DATA – Introduction to Data Science

Semester 2 | 2025/26

Lecture 4 : Foundations of Descriptive Statistics

Dr Xinwei Fang

Lecture Overview

- Identify and distinguish between different types of data
- Recognise common sources where data can be collected
- Understand the key considerations involved in data collection
- Understand the purpose of descriptive statistics and how they are used to summarise and explain data.

Can John and Jane Afford Melbourne?

User Profile

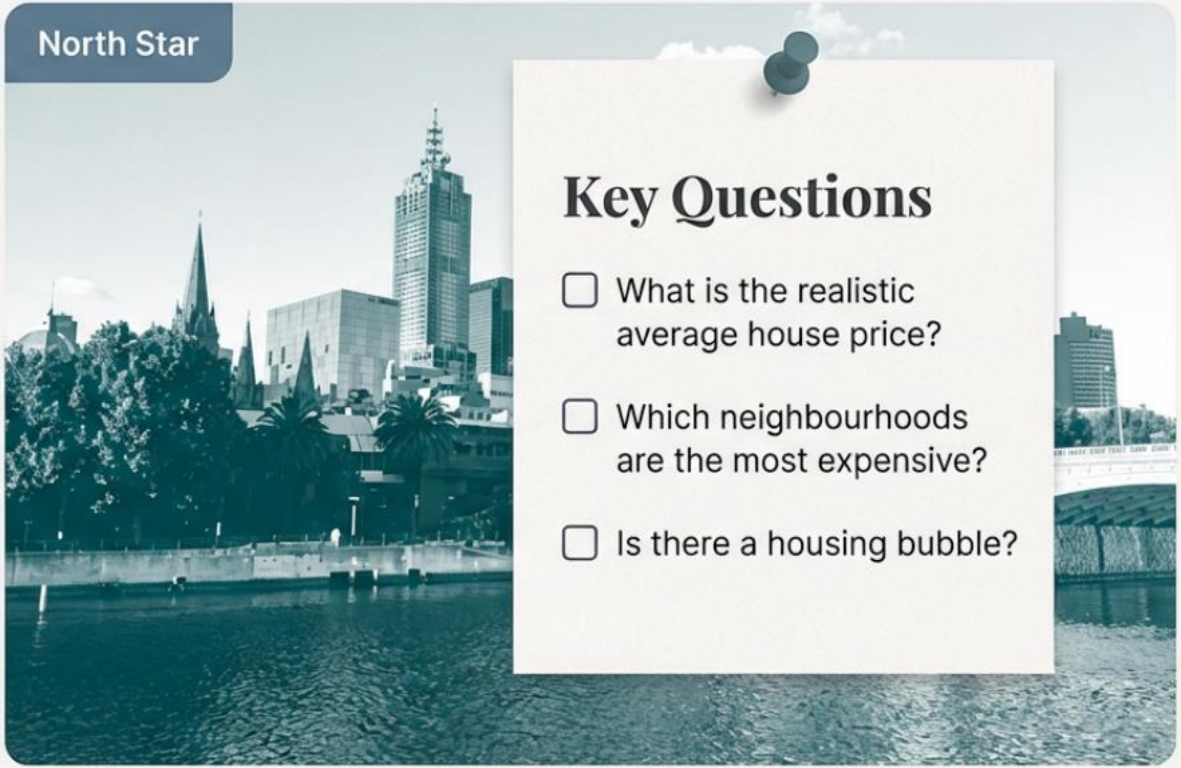
Profiles: John & Jane Doe

Role: Senior Data Scientists

Origin: York, UK

Destination: Melbourne, AU

Objective: Assess Cost of Living.



The Dataset: 63,000+ Entries from Domain.com.au

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Postco.	Regionname
0	Abbotsford	49 Lithgow St	3	h	$1,490,000.0	S	Jellis	1/04/2017	3067	Northern Metropolitan
1	Abbotsford	59A Turner St	3	h	$1,220,000.0	S	Marshall	1/04/2017	3067	Northern Metropolitan
2	Abbotsford	119B Yarra St	3	h	$1,420,000.0	S	Nelson	1/04/2017	3067	Northern Metropolitan
3	Aberfeldie	68 Vida St	3	h	$1,515,000.0	S	Barry	1/04/2017	3040	Western Metropolitan
4	Airport West	92 Clytedale Rd	2	h	$670,000.0	S	Nelson	1/04/2017	3042	Western Metropolitan

https://www.kaggle.com/anthonypino/melbourne-housing-market/data

Data Profile

- **Source:** Domain.com.au (Public Query)
- **Volume:** >63,000 Entries
- **Dimensions:** 13 Variables
- **Mix:** Categorical & Numerical

