Univariate visualization - interval and ratio variables

A bit of statistics

Categorical variables

Cardinal (interval and ratio) variables

Counts (frequencies)

- Relative (percentage)
- Absolute (number as such)

Too many different values

- Problems with showing just counts
- distribution

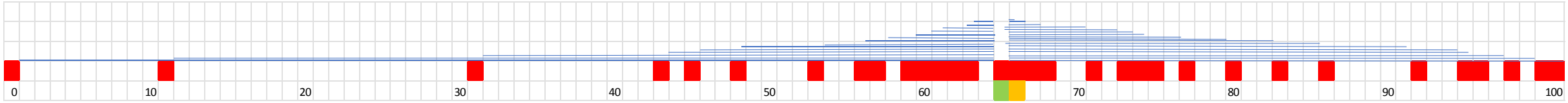
What can be good quantity to show?

- Central value : Average, Median
- Variation: standard deviation
- Other descriptive statistics: minimum, maximum

Average

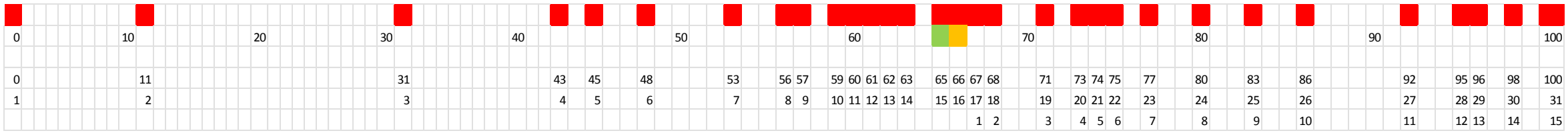
- The „geographical centre“ of data
- sum of distances to lower values = sum of distances to higher values

- movie rating 0-100 points
- Red dots: users evaluation
- Green dot: average value



Median

- Value standing in the middle
- Half of data are lower, half higher



Relation between mean and median

Can be the same

- symetric distribution
- Normal
- U-shape
- Low variace

Can be different

- Skewed distribution
- E.g. income
- Germany: household wage in 2022: 42,192 € 45,457 €
- Why is it different?

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

X - The Value in the data distribution

μ - The population Mean

N - Total Number of Observations

Standard deviation

- How far are data from average
- Average speed 50 KMpH (30 miles)
  - Because the car went whole time exactly 50 (sd=0)
  - Because car went half of journey 30 and other half 70 (sd=20)
  - Because car spent one hour in traffic jam and half hour went 150 (sd=70)
- higher deviation means higher variance

Other descriptive stats

Minimum, maximum (+ range)

World records, temperatures, ….

