EXPLORATORY DATA ANALYSIS

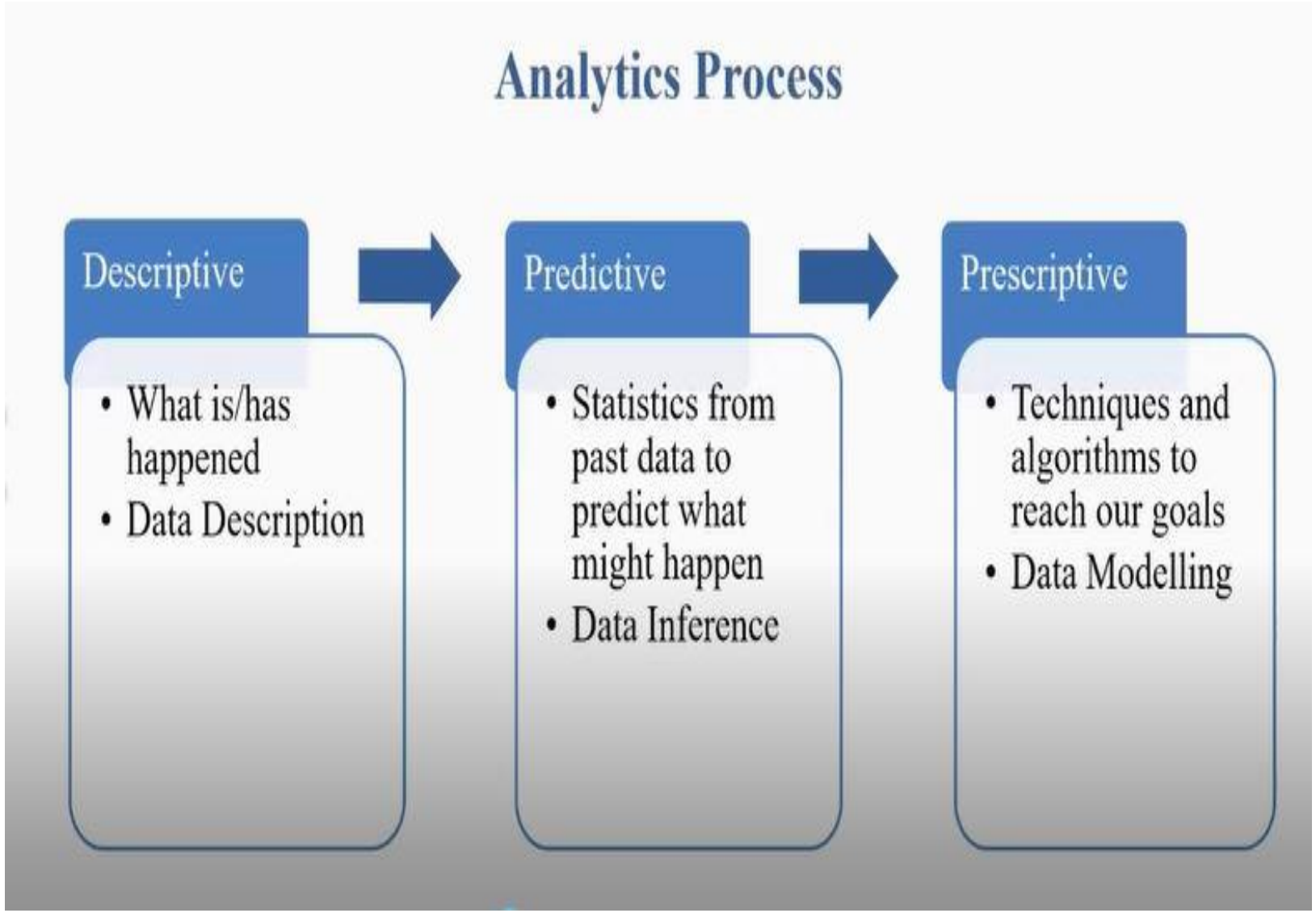
DATASCIENCE USING PYTHON

TOPIC: 3

ANALYTICS

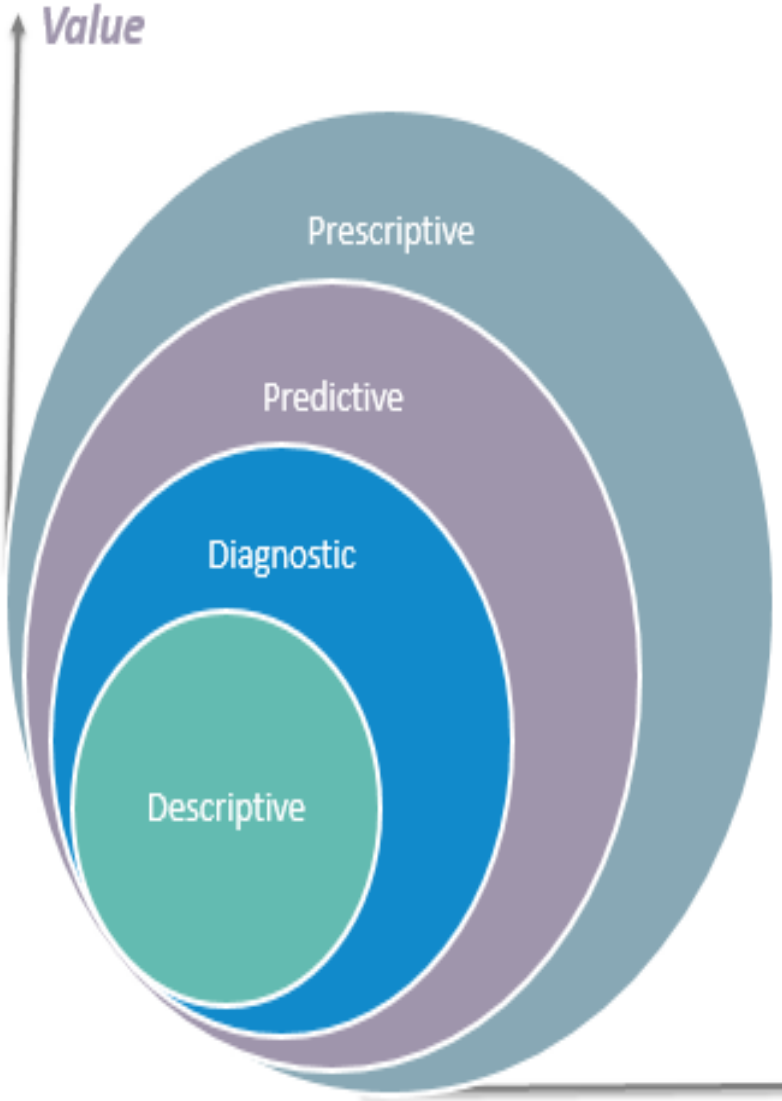
- Analytics is the process of transforming data into insights for making a better decisions:
- **Statistics**
- **Data Mining**
- **Technology**

Analytic Processes



TYPES OF DATA ANALYTICS

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

3 A's of Analytic Processes

3 A's of Analytics Process

Based on 3A's will develop new model to the problem-solving operation:

- **Analytics (What?)**
- **Attribution (How?)**
- **Algorithm (Solution?)**

Field where Business Analytics involved?

