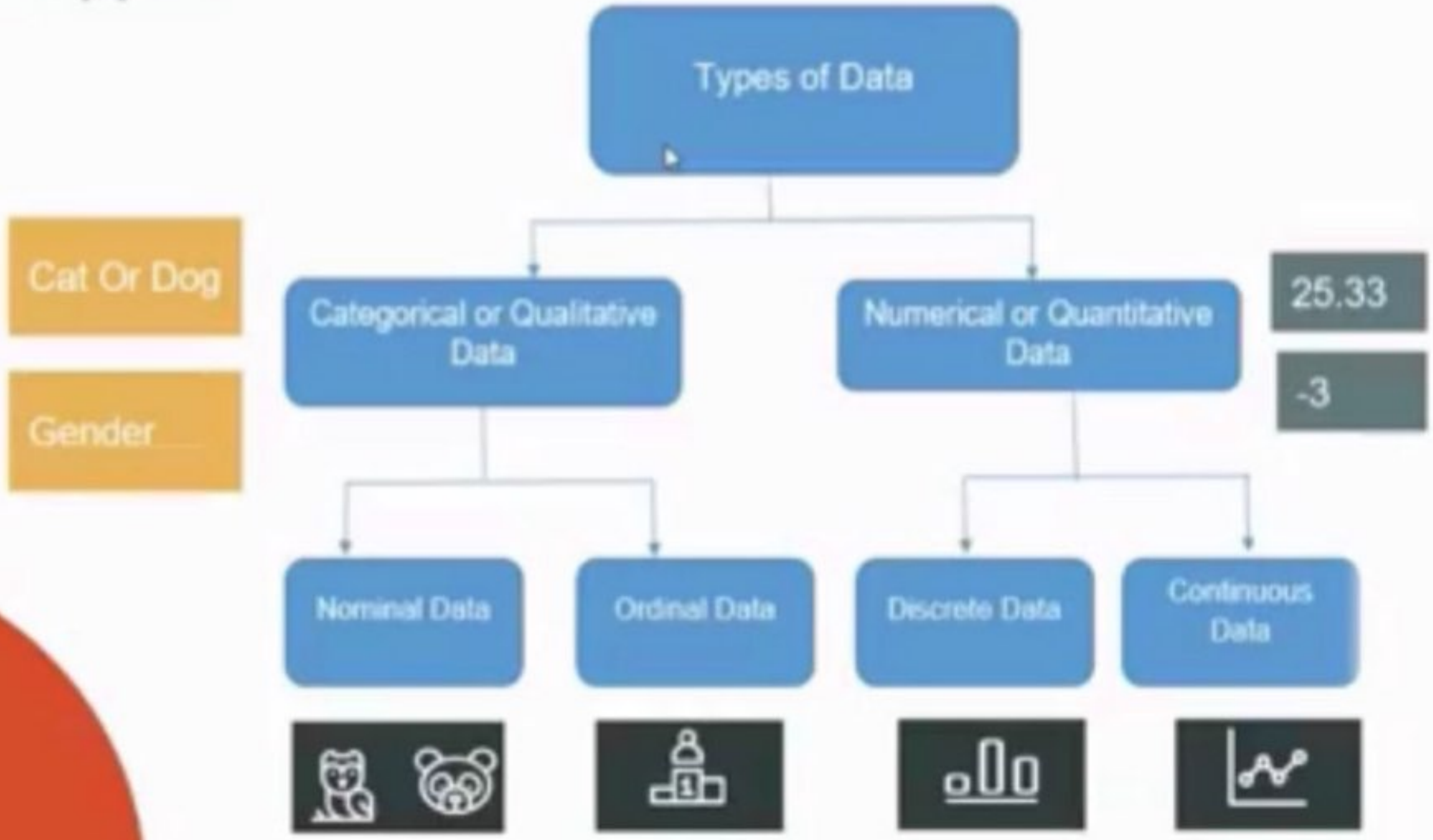
Data Types



Types of Data

Categorical or Qualitative Data

Numerical or Quantitative Data

Nominal Data

Ordinal Data

Discrete Data

Continuous Data

Cat Or Dog

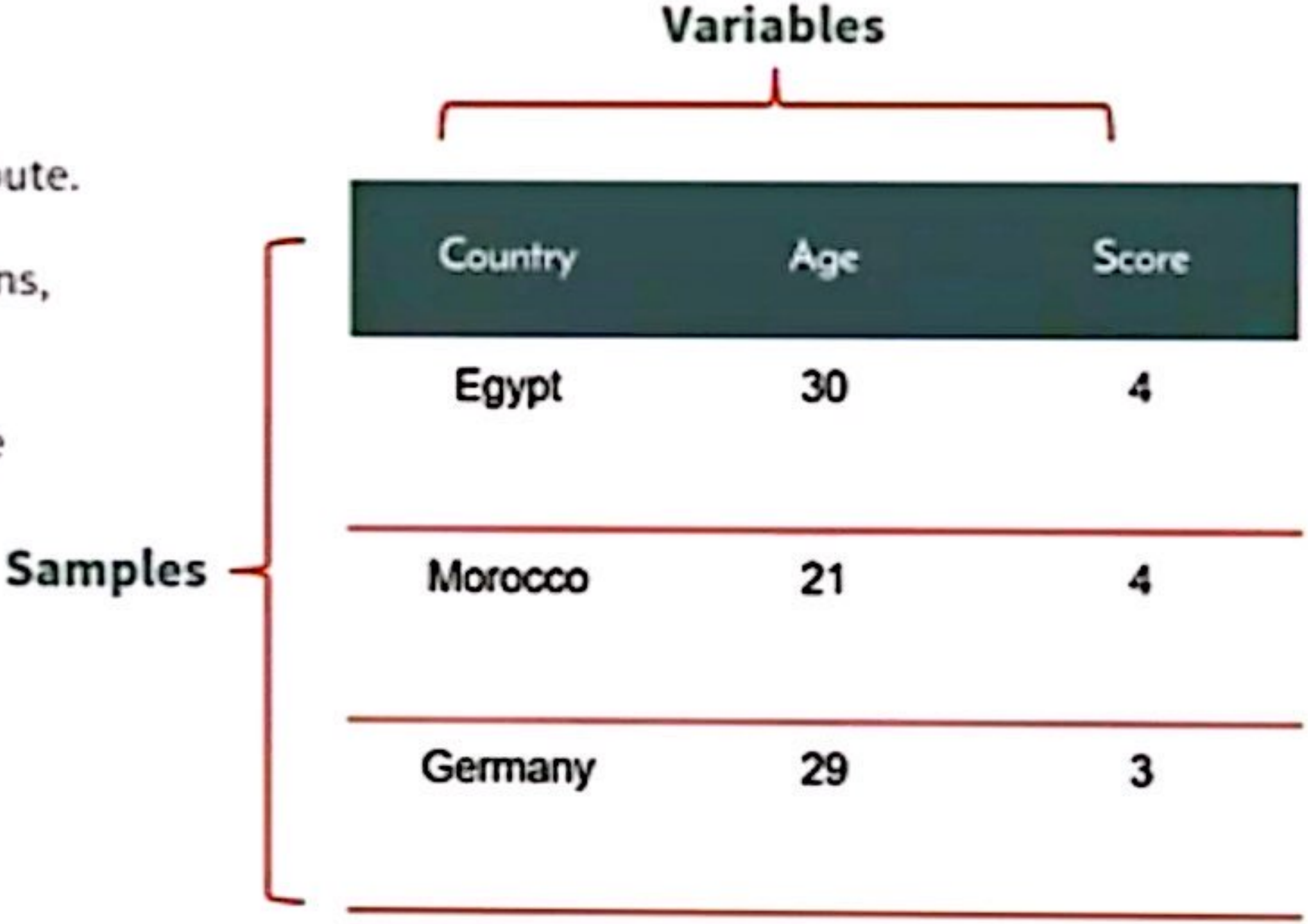
Gender

25.33

-3