Taxonomy of Data

Depending on what is being classified

How data is organised and stored

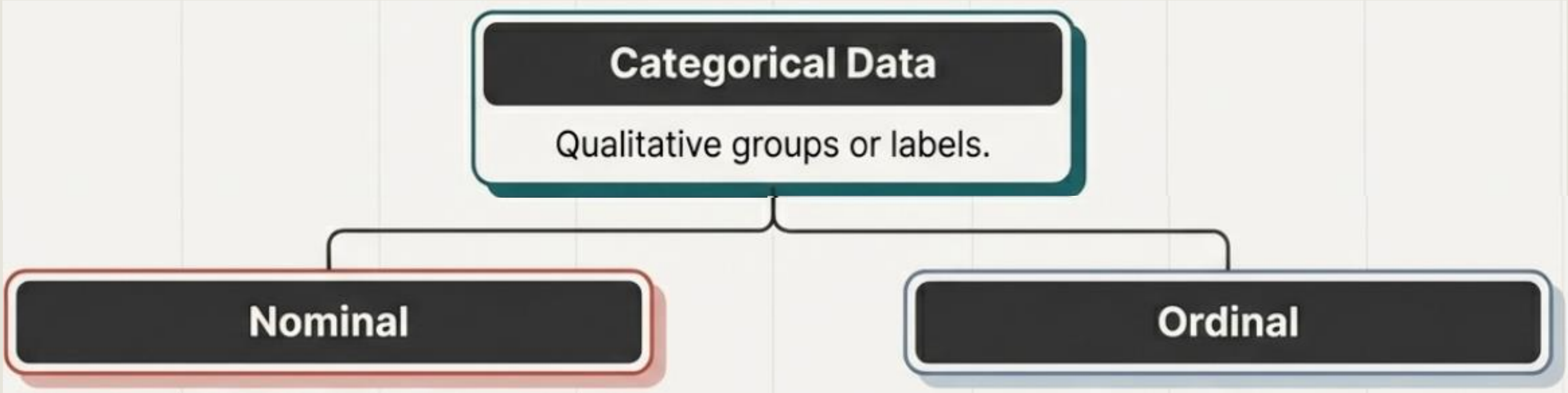
The nature of the data values

All Variables

Categorical

Numerical

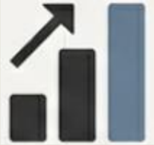
Categorical Data



No implied order

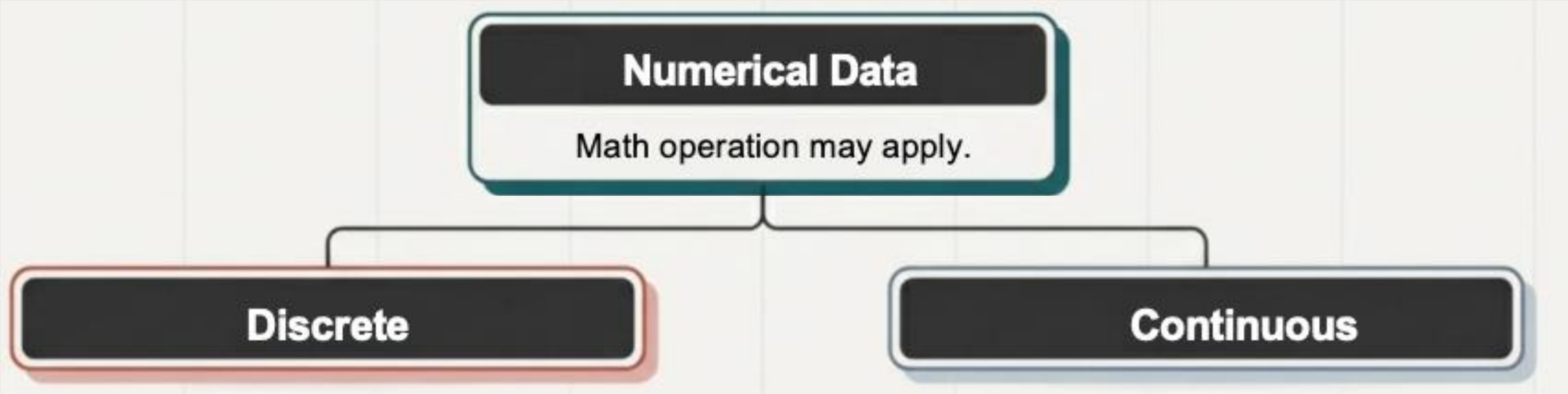
Clear ordering or rank

- Property Types (Flat, House)
- Cities (York, Leeds, London)
- Car Brands (BMW, Toyota)



- Survey Ratings (Low ->High)
- House Condition (Poor -> Excellent)
- Ranking (1st -> 5th)

Numerical Data



consists of countable, distinct values

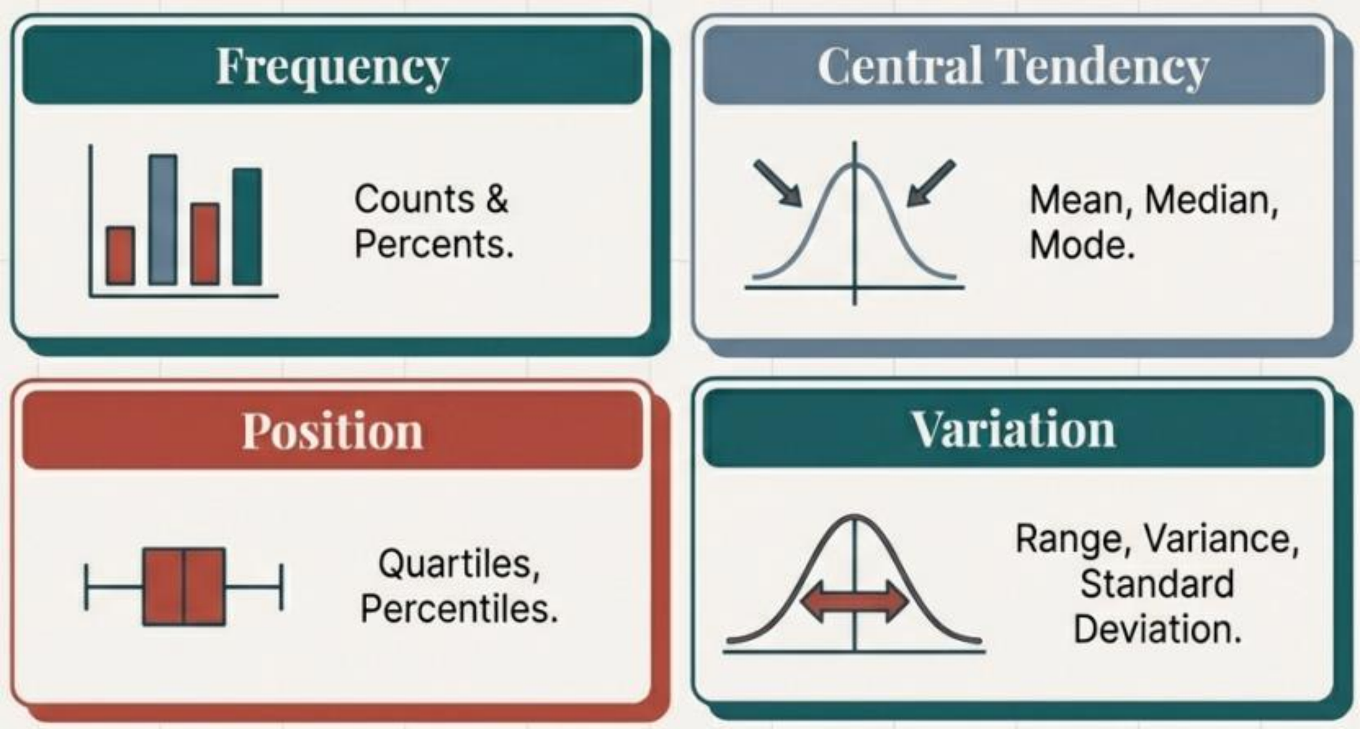
can take any value within a given range.

