Summary Statistics

Numerical

- Centrality measure (mean, median)
- Dispersion measure (range, percentiles, variance , standard deviation)

Categorical

- Total count
- Unique count
- Category Counts and proportions
- Per category statistics

Centrality Measure

One number to represent entire set of values

Number central to the data

Central tendency

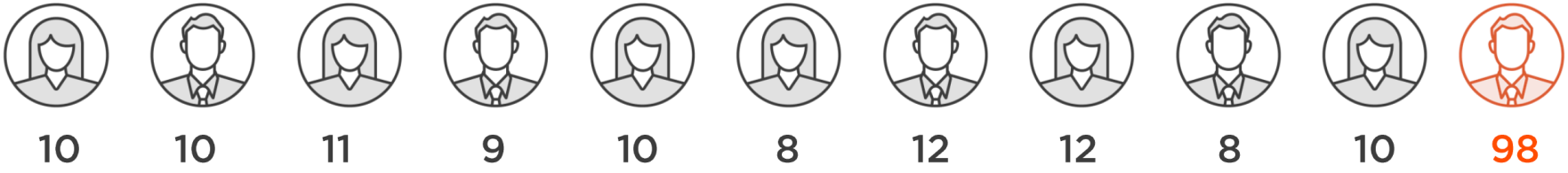
Mean / Average

Average behavior

Centrality Measure : Mean or Average

Mean age : sum of ages / count = 100 / 10 = 10

Problem : Affected by extreme values

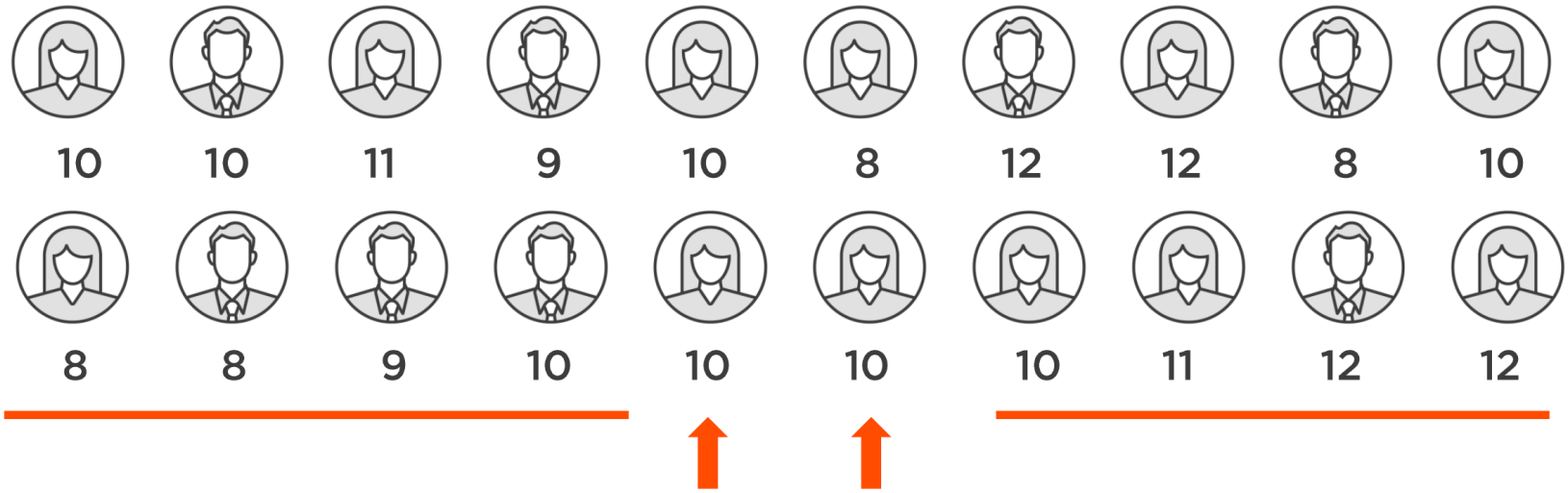