Data Terminologies

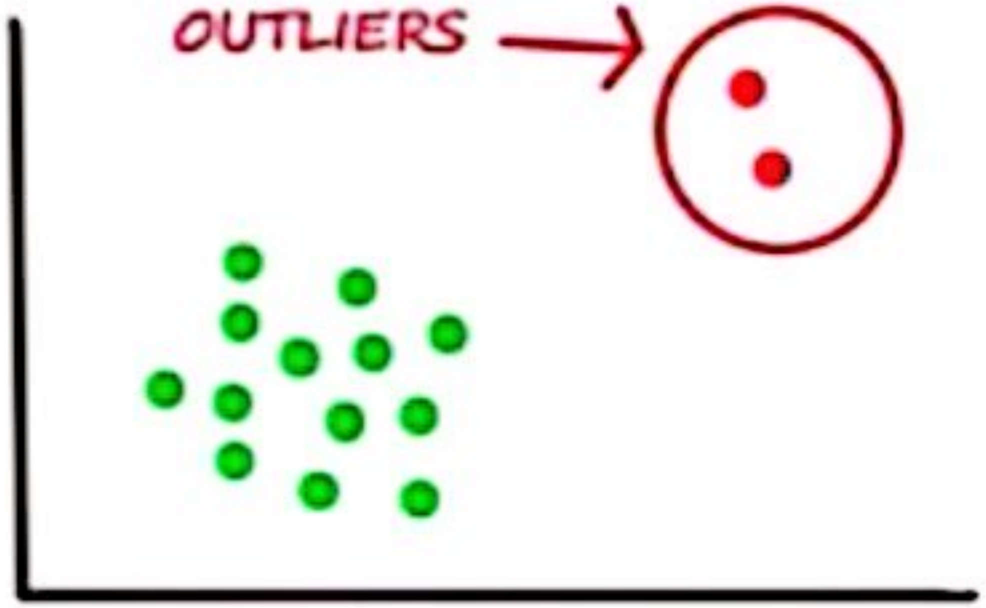
- Variable are also called Column, Feature, Dimension, field and Attribute.
- Samples are also called Observations, Records, Instances and rows.
- Variables and Samples make up the term "Data Set" or "Data Frame".