- Number of students in a class
- Number of cars in a car park
- Number of emails received in a day

- Height of a person
- Time taken to complete a task
- Temperature of a room

The Toolkit: Descriptive Statistics

Understand data without making predictions or inferences.



Frequency

Used with countable data

Count: how many times an event has happened or happened within a given time frame.

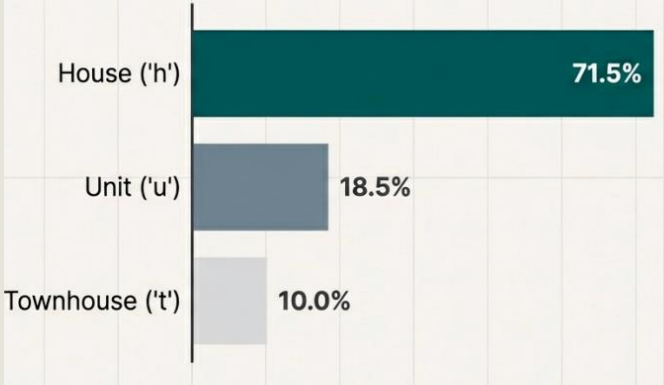
Percentage: The percentage of a particular category over the sample size

```
#Count houses of each type
uniqueT, countsT = np.unique(df['Type'], return_counts=True)
print("Number of houses of different types:")
print(uniqueT, countsT)
```

```
Number of houses of different types:
['h' 't' 'u'] [45053  6315 11655]
```

```
#Percentage of house of each type
print("Percentage of houses of different types:")
(uniqueT, countsT/len(df['Type'])*100)
```

```
Percentage of houses of different types:

(array(['h', 't', 'u'], dtype=object),
 array([71.48660013, 10.02015137, 18.4932485 ]))
```



Insight: The Melbourne market is heavily dominated (71.5%) by houses.

Central Tendency: The Mean Price

Captures the centre of the distribution and is calculated by summing all values and dividing by the count.

$$\overline{x} = \frac{1}{n}\sum_i x_i$$

For example, given the house values (in thousands) [70,60,80,85,92]

(70+60+80+85+92) / 5 = 77.4

If the data sample is drawn from a population, the mean of the sample is an unbiased estimate of the population mean

Population → the whole group
Sample → a subset of the population

Central Tendency: The Mean Price

In Python with NumPy Library

```
mean = np.mean(df['Price'])
```

Result

Townhouse Mean:
$911,147

House Mean:
$1,110,586

Overall Mean:
$997,898

Central Tendency: The Median Price

The middle score in an ordered data set which splits the data at the 50th percentile, and can be obtained by

$$\text{Median location} = (N+1)/2$$

Odd number of observations: [45, 30, 87, 67, 94, 102, 124], N=7

Median location = (7+1)/2 =4

1st	2nd	3rd	4th	5th	6th	7th
[30,	45,	67,	87,	94,	102,	124]

median

Even number of observations: [45, 30, 87, 67, 94, 124, 155, 102], N=8

Median location = (8+1)/2 =4.5

1st	2nd	3rd	4th	5th	6th	7th	8th
[30,	45,	67,	87,	94,	102,	124,	155]

median is here

Median = (87+94)/2 = 90.5 (the midpoint)

