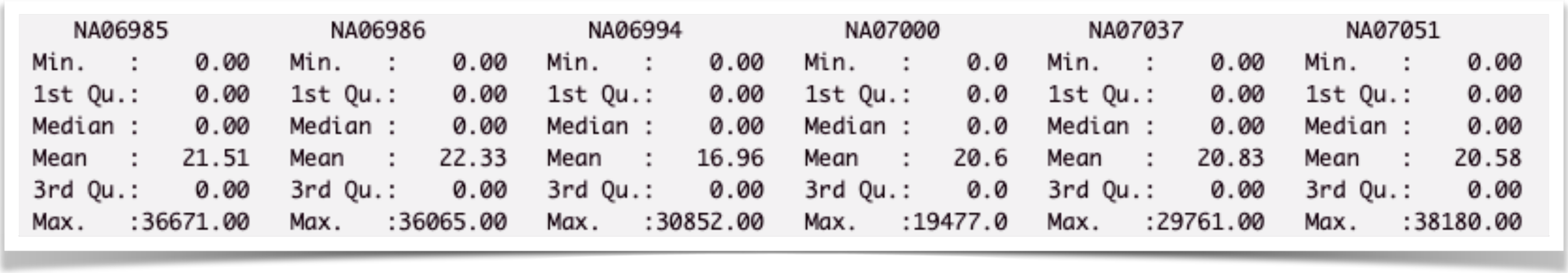
technical or biological outlier?



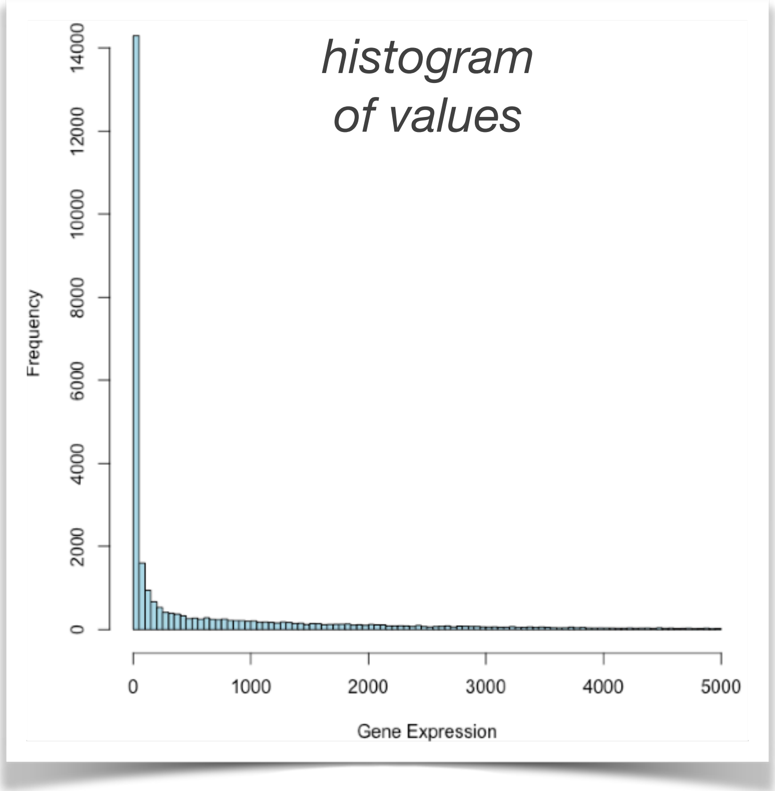
- Solutions:
  - if artifact, use measures that are less sensitive to outliers
  - **remove** outlier values from dataset
  - **replace** outlier value with more likely value (see imputation strategies)

Dealing with skewed data

- Many data type have **heavily skewed distributions**

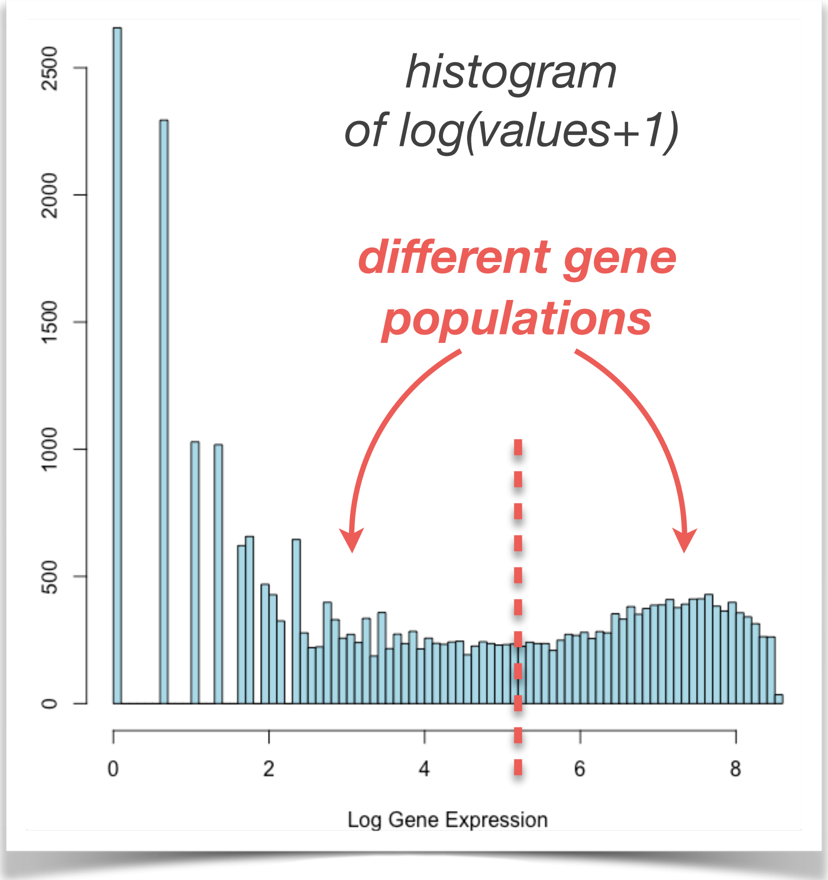


```
    NA06985           NA06986           NA06994           NA07000          NA07037           NA07051
 Min.   :    0.00  Min.   :    0.00  Min.   :    0.00  Min.   :    0.0  Min.   :    0.00  Min.   :    0.00
 1st Qu.:    0.00  1st Qu.:    0.00  1st Qu.:    0.00  1st Qu.:    0.0  1st Qu.:    0.00  1st Qu.:    0.00
 Median :    0.00  Median :    0.00  Median :    0.00  Median :    0.0  Median :    0.00  Median :    0.00
 Mean   :   21.51  Mean   :   22.33  Mean   :   16.96  Mean   :   20.6  Mean   :   20.83  Mean   :   20.58
 3rd Qu.:    0.00  3rd Qu.:    0.00  3rd Qu.:    0.00  3rd Qu.:    0.0  3rd Qu.:    0.00  3rd Qu.:    0.00
 Max.   :36671.00  Max.   :36065.00  Max.   :30852.00  Max.   :19477.0  Max.   :29761.00  Max.   :38180.00
```



histogram of values

log-transformation



histogram of log(values+1)

different gene populations