Country	Age	Score
Egypt	30	4
Morocco	21	4
Germany	29	3

Outliers

1. An outlier is a data point that differs **significantly** from other observations.
2. We usually tend to remove the outliers to make sure that we are making accurate analysis.
3. **Outliers** can cause serious problems in statistical analyses



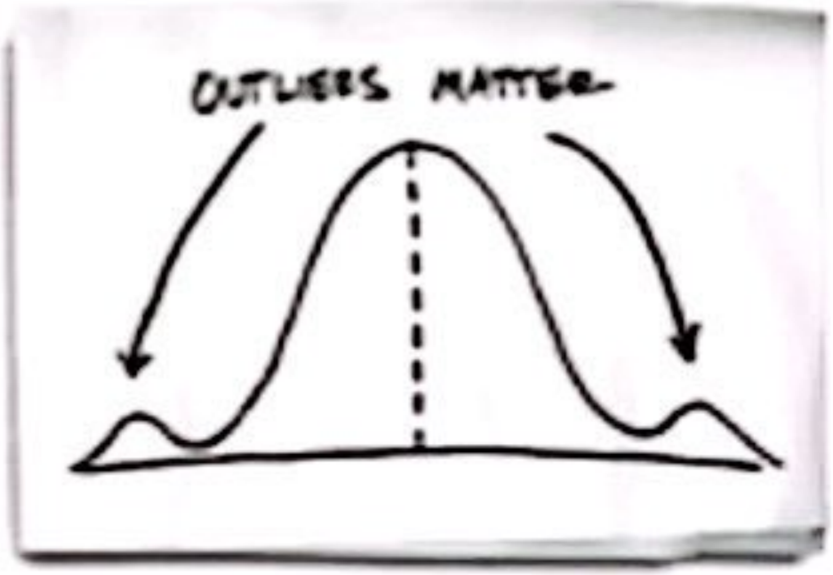
Removing Outliers

There are different techniques to capture outliers

1. We usually tend to remove the outliers to make sure that we are making accurate analysis, But other times we keep them. Example: **Fraud Detection Applications.**

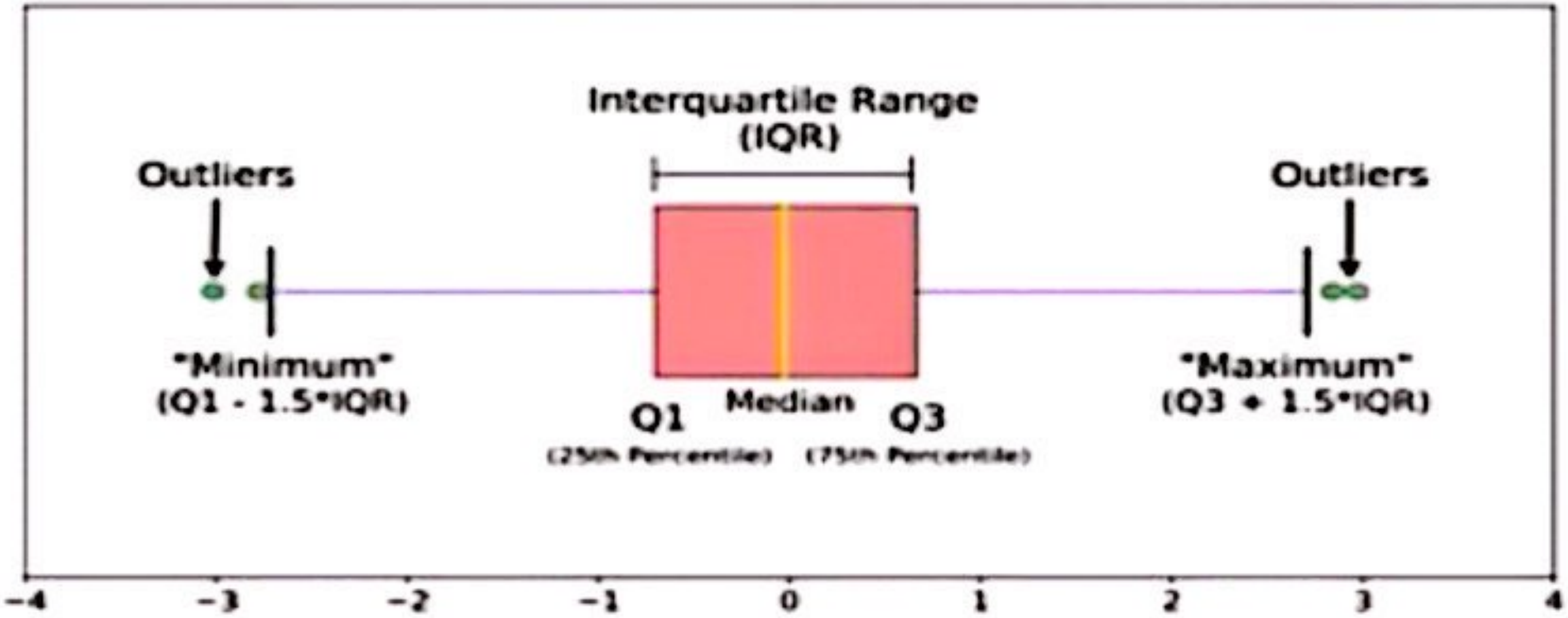
2. We can **normalize** our outliers to make them look like the majority, not to remove them if their removal will have bad effects like in in case of small datasets.