Finance

- Credit scoring, fraud detection, pricing, claims analysis

Retail

- Promotions, replenishment, demand forecasting, merchandising optimisation

Manufacturing

- Inventory replenishment, product customisation, supply chain optimisation

Health care

- Drug interaction, preliminary diagnosis, disease management

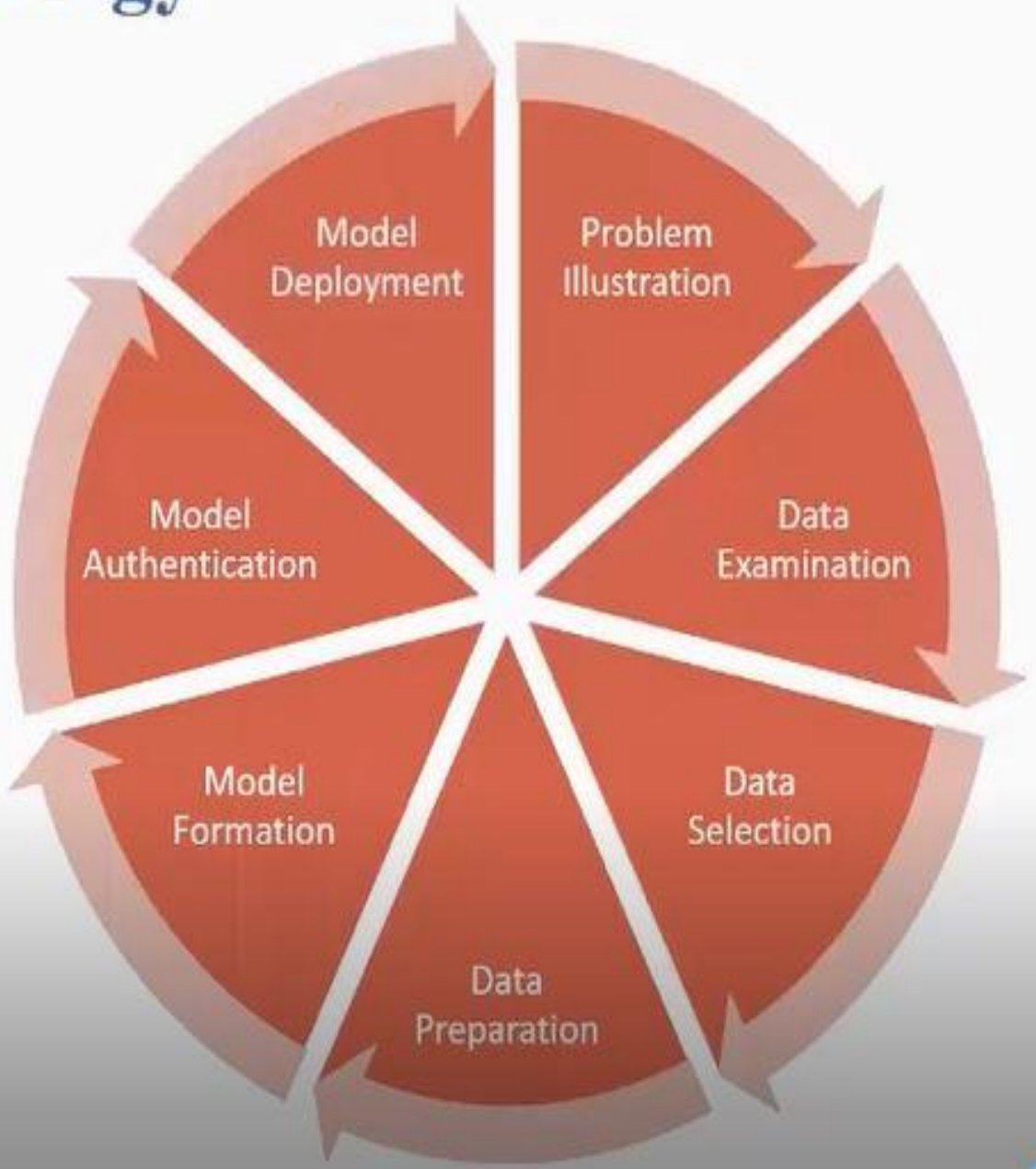
Energy

- Trading, supply, demand forecasting, compliance

Communications