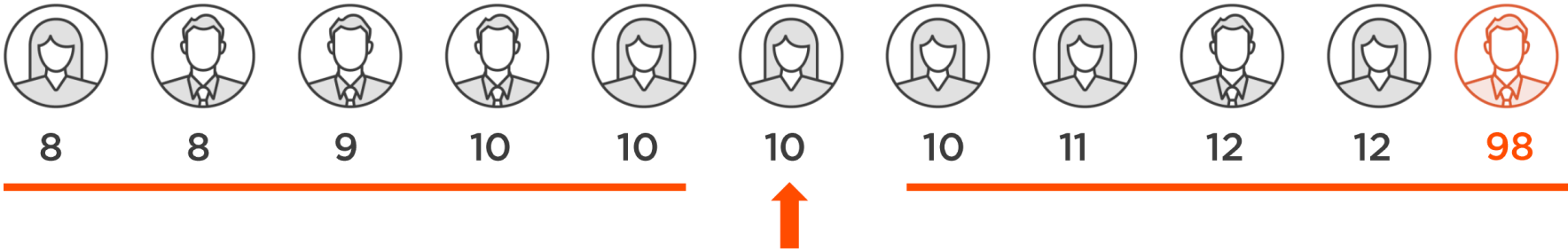
Mean age : sum of ages / count = 198 / 11 = 18

Median

Middle value in the sorted list

Centrality Measure : Median





Spread / Dispersion Measure

How spread out values are from central value

Variability

Range

Difference between maximum and minimum

Spread : Range

Age range : max – min = 12 – 8 = 4

Problem : Affected by extreme values

Age range : max – min = 98 – 8 = 90

Percentiles

x percentile is y means x% of values are below y

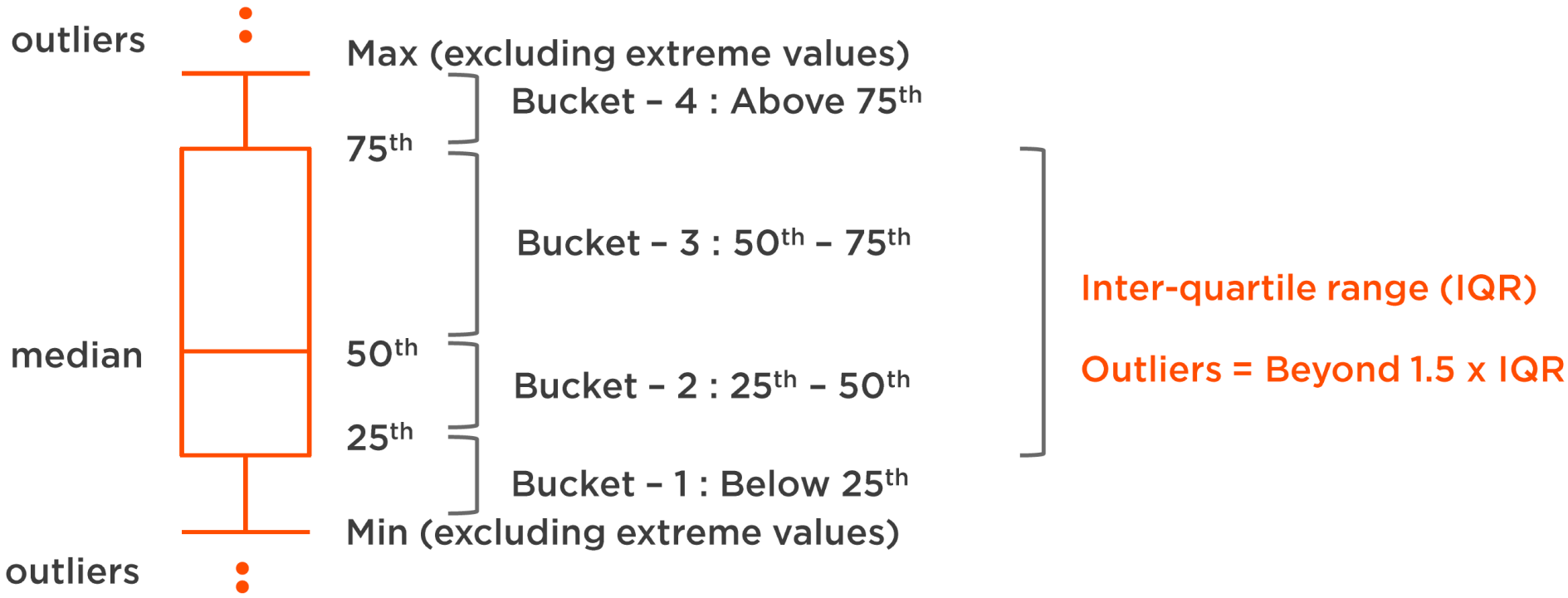
50 percentile is 10 means 50% of values are below 10

