

# Intro to Descriptive Statistics

By: Mahmoud Galal



Developer Student Clubs

Al-Azhar University

# AGENDA

01

## History



Knowing the past ...  
Respecting the present.

02

## Data Types and Terminologies



Knowing the type of the data  
is knowing what to do with it.

03

## Summary Statistics



Representing a great sum of  
data with just one number !

04

## Exploratory Data Analysis



Using the swords we mastered to  
slice down the beast, Just to know  
him better (We are not savages !)

-- It's said that data is the biggest beast of our age --



# 01

## History...

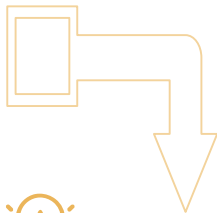
**Knowing** the past ... **Respecting** the present.



# How statistics evolved?



Until the 18<sup>th</sup> century data was collected and putted into dummy tables.



From the 19<sup>th</sup> century and onwards, Life got complicated, and people started to think about how to collect, summarize and present the data.



Nowadays, statistics has evolved to contribute in every field of study in our lives and even just Fun !

# Founders ...



Abu Youssef Yaaqob Al Kindi Developed the first code breaking algorithm based on **frequency analysis**. He wrote a book entitled "Manuscript on Deciphering Cryptographic Messages", containing detailed discussions on **statistics**.



Check rest of the founder like Bayes , Laplace and Gauss [here](#).



أبو يوسف يعقوب بن إسحاق الصبّاح الكندي

# Who got it?






# 02

## Data Types and Terminologies

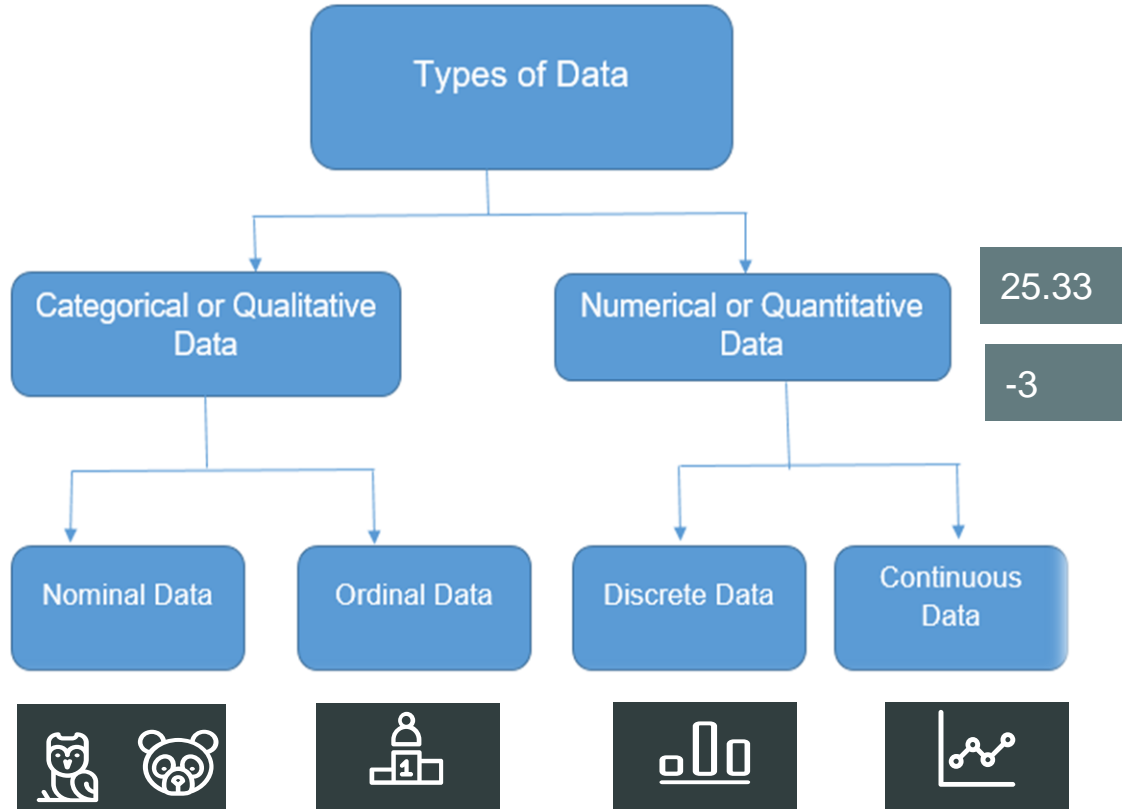
Knowing the type of the data  
is knowing what to do with it.