- Customer retention, capacity planning, network optimisation

Analytics Methodology

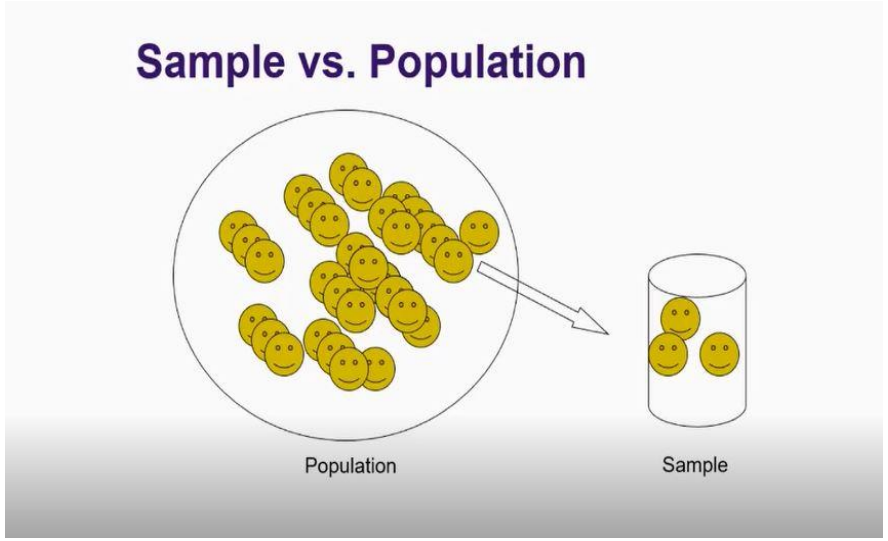


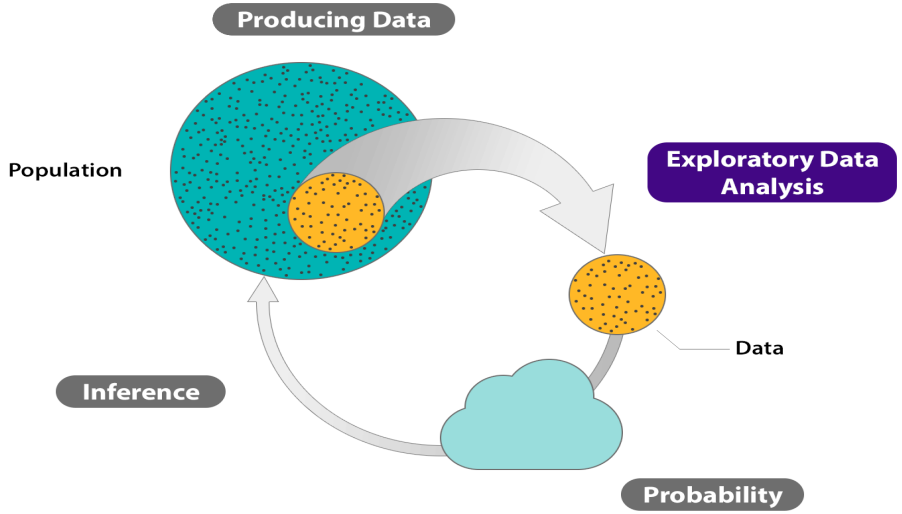
EDA

Exploratory Data Analysis (EDA) is an approach/ philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set;

- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

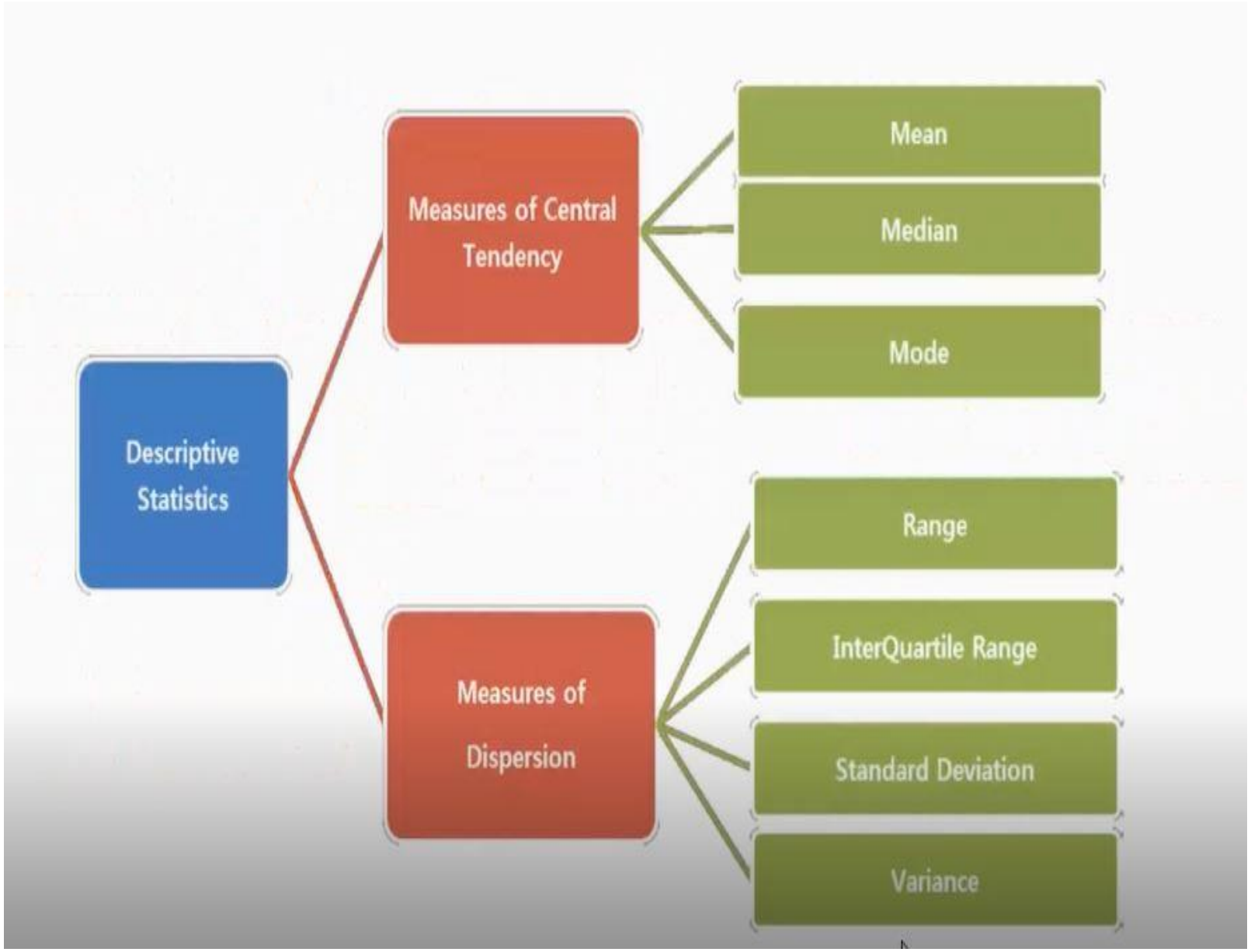
Sample And Population





Exploratory Data Analysis

DESCRIPTIVE STATISTICS



WHY STATISTICS?