3. So we have to always visualize our data and analyze it properly before making any moves



Removing Outliers – Tukey's Method

- One of the most used methods to detect and remove outliers is the Tukey's method.
- First We calculate the IQR like IQR = Q3 value – Q1 value
- Then any data points < (Q1 – 1.5 * IQR) and > (Q3 + 1.5 * IQR) is considered an outlier.
- Outliers are **X**, Where **(Q3 + 1.5 * IQR) < X < (Q1 - 1.5 * IQR)**

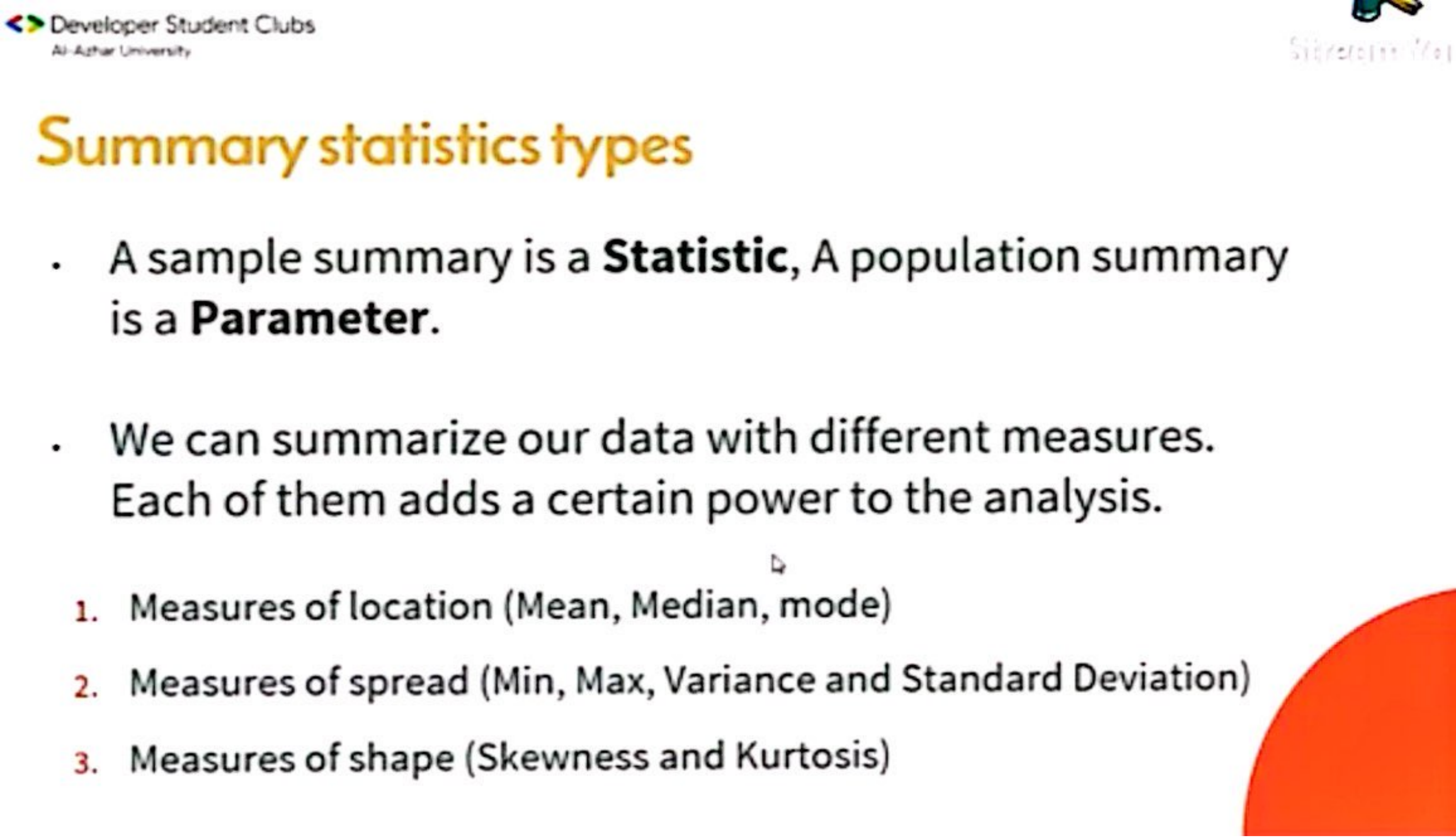


03

Summary Statistics

“ONE NUMBER TO RULE THEM ALL ,
ONE NUMBER TO FIND THEM ALL ,
ONE NUMBER TO BRING THEM ALL ”

-Gandalf The Grey

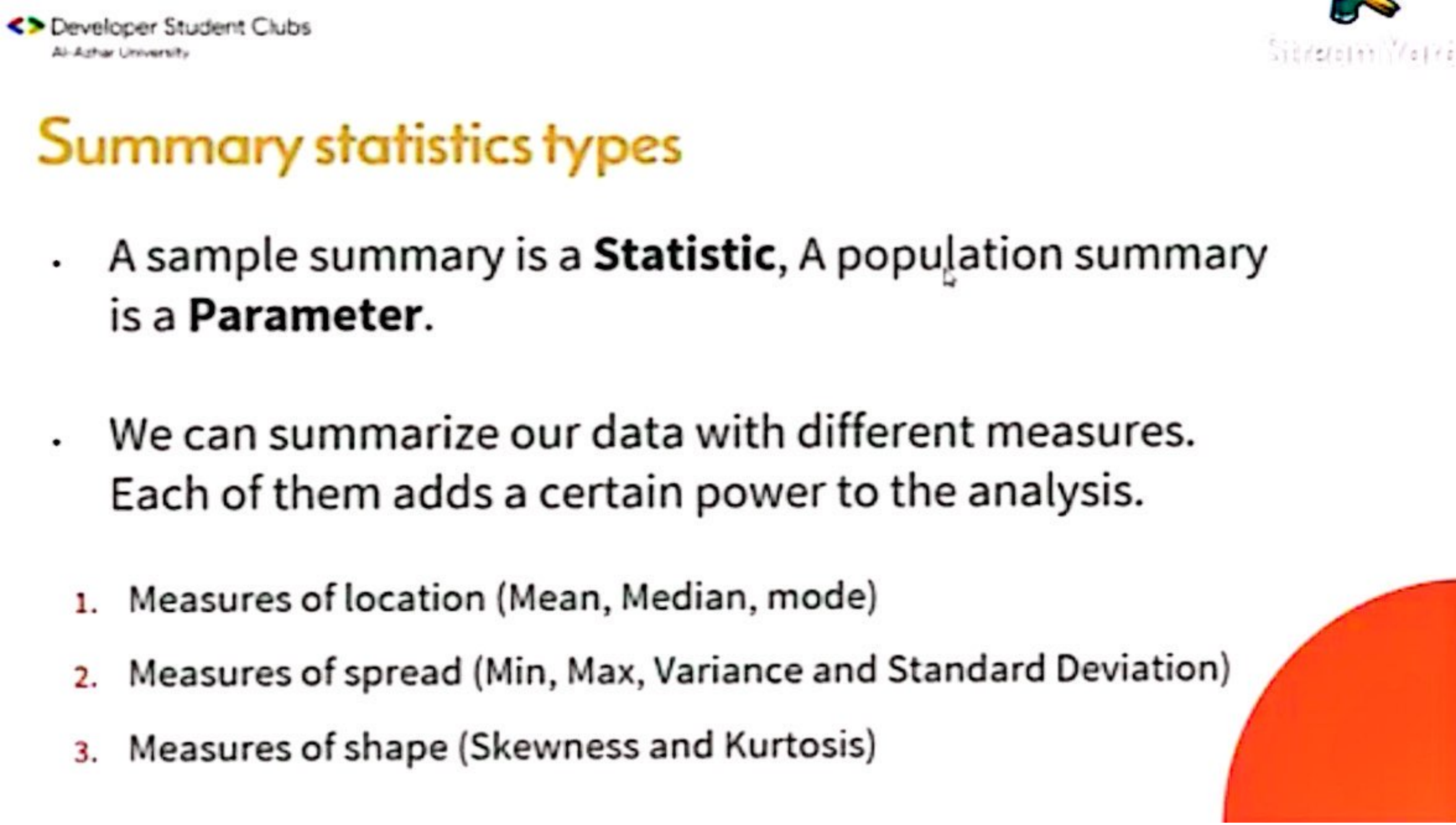


Summary statistics types

- A sample summary is a **Statistic**, A population summary is a **Parameter**.