# Data Types

Cat Or Dog

Gender





# 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”.

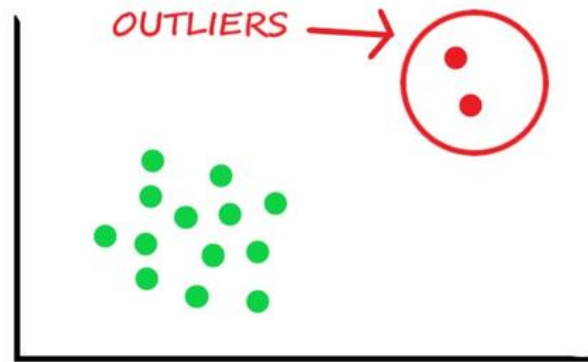
**Samples**

**Variables**

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

# Outliers

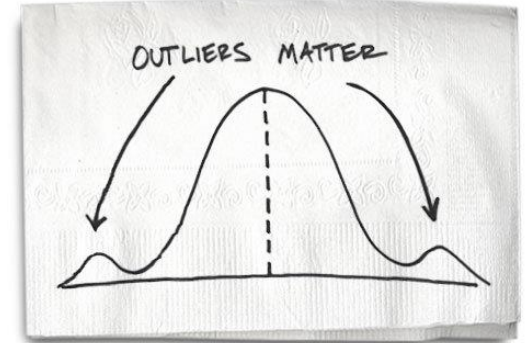
1. An outlier is a data point that differs **significantly** from other observations.
2. We usually tend to remove the outliers to make sure that we are making accurate analysis.
3. **Outliers** can cause serious problems in statistical analyses



# Removing Outliers

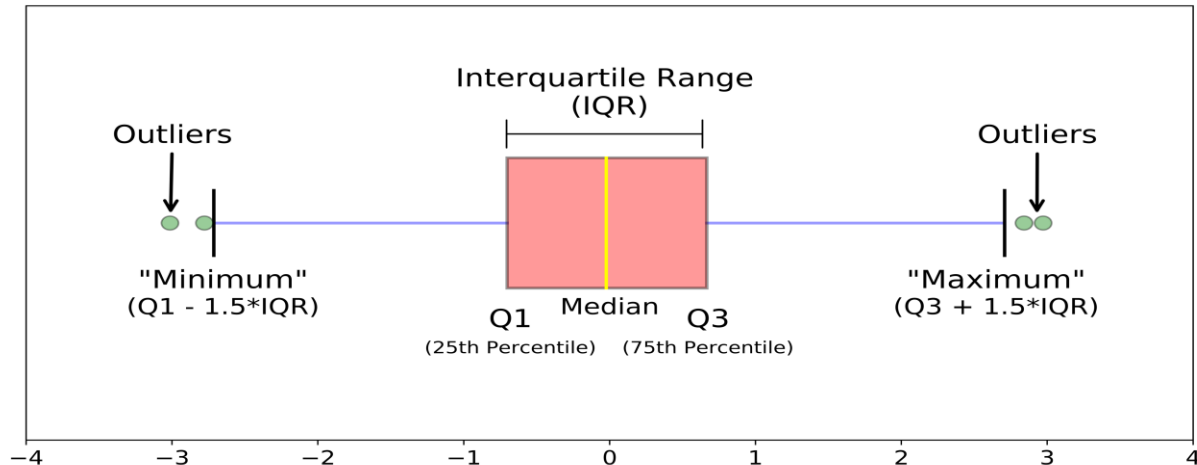
There are different techniques to capture outliers