25^{th}, 50^{th} , 75^{th}

- Bucket – 1 : Below 25^{th}
- Bucket – 2 : 25^{th} – 50^{th}
- Bucket – 3 : 50^{th} - 75^{th}
- Bucket – 4 : above 75^{th}

Quartiles

Box-Whisker Plot



Variance

Measure of variability

How far each value in list from mean value

Small variance = less spread

High variance = large spread

Variance = $\frac{sum((value - mean)^2)}{count}$

Affected by extreme values

Unit is not clear

Standard Deviation

Standard deviation = $\sqrt{variance}$

Unit is same as that of the feature

Low standard deviation = less spread

High standard deviation = large spread

Overview

(Concepts)

Exploratory data analysis

- Distributions
- Grouping
- Crosstabs
- Pivots

Distributions

Univariate

- Histogram
- Kernel Density Estimation (KDE) plot

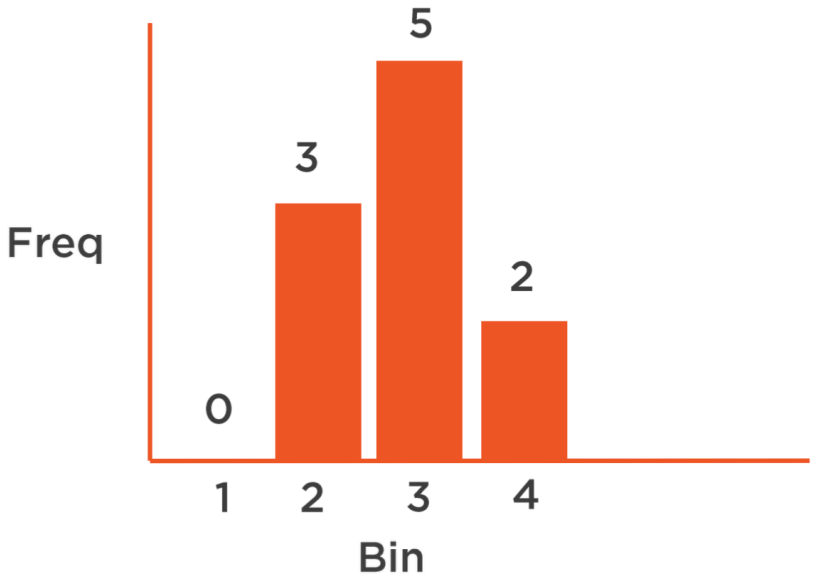
Bivariate

- Scatter plot

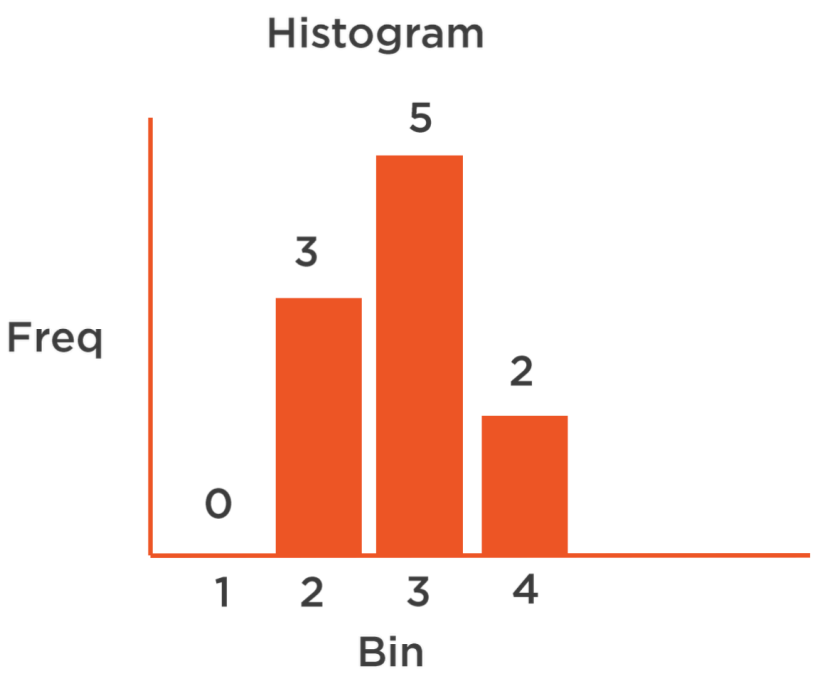
Histogram

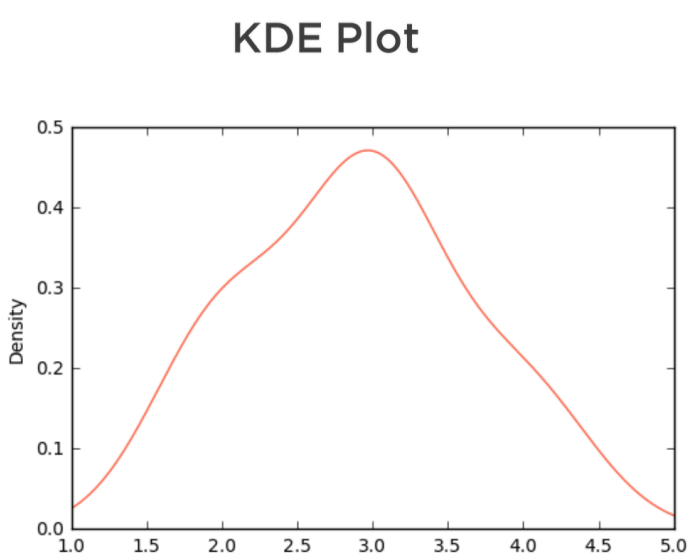
Age	10	10	11	9	10	8	12	12	8	10
Bin	3	3	3	2	3	2	4	4	2	3