- We can summarize our data with different measures. Each of them adds a certain power to the analysis.

1. Measures of location (Mean, Median, mode)
2. Measures of spread (Min, Max, Variance and Standard Deviation)
3. Measures of shape (Skewness and Kurtosis)



Summary statistics types

- A sample summary is a **Statistic**, A population summary is a **Parameter**.

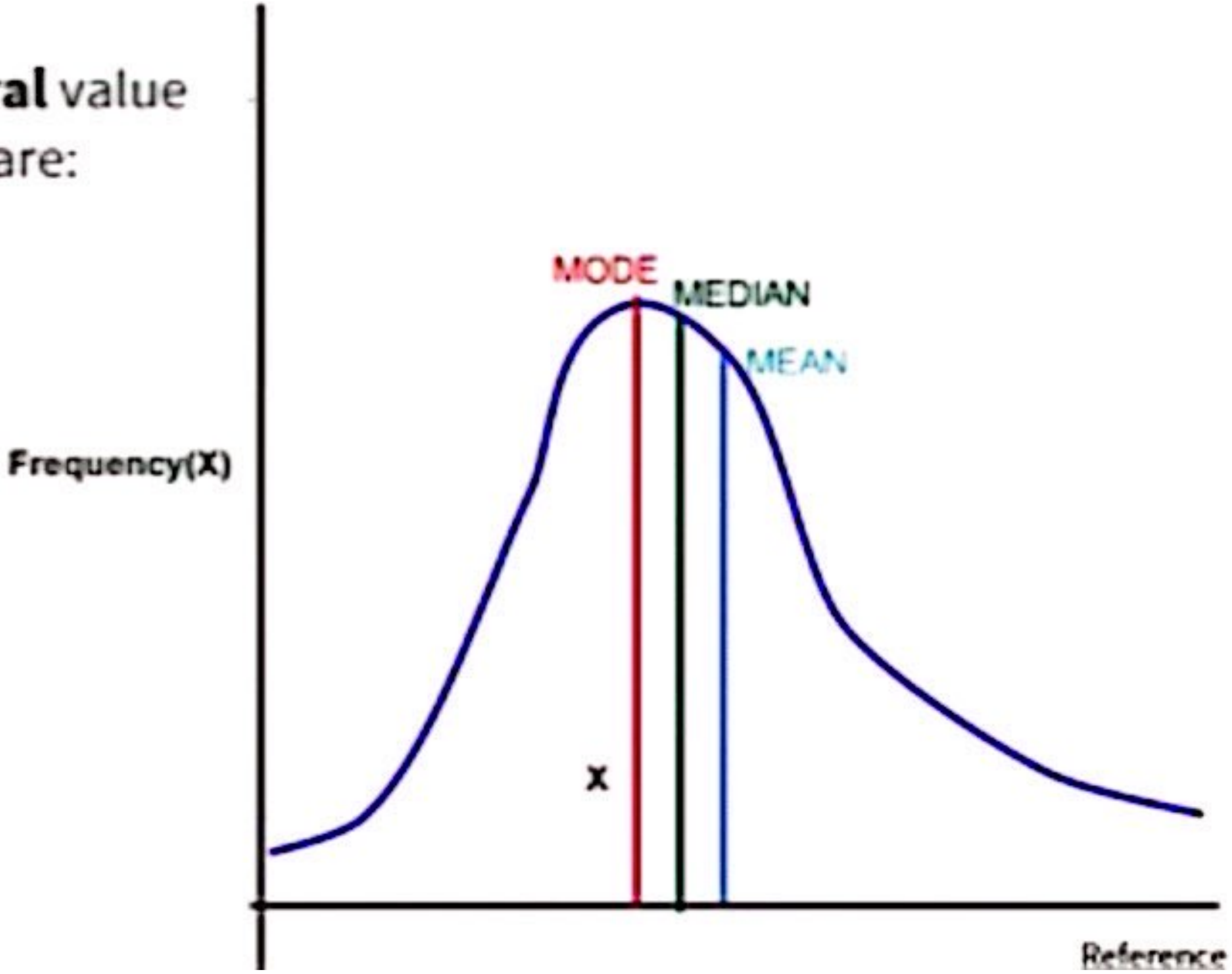
- We can summarize our data with different measures. Each of them adds a certain power to the analysis.