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  $IQR = Q3 \text{ value} - Q1 \text{ value}$
- Then any data points  $< (Q1 - 1.5 * IQR)$  and  $> (Q3 + 1.5 * IQR)$  is considered an outlier.
- Outliers are **X**, Where  **$(Q3 + 1.5 * IQR) < X < (Q1 - 1.5 * IQR)$**



# Real-World Example



# 03

## Summary Statistics

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

-Gandalf The Grey

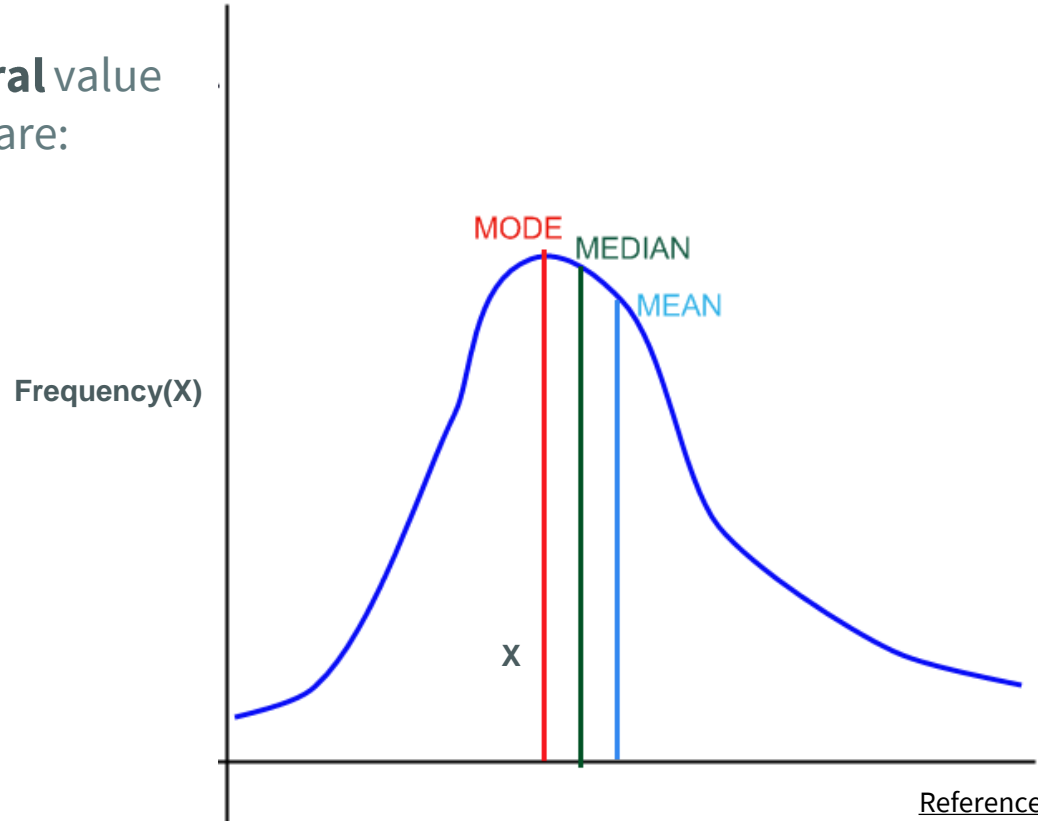
# Summary statistics types

- 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)

# 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





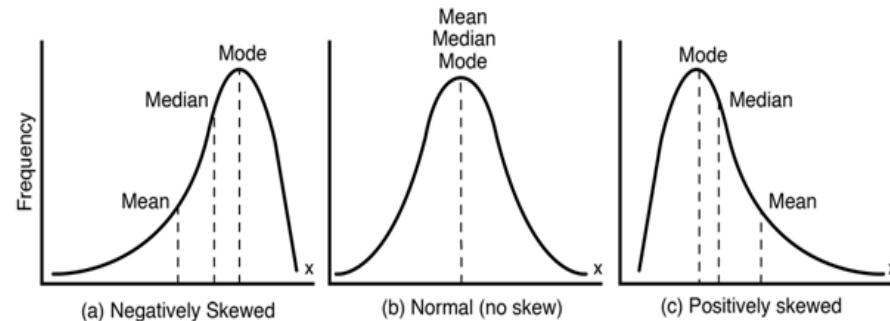
# Mean

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                                                                                  |
|----------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| $\mu = \frac{\sum_{i=1}^N x_i}{N}$ <p><math>N</math> = number of items in the population</p> | $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$ <p><math>n</math> = number of items in the sample</p> |