median

Central Tendency: The Median Price

In Python with NumPy Library

```
median = df['Price'].median()
```

Townhouse Median:
$830,000

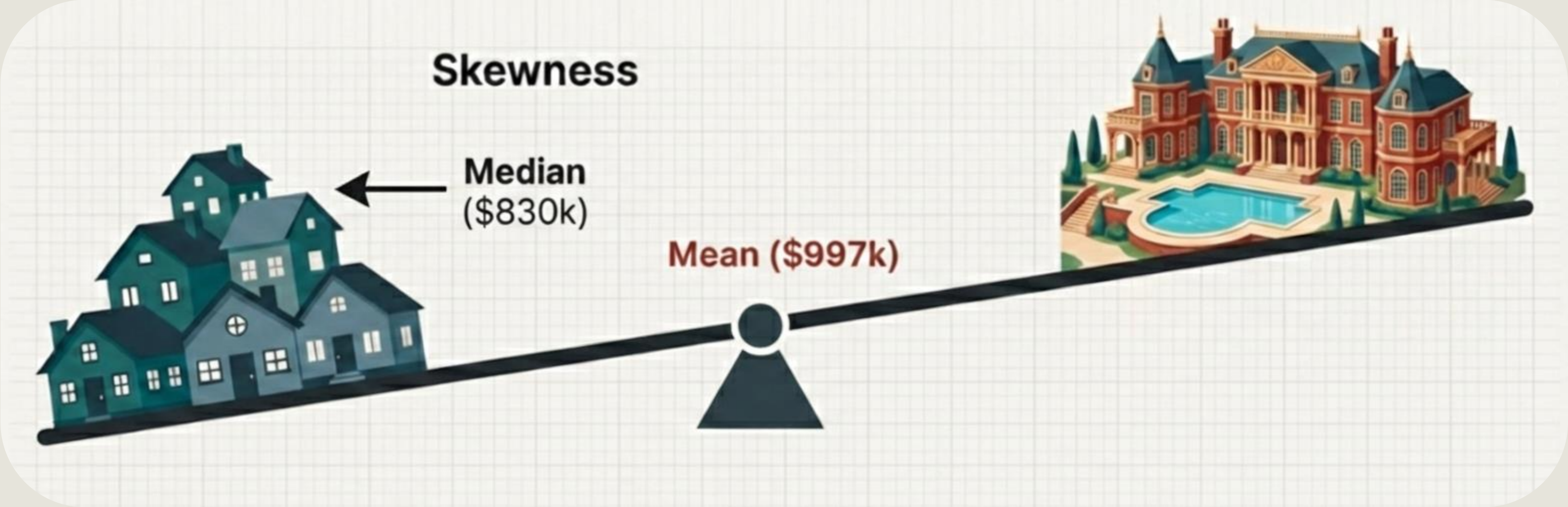
House Median:
$935,000

Overall Mean:
$997,898

Overall Median:
$830,000

$167K Price gap

The Conflict: Why Mean and Median Disagree

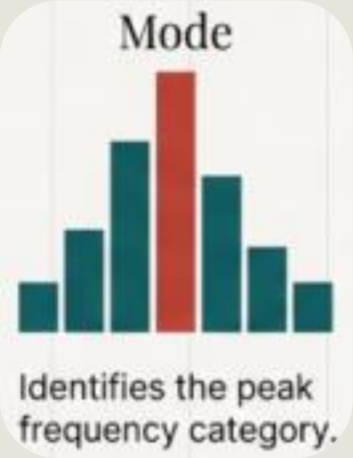


High-value luxury properties pull the mathematical average (Mean) upwards, distorting reality for the typical buyer. The Median is the 'honest' metric in this case.

[60, 70, 80, 85, 92, 500] → Mean increases, Median is stable.

Central Tendency: The Mode

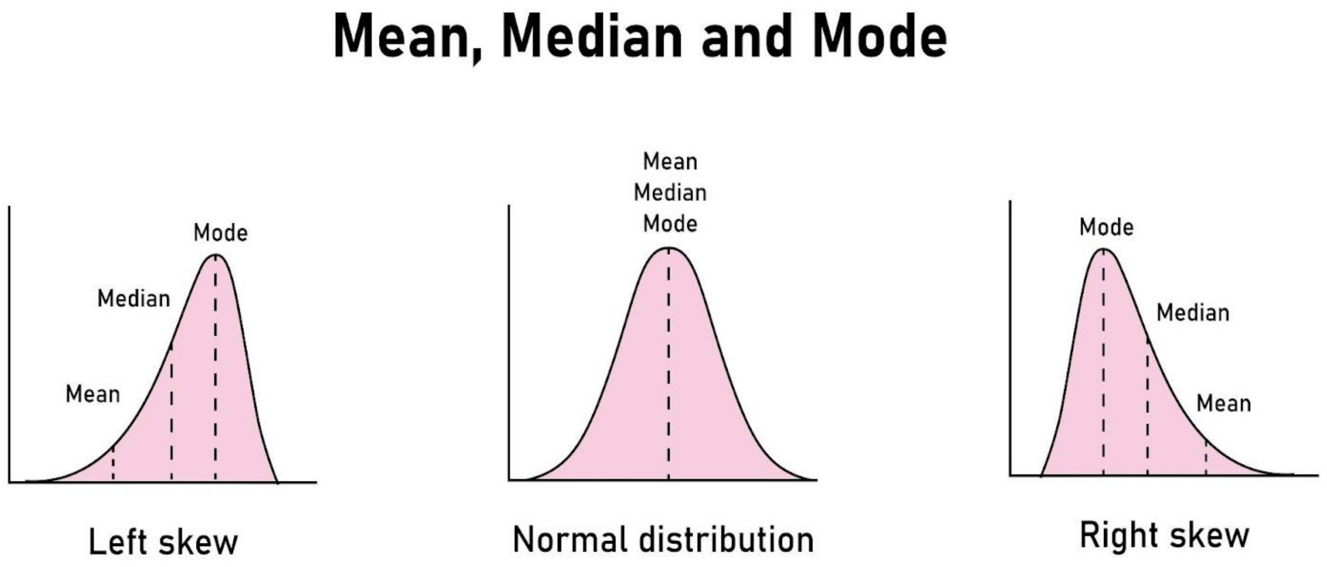
The most frequently occurring value

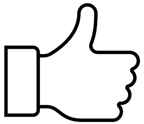



```python
#Mode - most commonly occurring house type

#Find the unique types and cardinality (how many per type)
uniqueT, countsT = np.unique(df['Type'], return_counts=True)
#Find the index of the type with the largest number
modeIndex = np.argmax(countsT)
#Find the house type
modeType  = uniqueT[modeIndex]
#Find its size
modeCount = countsT[modeIndex]

print("Most common type is ", modeType,
      " with ", modeCount, " houses")
```

```
Most common type is  h  with  45053  houses
```