1. Measures of location (Mean, Median, mode)
2. Measures of spread (Min, Max, Variance and Standard Deviation)
3. Measures of shape (Skewness and Kurtosis)

Measures of location (Center)

It's the measures that describes the **central** value of a data set, And Its most popular forms are:

1. Mean
2. Median
3. Mode



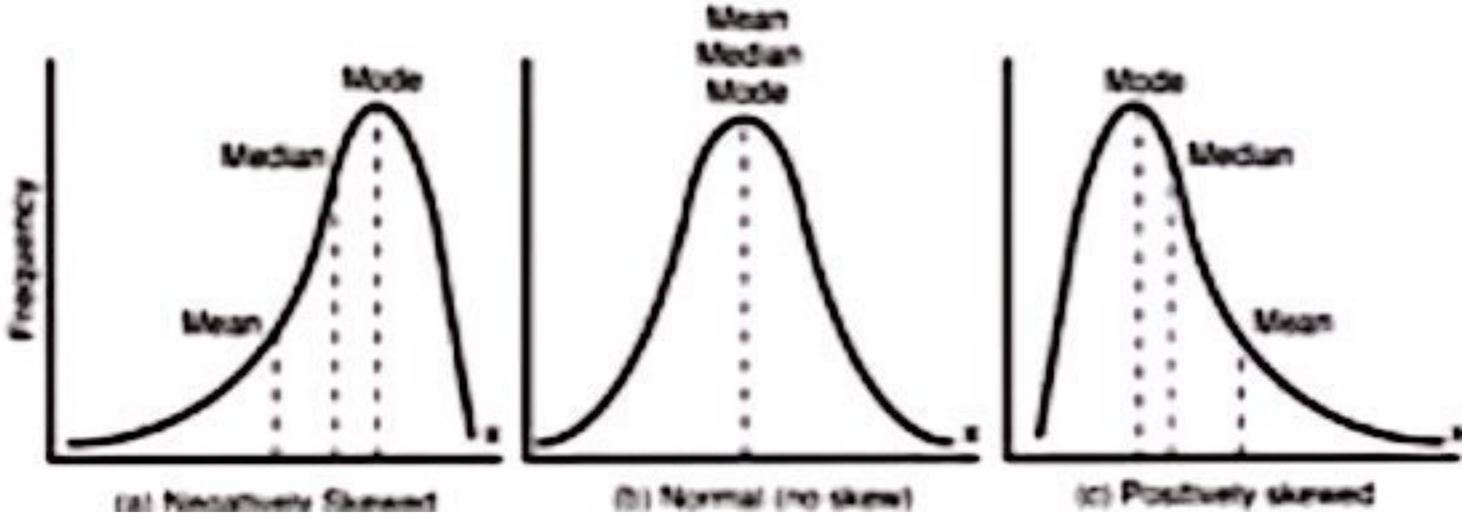
Mean

It's the sum of all values of the data set divided by its records number

1. It's the **simplest** computed summary statistic.
2. Suitable for general-purposed analysis.
3. **Can be computed algebraically**. Median and Mode can not be algebraically manipulated.
4. The mean is more widely used than median and mode.
5. Very Sensitive to outliers and **skewness**.
6. Can't handle **non-numeric** features.
7. **Catches the variability** of data points.

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	*n* = number of items in the sample

Reference



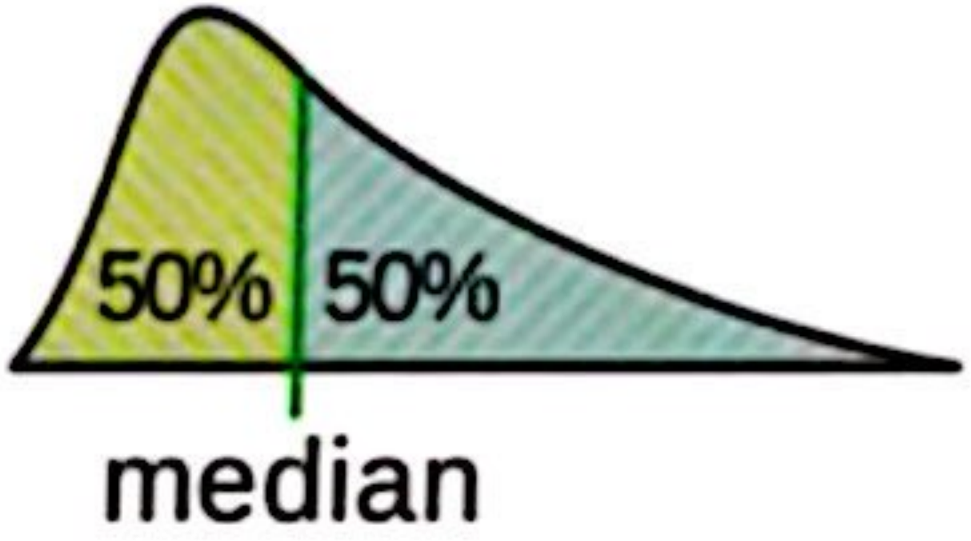
(a) Negatively Skewed

(b) Normal (no skew)

(c) Positively skewed

Median

It's the middle value of our data set.