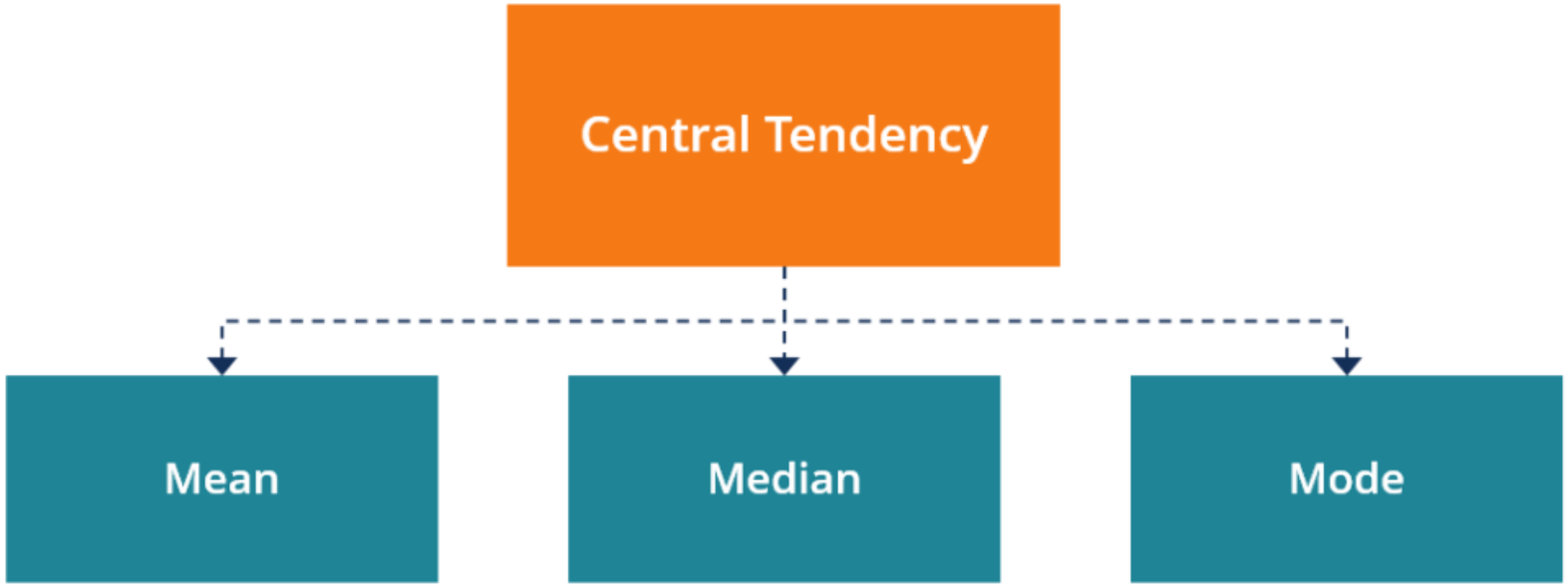
- **Enables classification and organization**
- This is a statistical method that's used by the same name in the data science and mining fields. Classification is used to categorize available data into accurate, observable analyses. Such an organization is key for companies who plan to use these insights to make predictions and form business plans. It's also the first step to making a massive dump of data usable.
- **Helps to calculate probability distribution and estimation**
- These statistical methods are key to learning the basics of machine learning and algorithms like logistic regressions. Cross-validation and LOOCV techniques are also inherently statistical tools that have been brought into the Machine Learning and Data Analytics world for inference-based research, A/B and hypothesis testing.
- **Finds structure in data**
- Companies often find themselves having to deal with massive dumps of data from a panoply of source, each more complicated than the last. Statistics can help to spot anomalies and trends in this data, further allowing researchers to discard irrelevant data at a very early stage instead of sifting through data and wasting time, efforts and resources.
- **Facilitates statistical modeling**
- Data is made up of series upon series of complex interactions between factors and variables. To model these or display them in a coherent manner, statistical modeling using graphs and networks is key. This also helps to identify and account for the influence of hierarchies in global structures and escalate local models to a global scene.