Central Tendency



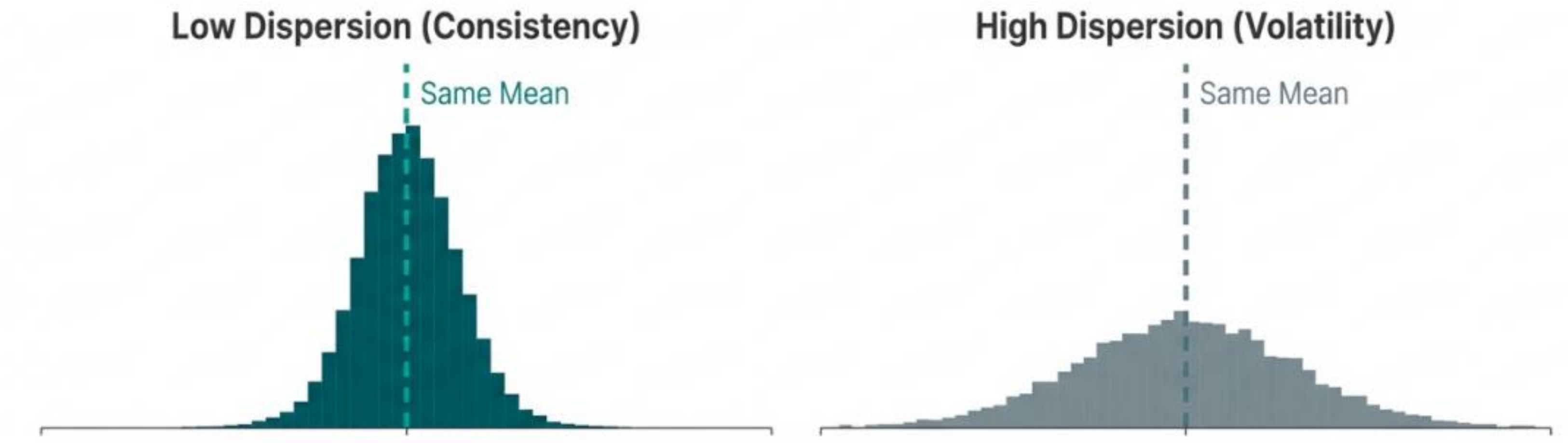


Mean: best for symmetric distributions without outliers
Median: useful for skewed distributions or data with outliers
Mode: useful for categorical (ordinal/nominal) and discrete data

Measures of Variability and Position

In data analysis, knowing the 'typical' value (the Mean) is only half the story.

Two datasets can share an average but tell completely different stories.



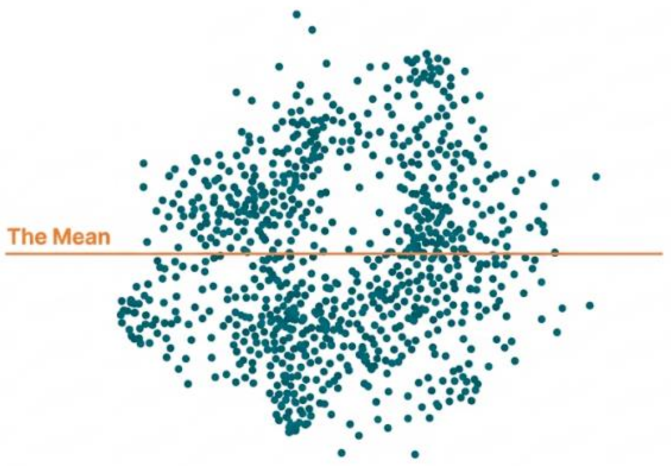
Low Dispersion: Data points are huddled close to the mean.
High Dispersion: Data points drift far from the mean.

Measures of Variability and Position

To truly understand the data, we must answer two deeper questions: 'How much does the data vary?' and 'Where does a specific data point sit relative to the others?'

Let's explore the statistical tools used to answer these questions:

- **Range,**
- **Variance,**
- **Standard Deviation,**
- **Percentiles.**



Range: The Quickest Estimate of Spread

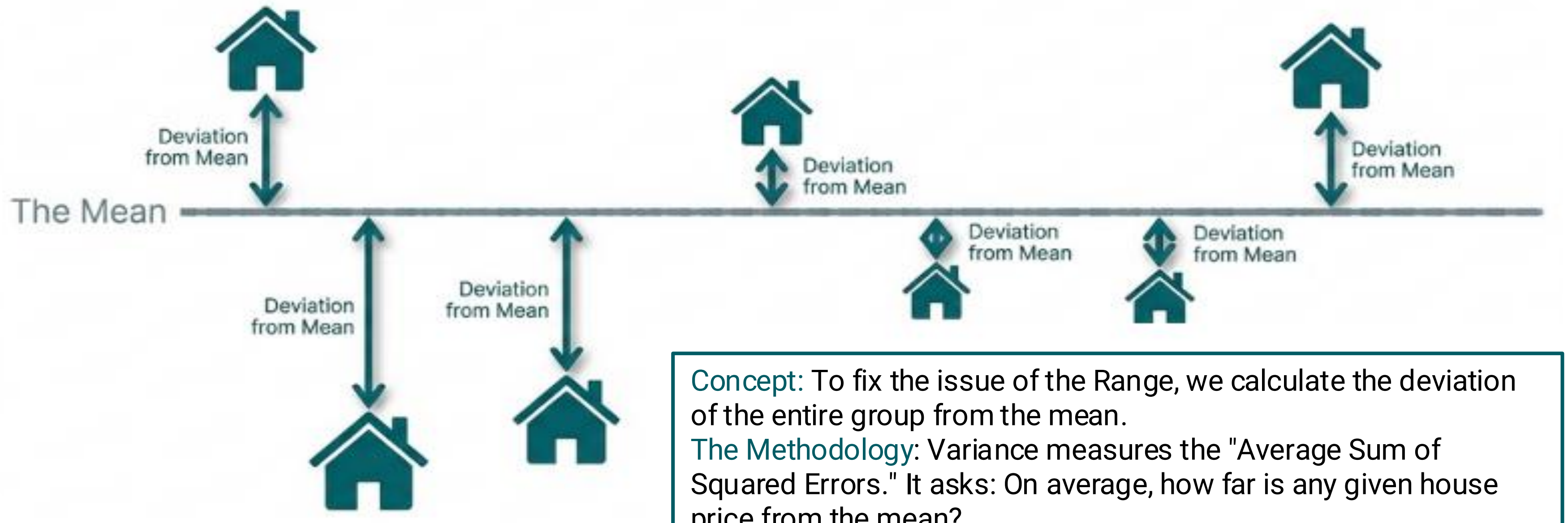
Definition: The distance from the lowest to the highest value

Calculation: Max – Min = Range

The Limitation: The range is not a reliable measure because it is extremely sensitive to extreme values. It is determined only by the smallest and largest values in the data.