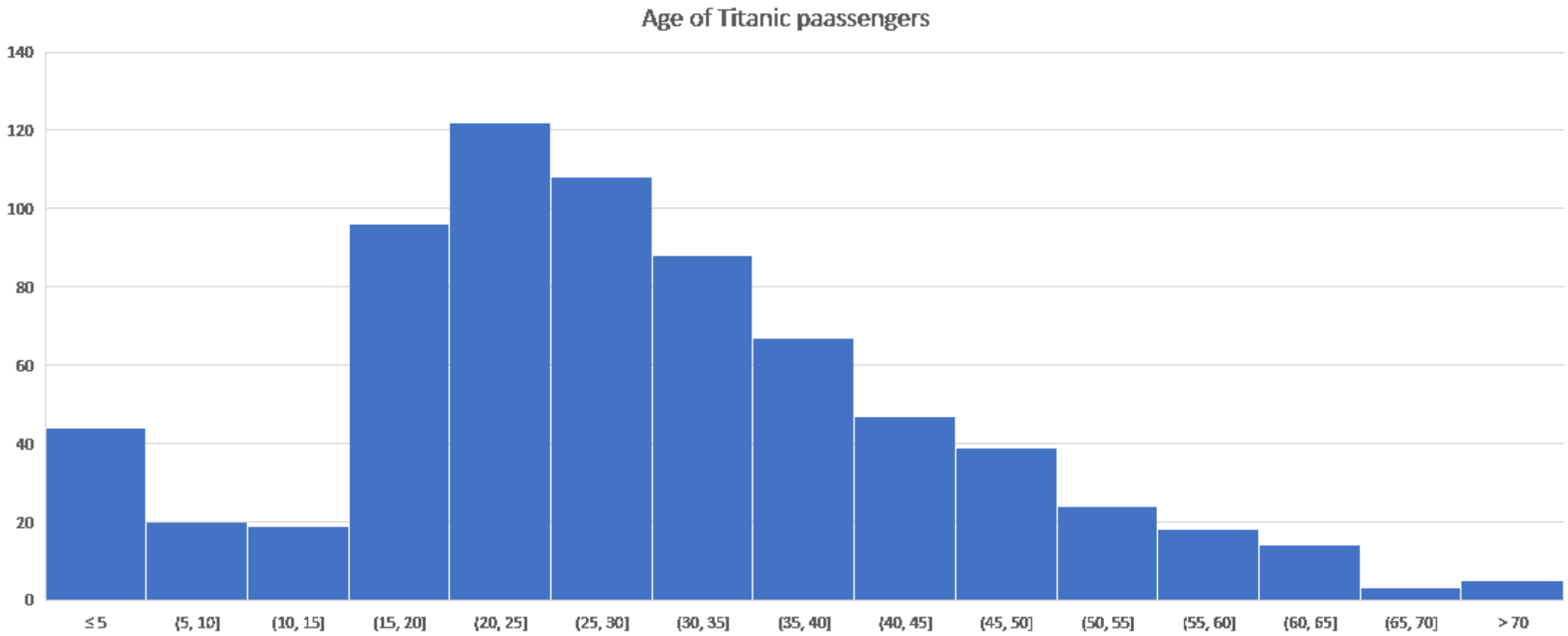
Quartiles

Way how to group values

Gini and variation coef.