WHY STATISTICS?

- **Aids data visualization**
- Visualization in data is the representation and interpretation of found structures, models and insights in interactive, understandable and effective formats. It's also crucial that these formats be easy to update– this way, nothing needs to undergo a huge overhaul each time there's a fluctuation in data. Beyond this, data analytics representations also use the same display formats as statistics– graphs, pie charts, histograms and the like. Not only does this make data more readable and interesting, but it also makes it much easier to spot trends or flaws and offset or enhance them as required.
- **Facilitates understanding of distributions in model-based data analytics**
- Statistics can help to identify clusters in data or even additional structures that are dependent on space, time and other variable factors. Reporting on values and networks without statistical distribution methods can lead to estimates that don't account for variability, which can make or break your results. Small wonder, then, that the method of distribution is a key contributor to statistics and to data analytics and visualization as a whole.
- **Aids in mathematical analysis and reduces assumptions**
- The basics of mathematical analysis– differentiability and continuity– also form the base of many major ML/ AI/ data analytics algorithms. Neural networks in deep learning are effectively guided by the shift in perspective that is differential programming.

MEASURE OF CENTRAL TENDANCIES

- One number that best summarizes the entire set of measurements.
- It is also called measure of central location.
- A number that is in some way "central" to the set.
- Depending upon the application various methods are used.

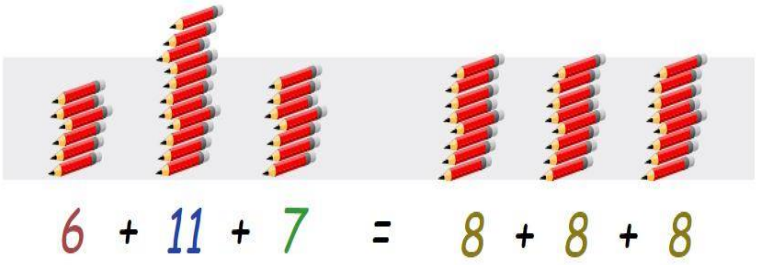


mean

The Mean

$$\overline{x} = \frac{\sum x}{N}$$

Here,
$\sum$ represents the summation
X represents scores
N represents number of scores.