Bucket (bin) number	Bucket (bins)	Frequency
1	6-8	0
2	8-10	3
3	10-12	5
4	12-14	2

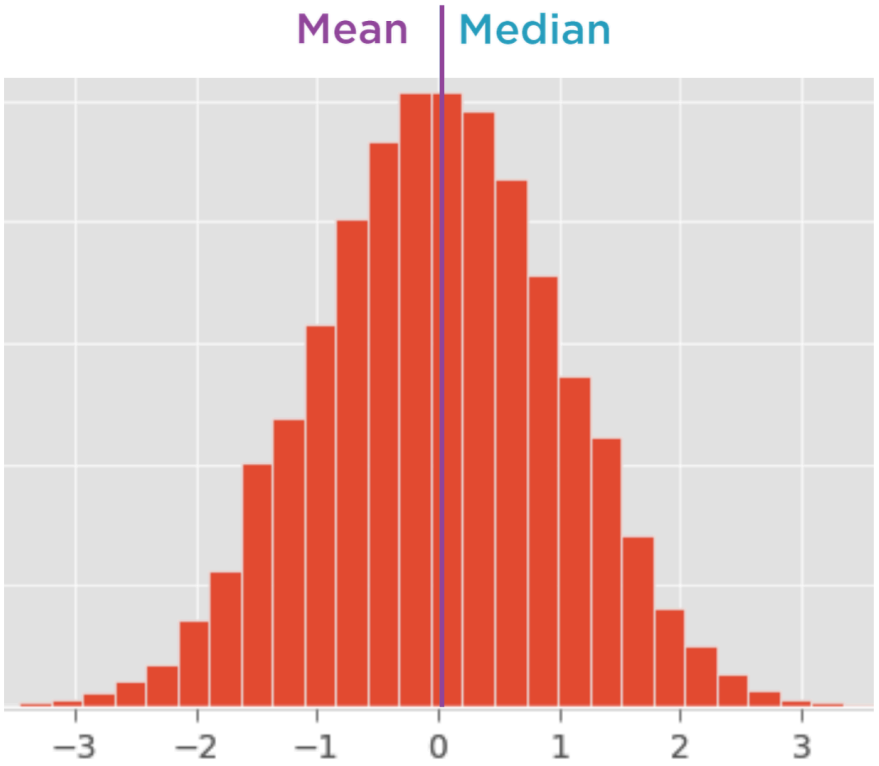


Kernel Density Estimation (KDE) Plot



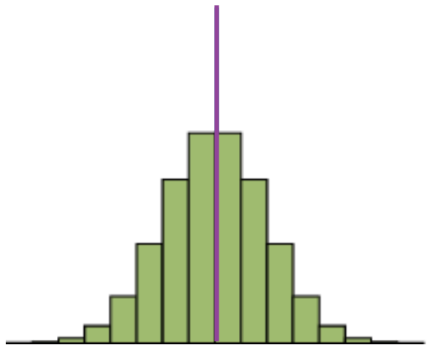


Normal Distribution

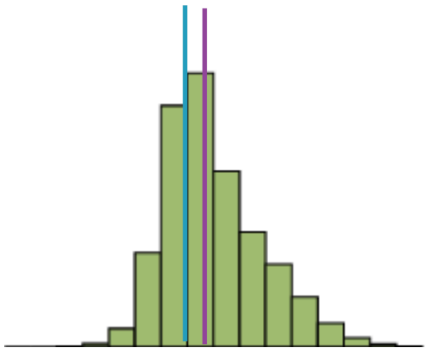


Skewness : zero

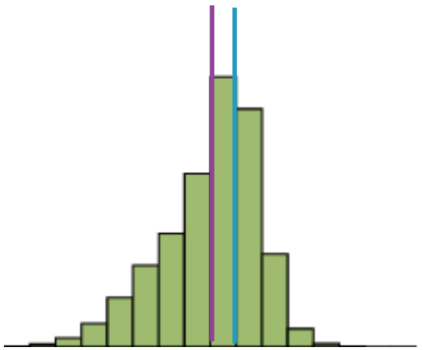
Univariate Distribution : Skewness



Normal distribution



Right (positive) skewed



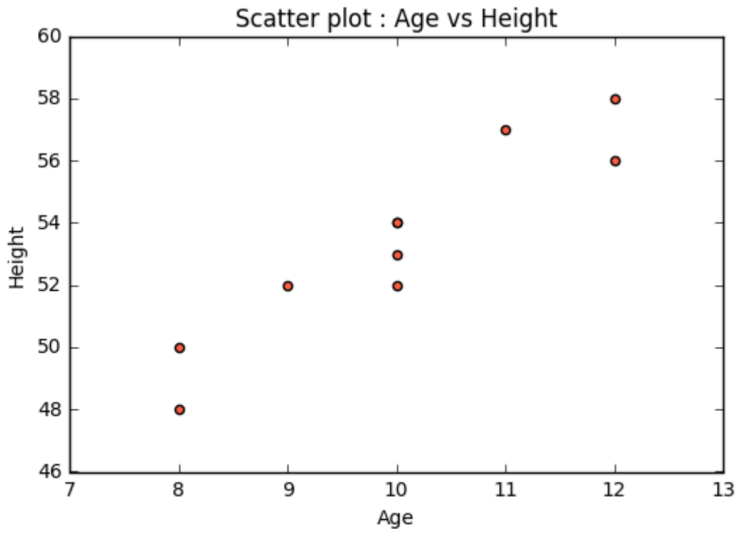
Left (negative) skewed

— Median

— Mean

Scatter Plot

Age	10	10	11	9	10	8	12	12	8	10
Height	54	53	57	52	52	50	58	56	48	54



Grouping