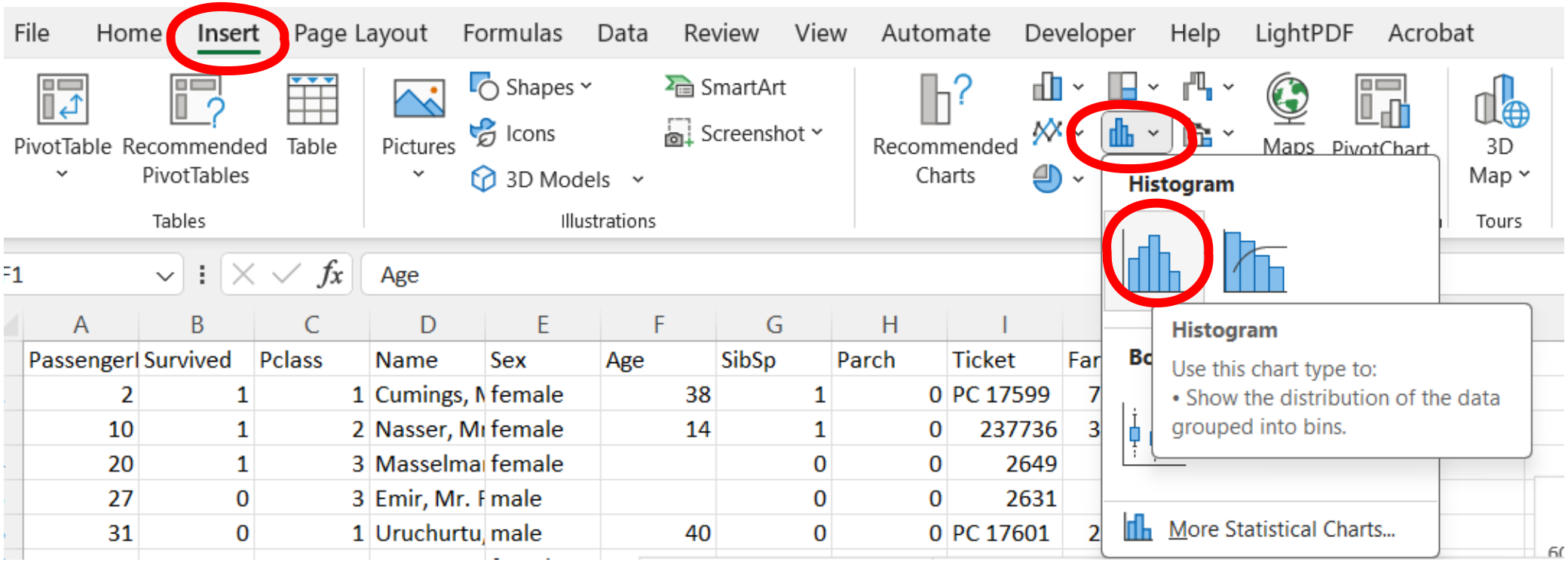
Measures of concentration

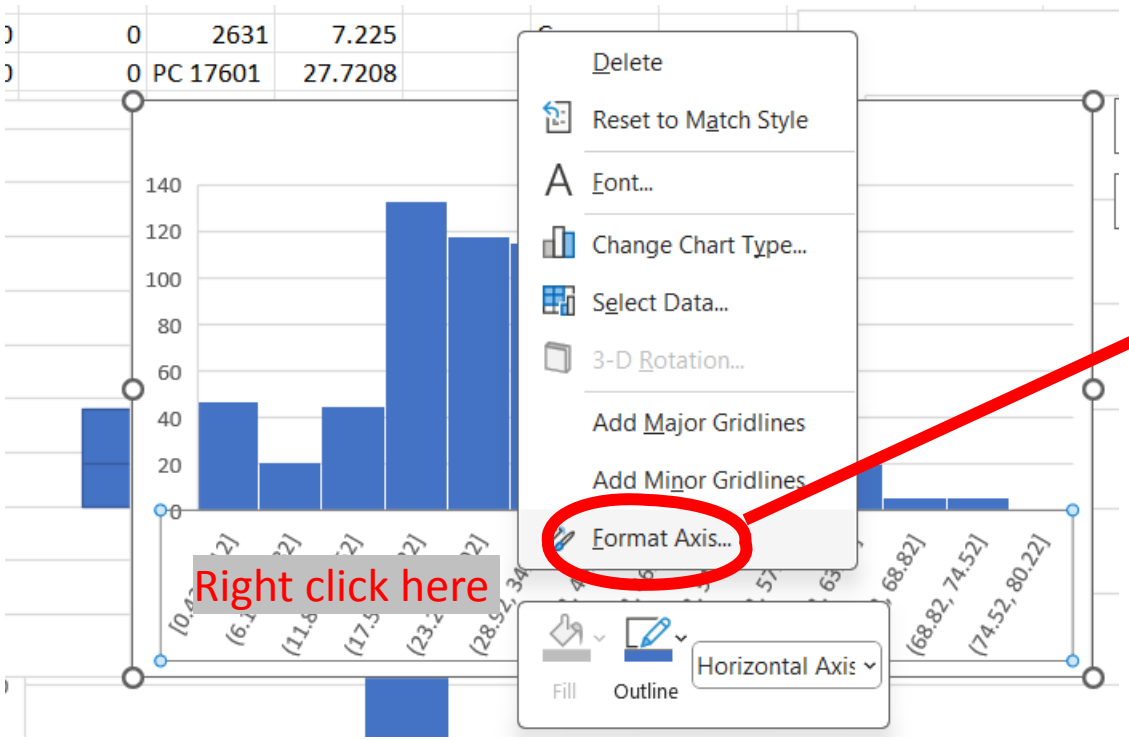
Histogram

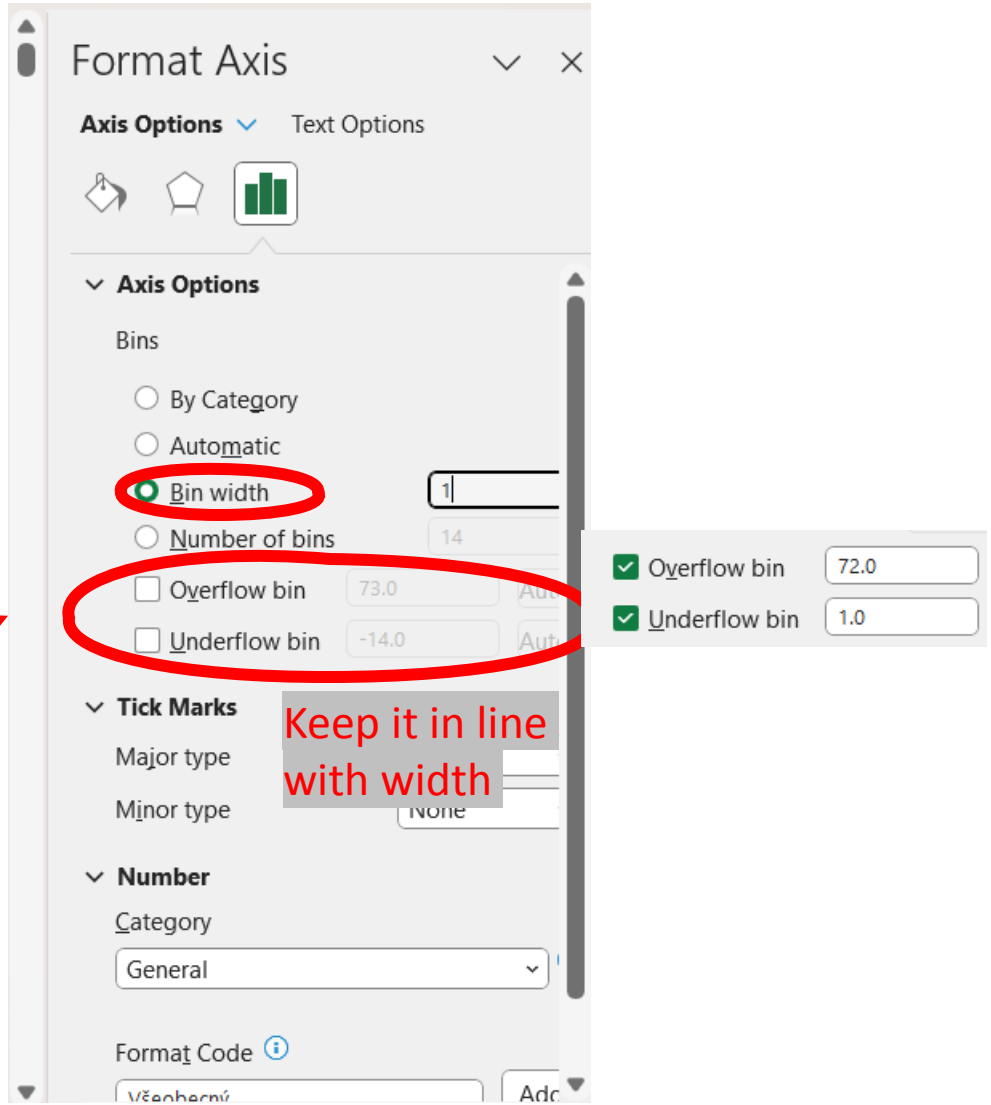

Age of Titanic paassengers
140
120
100
80
60
40
20
0
≤ 5
(5, 10]
(10, 15]
(15, 20]
(20, 25]
(25, 30]
(30, 35]
(35, 40]
(40, 45]
(45, 50]
(50, 55]
(55, 60]
(60, 65]
(65, 70]
> 70