6 + 11 + 7 = 8 + 8 + 8

It is like you are "flattening out" the numbers

Find the Mean

1 2 3 4 5 6 7 8

46.4, 29.3, 48.2, 35.1, 46.4, 39.5, 41.3, 25.2

$$\frac{46.4+29.3+48.2+35.1+46.4+39.5+41.3+25.2}{8}$$

$$\text{mean} = \frac{311.4}{8} =$$

median

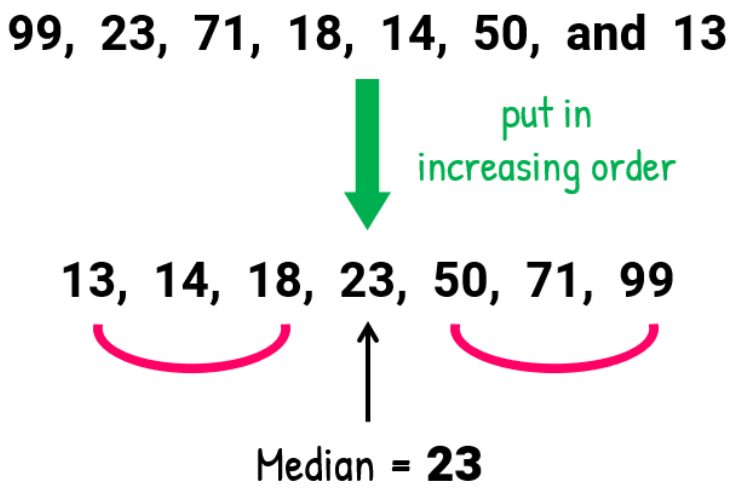
The Median

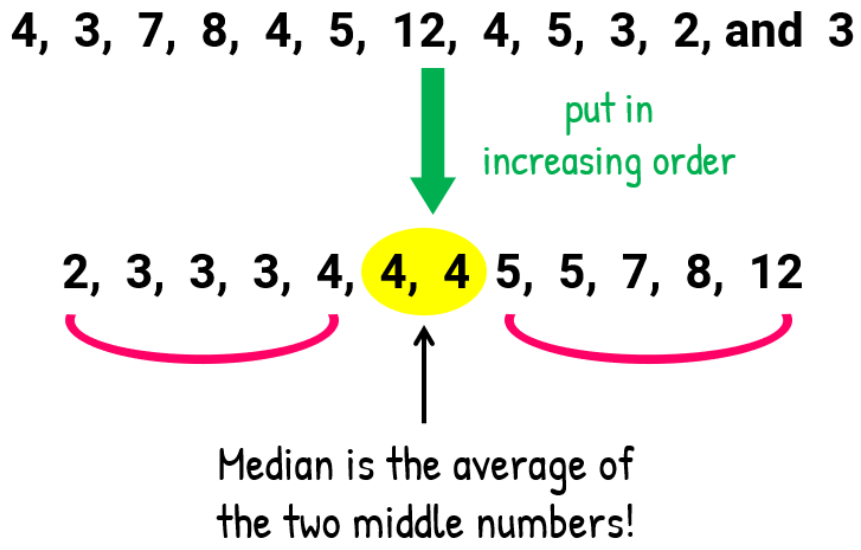
If the total number of numbers(n) is an odd number, then the formula is given below:

$$Median = \left(\frac{n+1}{2}\right)^{th} term$$

If the total number of the numbers(n) is an even number, then the formula is given below:

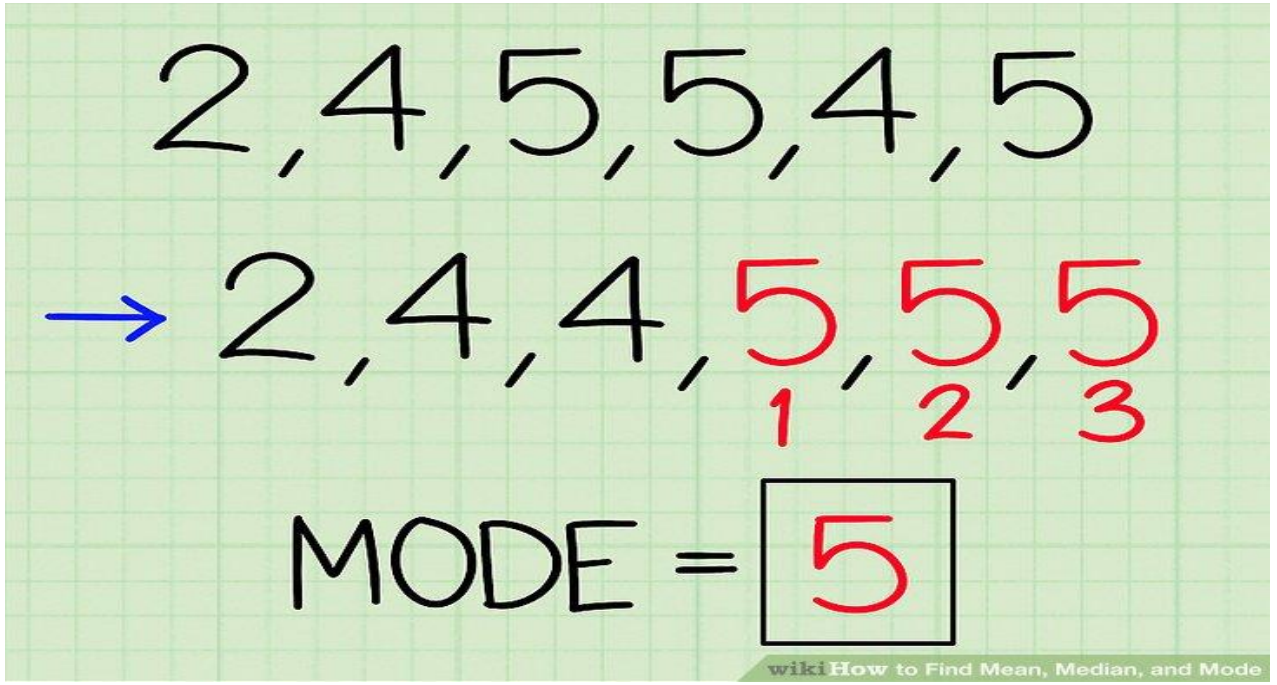
$$Median = \frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2}$$





mode

- The more frequently occurred score or value



Standard Deviation

- Standard deviation is the squared root of variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n}}$$

Note: If the data points are too far from the mean, there is higher deviation within the data set.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
6	$6 - 9 = -3$	9
7	$7 - 9 = -2$	4
10	$10 - 9 = 1$	1
12	$12 - 9 = 3$	9
13	$13 - 9 = 4$	16
4	$4 - 9 = -5$	25
8	$8 - 9 = -1$	1
12	$12 - 9 = 3$	9
		$\sum(x_i - \bar{x})^2 = 74$