Reference



# Practice

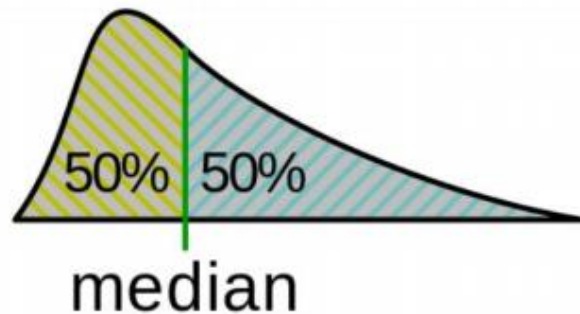
# 5-Min Break

Please submit your attendance while  
Making yourself something hot ☕

# 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 Algebraically.
4. Before applying the law of the median, **the data must be sorted first.**

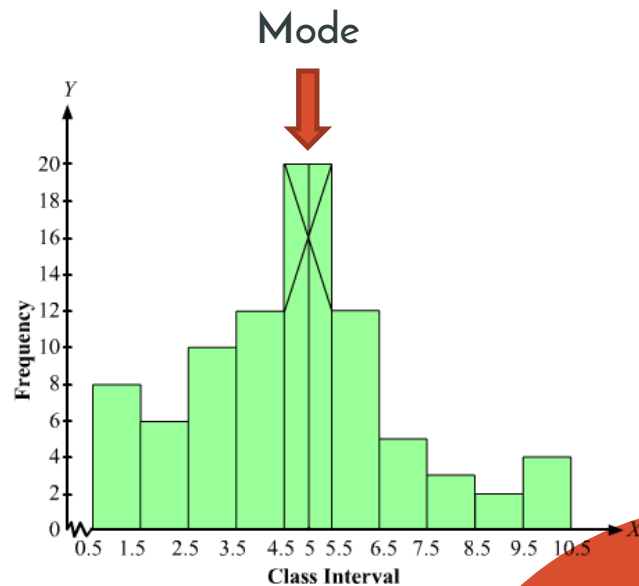


$$\text{Median} = \begin{cases} \frac{(N+1)^{\text{th}}}{2} \text{ term; when } N \text{ is odd} \\ \frac{\frac{N}{2} \text{ term} + \left(\frac{N}{2} + 1\right) \text{ term}}{2}; \text{ when } N \text{ is even} \end{cases}$$

# 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**.



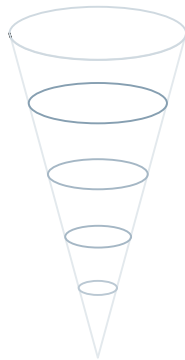
# Practice

# Pros and Cons

|        | Outliers Sensitive ? | Algebraic Manipulation | Qualitative Expression | Fluctuations of sampling |
|--------|----------------------|------------------------|------------------------|--------------------------|
| Mean   | ✓                    | ✓                      | ✗                      | ✗                        |
| Median | ✗                    | ✗                      | ✓                      | ✓                        |
| Mode   | ✗                    | ✗                      | ✓                      | ✓✓                       |

# 7,540,860,914

Just imagine a dataset with this number of rows...



1



**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

## Visuals

Box plot,  
histograms ..etc

## Measures of Shape

Skewness and  
Kurtosis

## Full EDA

Practising what we  
have learned.

---

“Statistics is the grammar of Sceince

–Karl Pearson (Who is he?)

# THANKS

Any Questions? ...

Head to google meet !