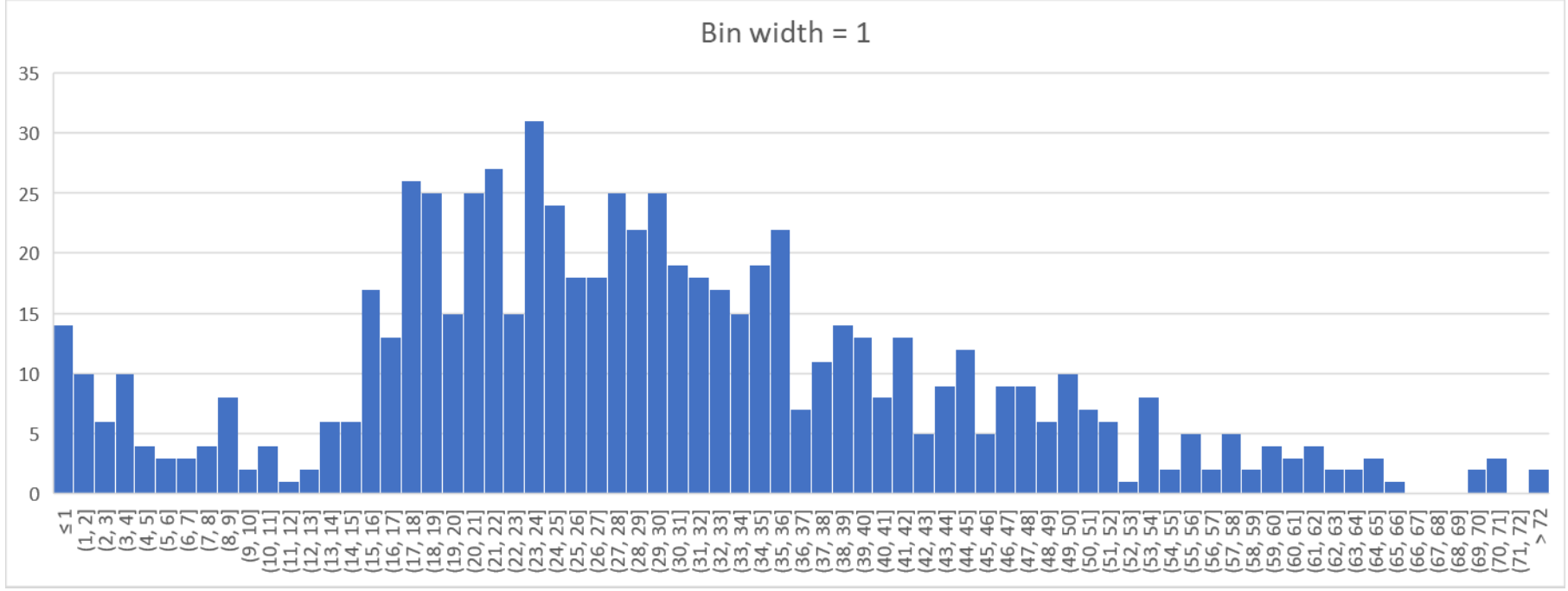
How to make histogram

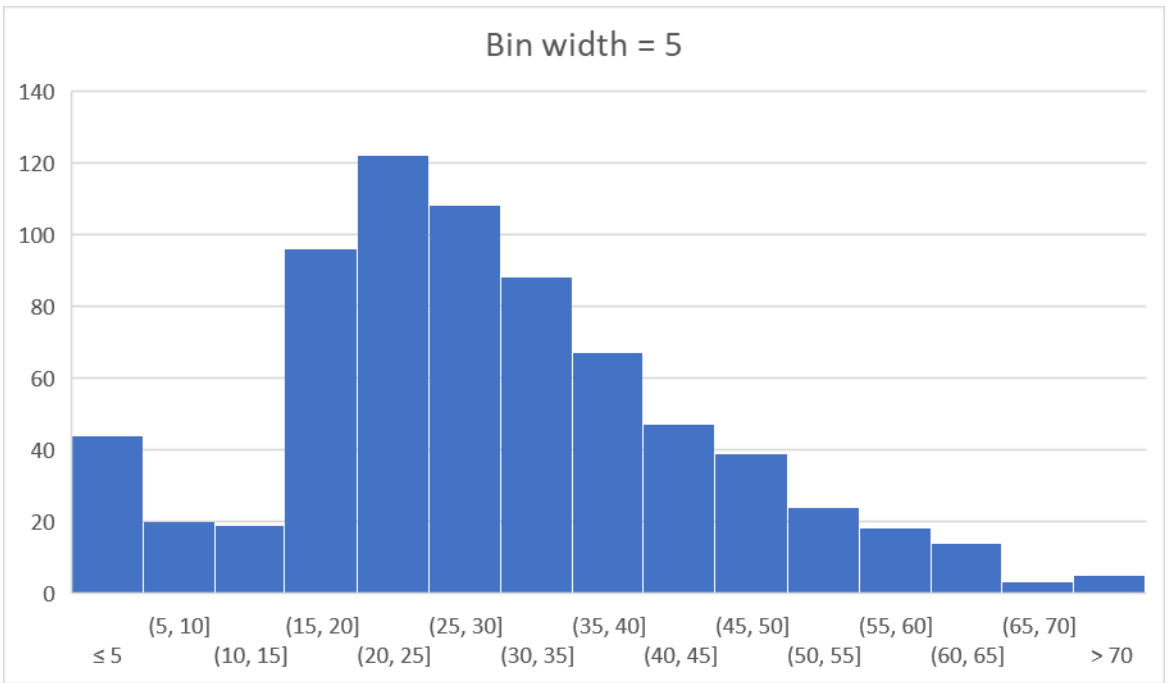
- No need for any computation!
- Select column and click on histogram