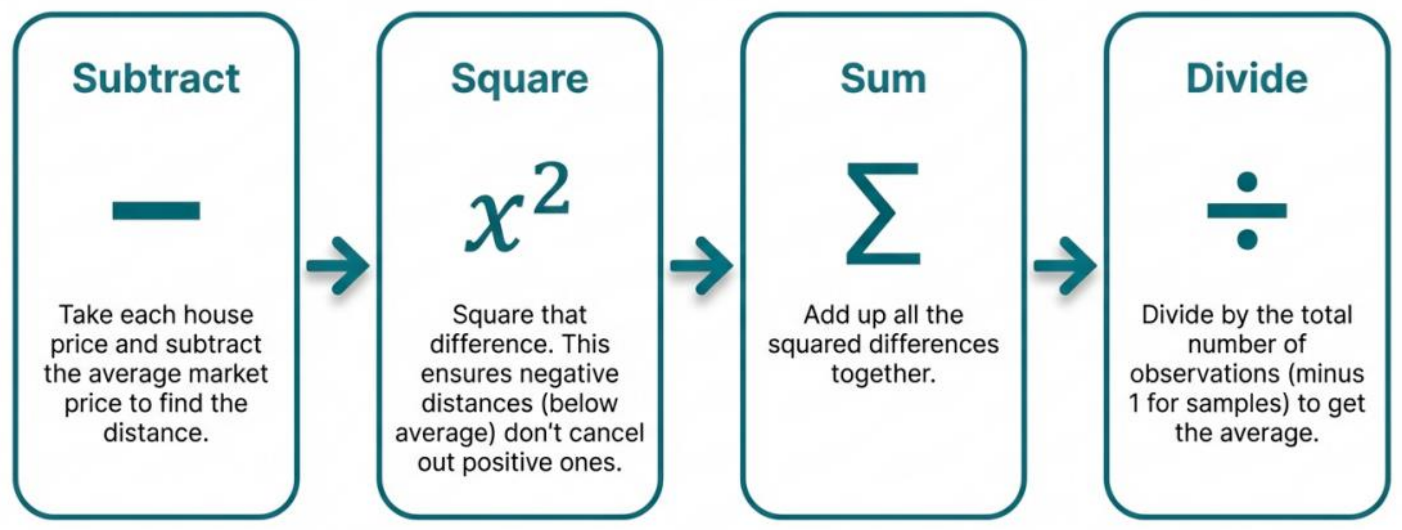
Variance: Accounting for Every Data Point

A robust metric that measures the distance of every observation from the mean.



Concept: To fix the issue of the Range, we calculate the deviation of the entire group from the mean.
The Methodology: Variance measures the "Average Sum of Squared Errors." It asks: On average, how far is any given house price from the mean?
Key Shift: Unlike Range, which uses only 2 numbers (min/max), Variance uses all N numbers in the dataset.

Calculating of Sample Variance

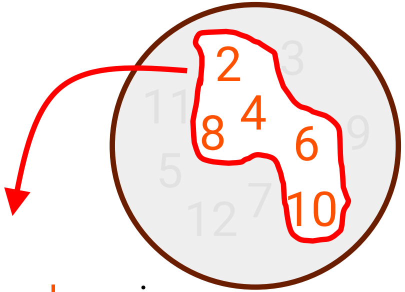


$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} \qquad \sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

The "N-1" Correction

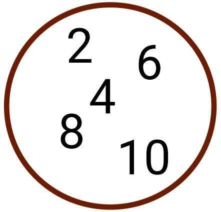
Think Like a Data Scientist

Sample vs. Population. In 99% of data analysis, we work with a SAMPLE (a portion of data), not the POPULATION (all data in existence).



Sample variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad =10$$



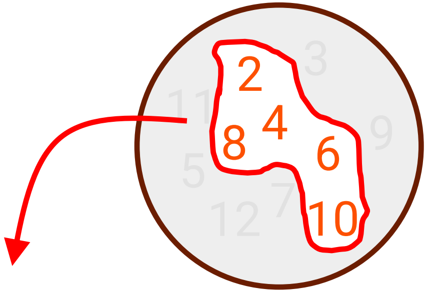
population variance:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad = 8$$

The “N-1” Correction

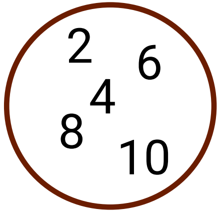
Think Like a Data Scientist

Sample vs. Population. In 99% of data analysis, we work with a **SAMPLE** (a portion of data), not the **POPULATION** (all data in existence).



Sample variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad =10$$



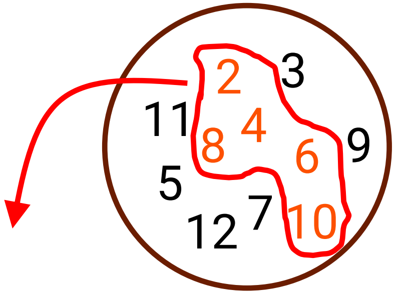
population variance:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad = 8$$

The "N-1" Correction

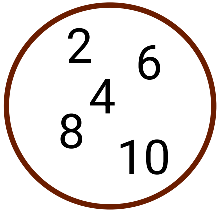
Think Like a Data Scientist

Sample vs. Population. In 99% of data analysis, we work with a **SAMPLE** (a portion of data), not the **POPULATION** (all data in existence).