Age	10	10	11	9	10	8	12	12	8	10
Age	10	11	10	8	12	10	10	9	12	8

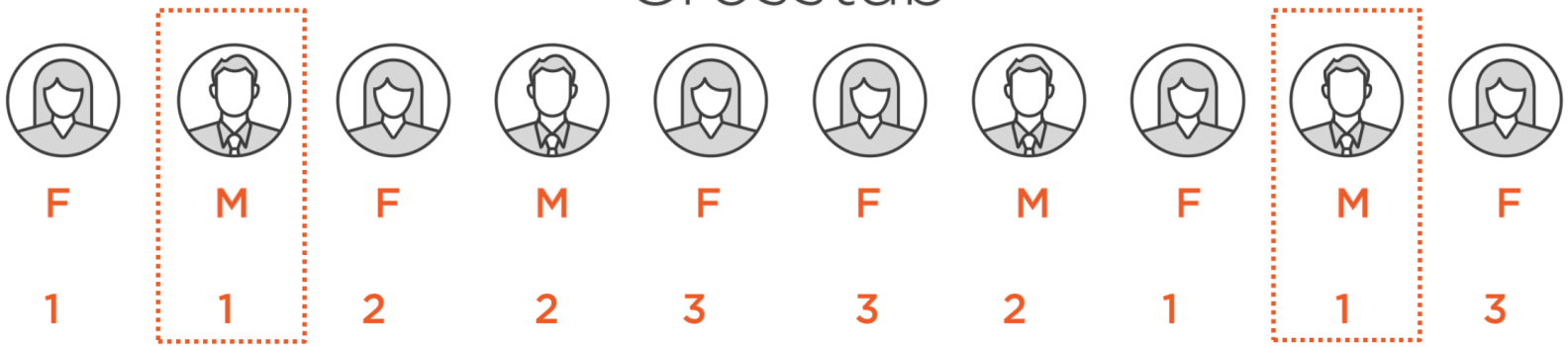
	Group : F	Group : M
Mean age	10.17	9.75
Median age	10	9.5
Count	6	4

Grouping

Age	10	10	11	9	10	8	12	12	8	10
Class	1	1	2	2	3	3	2	1	1	3

Group	Summary
M - 1	...
M - 2	...
M - 3	...
F - 1	...
F - 2	...
F -3	...

Crosstab



Gender	F	M	F	M	F	F	M	F	M	F
Class	1	1	2	2	3	3	2	1	1	3

Class	1	2	3
Gender			
M	2	2	0
F	2	1	3

Pivot Table

Gender	F	M	F	M	F	F	M	F	M	F
Class	1	1	2	2	3	3	2	1	1	3
Age	10	10	11	9	10	8	12	12	8	10

3 Age 4 Mean

2

Class	1	2	3
Gender			
M	9.0	10.5	NaN
F	11.0	11.0	9.33

1