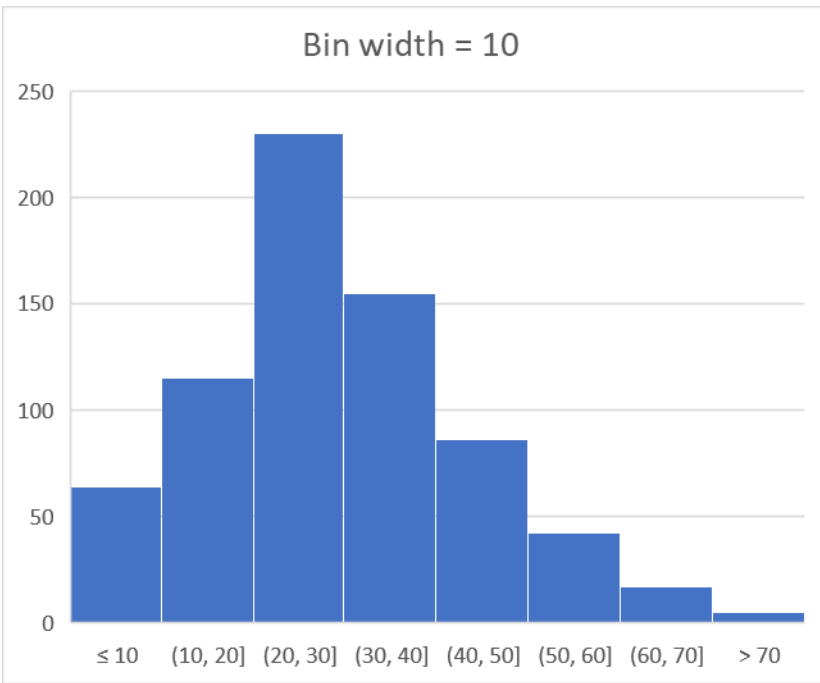


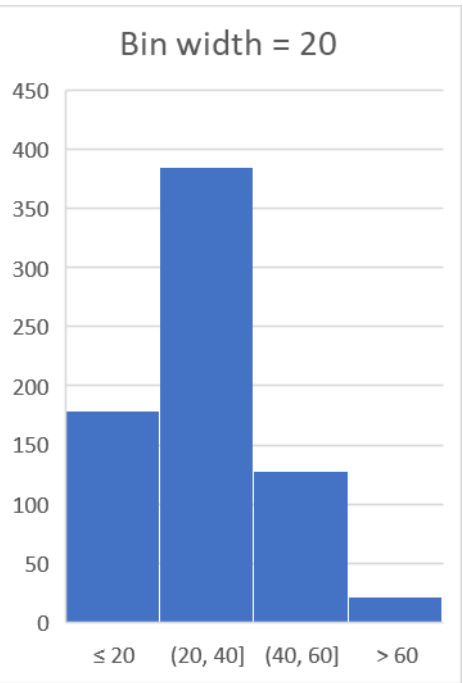

0 2631 7.225
0 PC 17601 27.7208
140
120
100
80
60
40
20
0
Delete
Reset to Match Style
Font...
Change Chart Type...
Select Data...
3-D Rotation...
Add Major Gridlines
Add Minor Gridlines
Format Axis...
Right click here
Fill
Outline
Horizontal Axis


Format Axis
Axis Options
Text Options
Axis Options
Bins
By Category
Automatic
Bin width
1
Number of bins
14
Overflow bin
73.0
Underflow bin
-14.0
Overflow bin
72.0
Underflow bin
1.0
Tick Marks
Major type
Minor type
None
Keep it in line with width
Number
Category
General
Format Code


Bin width = 1
35
30
25
20
15
10
5
0
≤ 1
(1, 2]
(2, 3]
(3, 4]
(4, 5]
(5, 6]
(6, 7]
(7, 8]
(8, 9]
(9, 10]
(10, 11]
(11, 12]
(12, 13]
(13, 14]
(14, 15]
(15, 16]
(16, 17]
(17, 18]
(18, 19]
(19, 20]
(20, 21]
(21, 22]
(22, 23]
(23, 24]
(24, 25]
(25, 26]
(26, 27]
(27, 28]
(28, 29]
(29, 30]
(30, 31]
(31, 32]
(32, 33]
(33, 34]
(34, 35]
(35, 36]
(36, 37]
(37, 38]
(38, 39]
(39, 40]
(40, 41]
(41, 42]
(42, 43]
(43, 44]
(44, 45]
(45, 46]
(46, 47]
(47, 48]
(48, 49]
(49, 50]
(50, 51]
(51, 52]
(52, 53]
(53, 54]
(54, 55]
(55, 56]
(56, 57]
(57, 58]
(58, 59]
(59, 60]
(60, 61]
(61, 62]
(62, 63]
(63, 64]
(64, 65]
(65, 66]
(66, 67]
(67, 68]
(68, 69]
(69, 70]
(70, 71]
(71, 72]
> 72