Sample variance:

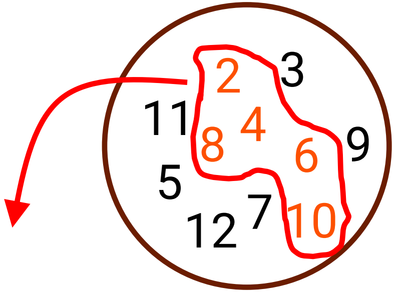
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad =10$$



population variance:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad = 8$$

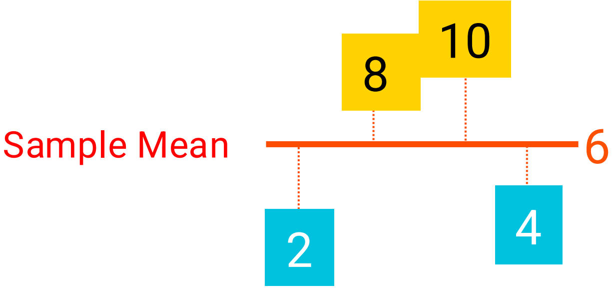
Why we divide by N-1 for samples.



Sample variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad =10$$

Once the mean is calculated, the deviations from that mean must satisfy an important constraint **they always add up to zero**. This is not true when measuring spread around the true population mean.



6 – 10 = - 4

6 – 8 = - 2

6 – 4 = 2

6 – 2 = 4

This restriction **limits how freely the data can vary around the mean**; and **the distances from the mean appear smaller than they really are**.

The Problem with Variance

Mathematically robust, but semantically confusing.

3522336941 51.18

$£^2$ **(Pounds Squared)**

The issue: What does 'Pounds Squared' mean? Because we squared the differences in Step 2, the unit has changed. It is no longer money; it is 'money squared'. This disrupts our intuitive understanding of the market. We need to convert it back.

Standard Deviation : Returning to Reality

The process that restores the original units.

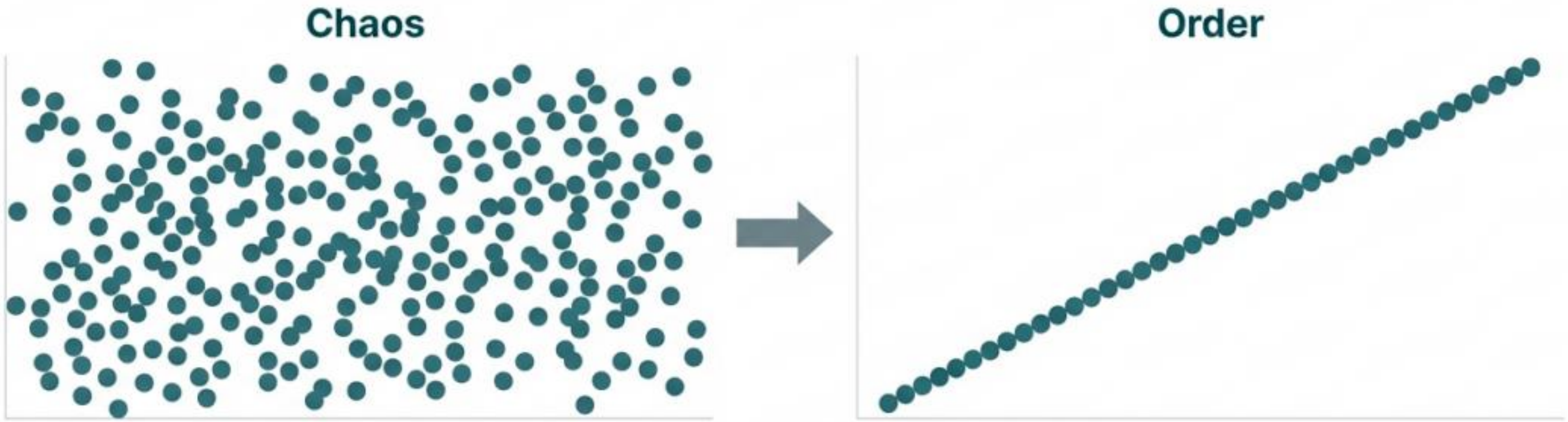
$$\sqrt{352{,}233{,}694{,}151} \longrightarrow \text{£}593{,}492$$

The Fix: To solve the unit problem, we simply use **Standard Deviation which is the square root of the Variance.**

The Result: The Standard Deviation is approx. £593,500.

Why it is important: We are back in the original units (Pounds). We can now say: The typical variation in house prices from the average is roughly £600k: This is an actionable, understandable insight.

Measure of Position: From Chaos to Order
(Determining relative standing within the dataset.)



We have measured the spread (Variability). Now, we need to understand relative standing. Measures of position tell us the point at which a certain percentage of the data falls below.

Two Key Tools:

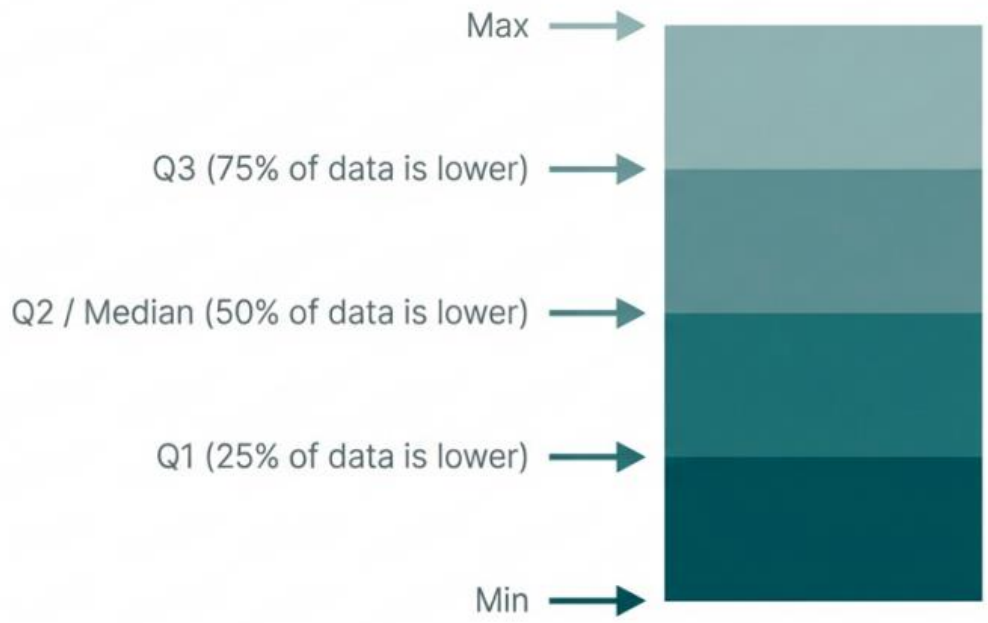
1. **Quartiles:** Splitting the sorted data into four equal chunks.