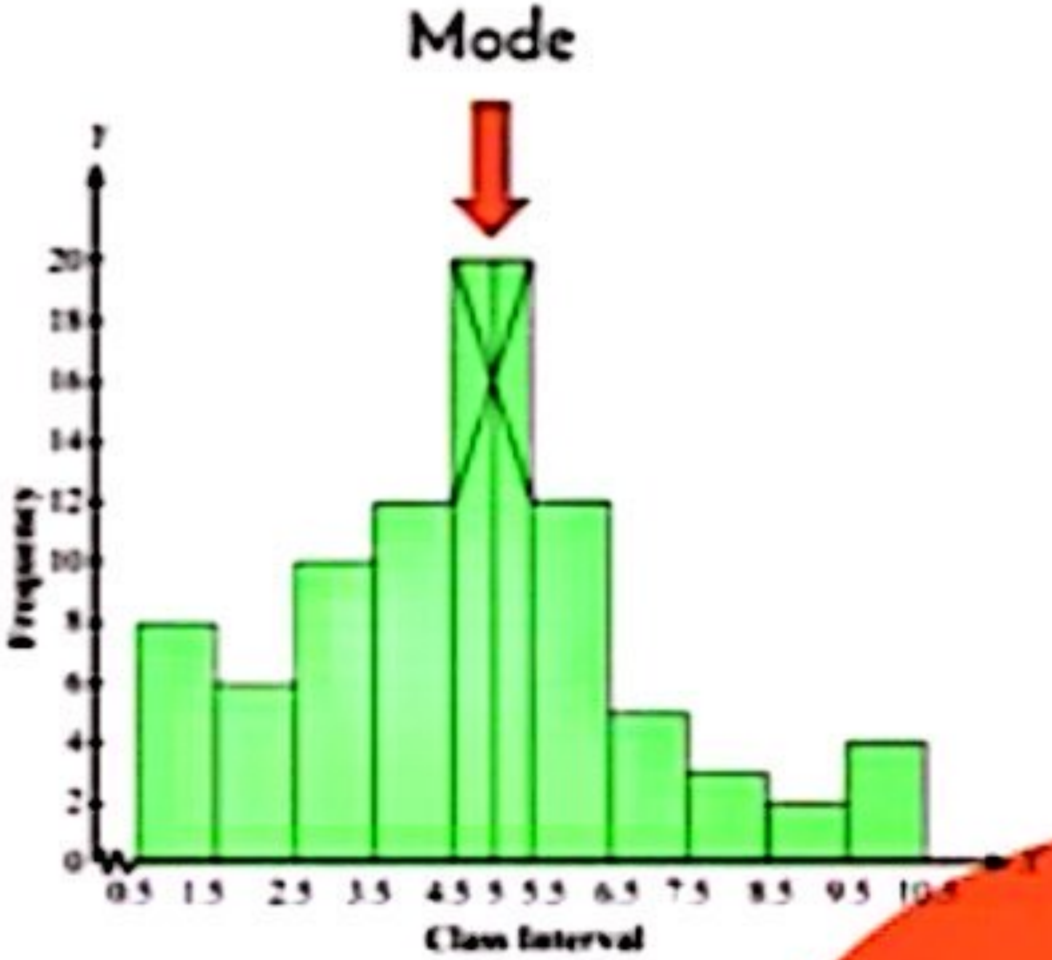
1. The median value is fixed by its position and is not reflected by the individual value.
2. Can be used to determine an approximate average if there were outliers in the data.
3. Can't be computed Algebarically.
4. Before applying the law of the median, **the data must be sorted first**.

$$\text{Median} = \begin{cases} \dfrac{(N+1)}{2}^{th} \text{ term; when N is odd} \\ \dfrac{\dfrac{N}{2}^{th} \text{ term} + \left(\dfrac{N}{2}+1\right) \text{term}}{2} \text{ ; when N is even} \end{cases}$$

Mode

It's the element that appeared the most in our dataset.

1. We can have **multiple modes** in the dataset.
2. Unlike mean, it has no mathematical property
3. Unlike mean, Mode is **affected by sampling fluctuations**.
4. It's the most suitable measure for **nominal data**.



Pros and Cons

	Outliers Sensitive ?	Algebric Manipulation	Qualitative Expression	Fluctuations of sampling
Mean	✓	✓	✗	✗
Median	✗	✗	✓	✓
Mode	✗	✗	✓	✓

CONCLUSION

To summarize our lecture we can say:

1. It's strongly believed that Arabs are the pioneers of statistics.
2. Outliers are generally bad for our analysis but sometimes they are the most important.
3. Summary statistics is a must when working with data.
4. Mean is the most popular measure of location or center.
5. We can use the other summaries like median and mode in special cases like outliers presence.
6. When data are on interval scale the suitable measure of central tendency is mean. Median is suitable when data are on ordinal scale. Mode is calculated when data are on nominal scale.

What is coming ?

Measures of Spread

Dataset Variability

Measures of Shape

Skewness and Kurtosis

Visuals

Box plot, histograms ..etc

Full EDA

Practising what we have learned.