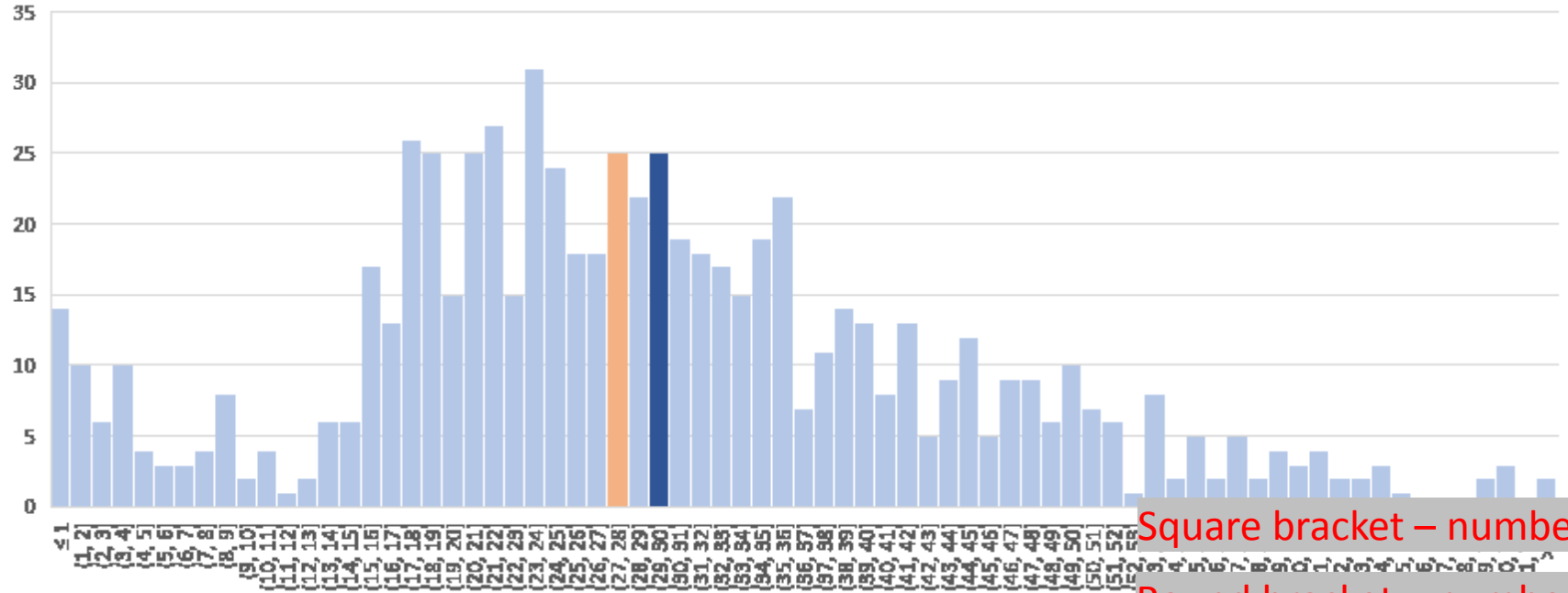

Bin width = 5
140
120
100
80
60
40
20
0
≤ 5
(5, 10]
(10, 15]
(15, 20]
(20, 25]
(25, 30]
(30, 35]
(35, 40]
(40, 45]
(45, 50]
(50, 55]
(55, 60]
(60, 65]
(65, 70]
> 70


Bin width = 10
250
200
150
100
50
0
≤ 10
(10, 20]
(20, 30]
(30, 40]
(40, 50]
(50, 60]
(60, 70]
> 70


Bin width = 20
450
400
350
300
250
200
150
100
50
0
≤ 20
(20, 40]
(40, 60]
> 60

Highlighting mean

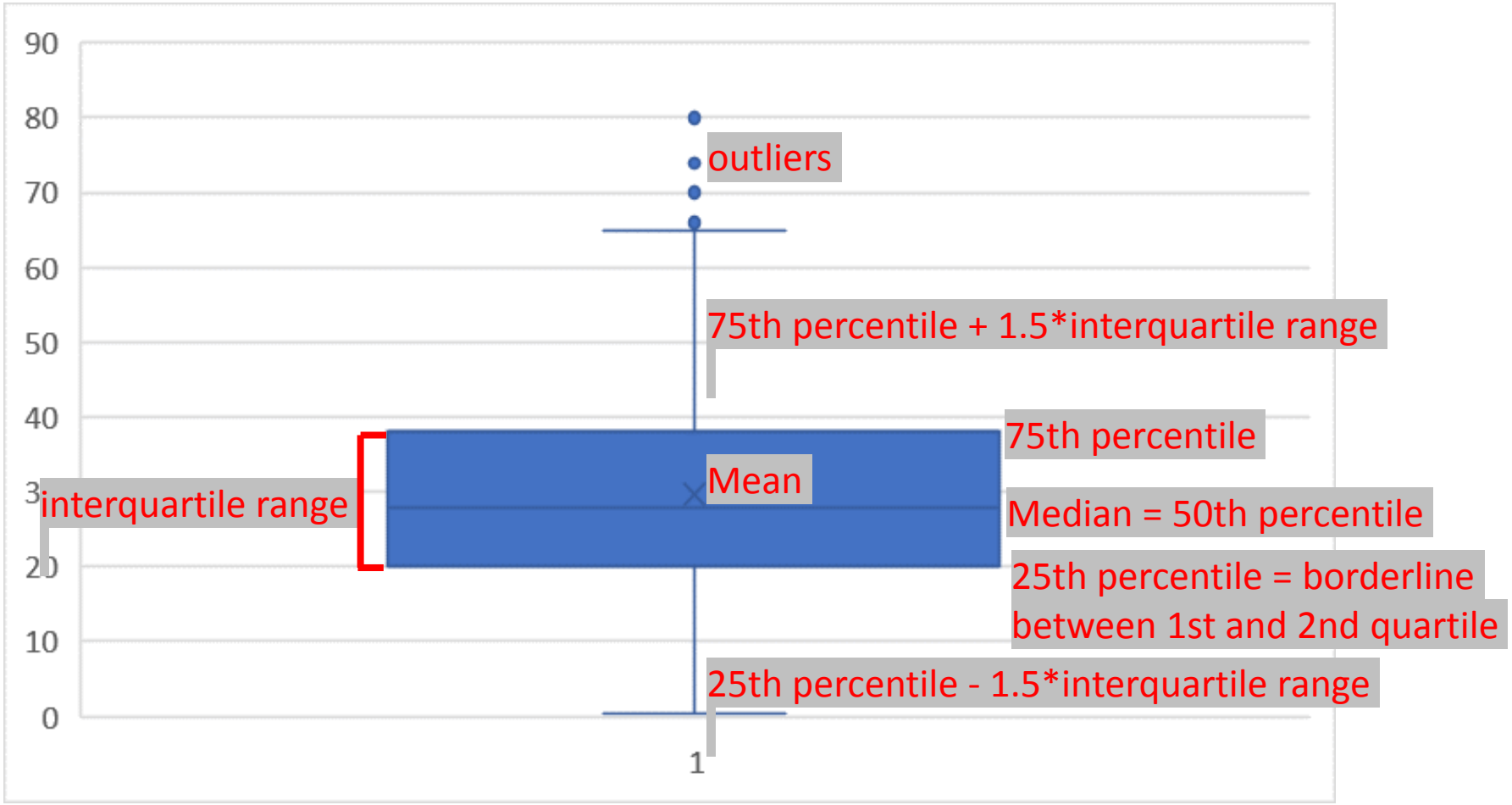
- =AVERAGE(*column*)
- = MEDIAN(*column*)
- Change of color mave to be done manually