2. **Percentiles:** A fine-grained scale from 0 to 100.

UNIVERSITY of York

Quartiles: Slicing the Data into Four

Structuring the distribution



Lower quartile Q1 is the median of the first half
Middle quartile Q2 is the median
Upper quartile Q3 is the median of the second half

Example:
Consider a list of house values (in 100K):
[60, 70, 80, 85, 92, 101, 125, 150].
1st 2nd 3rd 4th 5th 6th 7th 8th

Min = 60
Q1 = 75
Q2 = 88.5
Q3 = 113
Max = 150

The Interquartile Range (IQR)

Focusing on the “core” of the data

Calculation: IQR = Q3 - Q1

Significance: The IQR ignores the cheapest 25% and the most expensive 25%.

Why it matters: Unlike the standard Range, the IQR is resistant to outliers. It ignores the £11m mansions and tells us where the core of the market actually is.

Percentiles: Fine-Grained Analysis

Answering the question: "Is this house in the top 10%?"

Definition: The n-th percentile is the value such that n% of the data falls at or below it.

Mapping Quartiles to Percentiles:

- Min = 0th Percentile
- Q1 = 25th Percentile
- Median = 50th Percentile
- Q3 = 75th Percentile
- Max = 100th Percentile

UNIVERSITY of York

Calculating in Python

Implementing the concepts with NumPy.

Variance

```
# Variance calculation
var = np.var(df['Price'])
print("Houses price variance: %.2f" % (var))
# Output: 352233694151.18
```

```
#Variance calculation
var = np.var(df['Price'])
print("Houses price variance: %.2f" % (var))
```

```
Houses price variance: 3522336694151.18
```

Standard Deviation

```
# Standard Deviation
std = np.std(data['Price'])
print("Houses price std:", std)
# Output: 593492.8
```

Percentiles

```
# Percentiles
q75 = np.percentile(df['Price'], 75)
print("75th Percentile:", q75)
```

Sample or Population?

Python NumPy Documentation

numpy.var

```
numpy.var(a, axis=None, dtype=None, out=None, ddof=0, keepdims=<no value>,
*, where=<no value>, mean=<no value>, correction=<no value>)    [source]
```

Compute the variance along the specified axis.

Returns the variance of the array elements, a measure of the spread of a distribution. The variance is computed for the flattened array by default, otherwise over the specified axis.

Function name

Function return

Required parameters

A parameter **without =** is required.
A parameter **with =** is optional because it has a default value.

Keyword Only Separator

Everything **before *** can be passed positionally.
Everything **after *** must be passed by keyword.

https://numpy.org/devdocs/reference/generated/numpy.var.html

Python NumPy Documentation

UNIVERSITY of York

Parameters:

a : *array_like*

Array containing numbers whose variance is desired. If *a* is not an array, a conversion is attempted.

ddof : *{int, float}*, *optional*

"Delta Degrees of Freedom": the divisor used in the calculation is `N - ddof`, where `N` represents the number of elements. By default *ddof* is zero. See notes for details about use of *ddof*.

- Must be an integer or a float.
- Optional. Default value is 0.
- The divisor used in the calculation is N − ddof.

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

ddof=1 → sample variance

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

ddof=0 → population variance

Different values of the argument *ddof* are useful in different contexts. NumPy's default `ddof=0` corresponds with the expression:

$$\frac{\sum_i |a_i - \bar{a}|^2}{N}$$

which is sometimes called the "population variance" in the field of statistics because it applies the definition of variance to *a* as if *a* were a complete population of possible observations.

Many other libraries define the variance of an array differently, e.g.:

$$\frac{\sum_i |a_i - \bar{a}|^2}{N - 1}$$

In statistics, the resulting quantity is sometimes called the "sample variance" because if *a* is a random sample from a larger population, this calculation provides an unbiased estimate of the variance of the population. The use of $N - 1$ in the denominator is often called "Bessel's correction" because it corrects for bias (toward lower values) in the variance estimate introduced when the sample mean of *a* is used in place of the true mean of the population. For this quantity, use `ddof=1`.

UNIVERSITY of York

Numpy Explanation

Choosing the right measure

Measure	Use When…	Watch out for…
Range	You need a quick summary of the spread between the smallest and largest values	Highly sensitive to outliers, which can give a misleading impression of variability.
Variance	You want a measure of spread that uses all observations in the dataset.	Results are in squared units, which can be hard to interpret. Do $n - 1$ correction for sample data.
Standard Deviation	You want to describe typical variability in the original units of the data.	Remember to apply the $n - 1$ correction for sample data.
Percentiles / IQR	You need a measure of relative position or that is resistant to outliers.	Results can vary slightly depending on the interpolation method used.

Summary

- Data can also be classified into **categorical** and **numerical** data types, which often determine the choice of appropriate statistical methods.

- Measures of **frequency**, **central tendency**, **variation**, and **position** are fundamental descriptive statistics that help us to summarise and interpret datasets effectively.

Further Readings

- Introduction to Data Science A Python Approach to Concepts, Techniques and Applications (Chapter 3) – available on VLE
- Introduction to NumPy from the Python Data Science Handbook (Chapter 2)
- Data Manipulation with Pandas from the Python Data Science Handbook (Chapter 3)