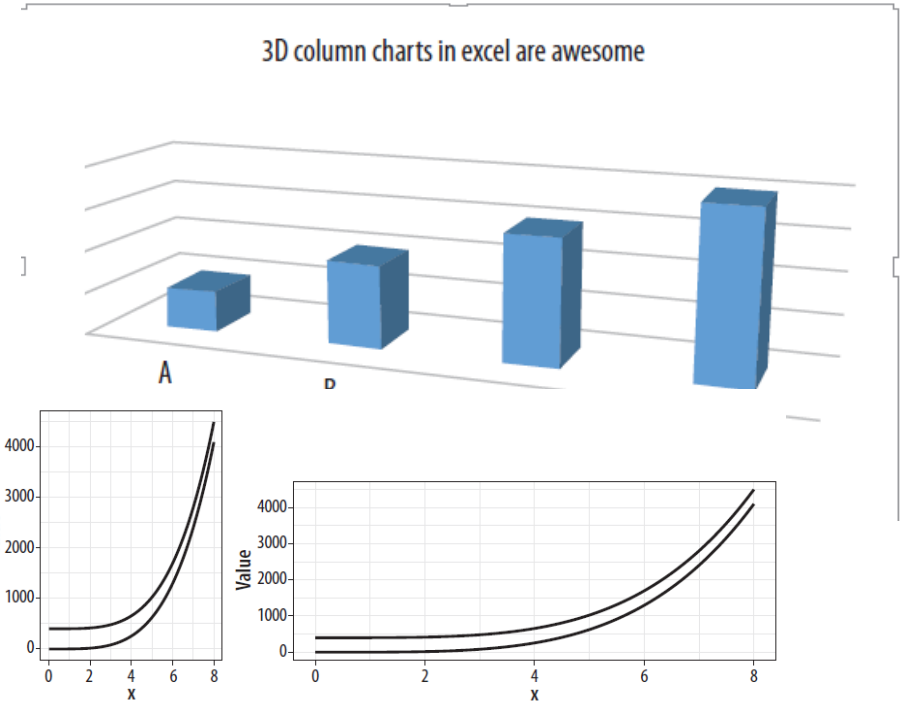
Square bracket – number belongs to interval

Round bracket – number does not belong to interval

Box plot


90
80
70
60
50
40
30
20
10
0
1
outliers
75th percentile + 1.5*interquartile range
75th percentile
Mean
interquartile range
Median = 50th percentile
25th percentile = borderline between 1st and 2nd quartile
25th percentile - 1.5*interquartile range

- Avoid the usage of 3D versions
  - They can be very misleading



- Be carefull with ratio aspect

- Always make titles, subtitles and labels as parsimonous as possible
  - (parsimony means to be maximally simplistics and maximally informative simultaneously)