Standard deviation of above case
= root(variance)
= root(9.25)
=3.041

x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
4	3	4 – 14 = –10	$(-10)^2 = 100$	3 × 100 = 300
8	5	8 – 14 = –6	$(-6)^2 = 36$	5 × 36 = 180
11	9	11 – 14 = –3	$(-3)^2 = 9$	9 × 9 = 81
17	5	17 – 14 = 3	$(3)^2 = 9$	5 × 9 = 45
20	4	20 – 14 = 6	$(6)^2 = 36$	4 × 36 = 144
24	4	24 – 14 = 10	$(10)^2 = 100$	4 × 100 = 400
32	1	32 – 14 = 18	$(18)^2 = 324$	1 × 324 = 324
	$\sum f_i = 30$			$\sum f_i(x_i - \bar{x})^2 = 1374$

Mean($\bar{x}$) = $\frac{\sum f_i x_i}{\sum f_i}$

$\bar{x} = \frac{420}{30}$

$\bar{x} = 14$

Now, finding variance

$\sum f_i(x_i - \bar{x})^2 = 1374$

$\sum f_i = 30$

Variance (σ^2) = $\frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i}$

$= \frac{1374}{30}$

= 45.8

Standard deviation (σ) = $\sqrt{45.8}$

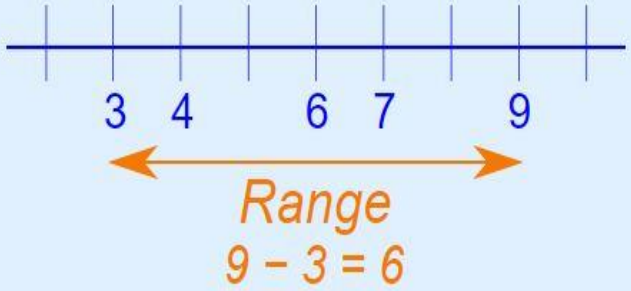
= 6.76

Range

- It is the difference between the highest value and the lowest value.

Example: In **{4, 6, 9, 3, 7}** the lowest value is 3, and the highest is 9.

So the range is 9 – 3 = **6**.



Percentile

- The value below which the percentage of data falls
- Is a way to represent position of a value in the data set
- The data should be in order

Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

That means you are at the **80th percentile**.

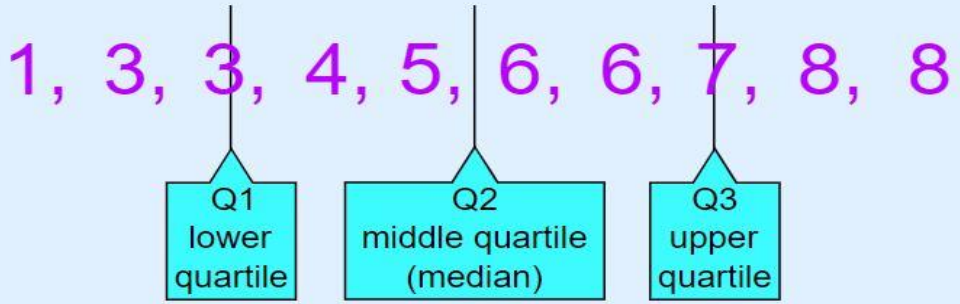
If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Quartiles

- splits the data into quarters

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are in order. Cut the list into quarters:



In this case Quartile 2 is half way between 5 and 6:

$$Q2 = (5+6)/2 = \mathbf{5.5}$$

And the result is:

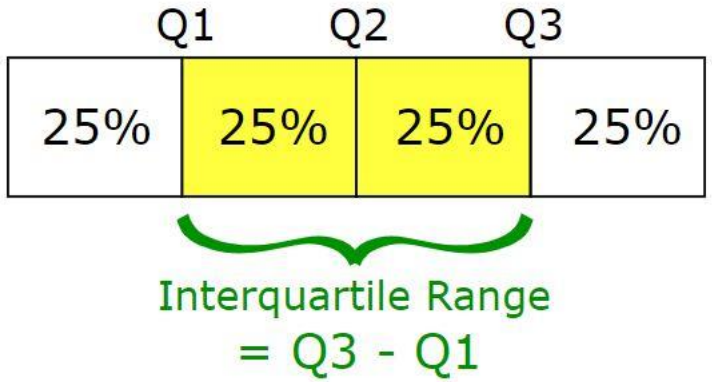
- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

- The Quartiles also divide the data into divisions of 25%, so:
- Quartile 1 (Q1) can be called the **25th percentile**
- Quartile 2 (Q2) can be called the **50th percentile**
- Quartile 3 (Q3) can be called the **75th percentile**

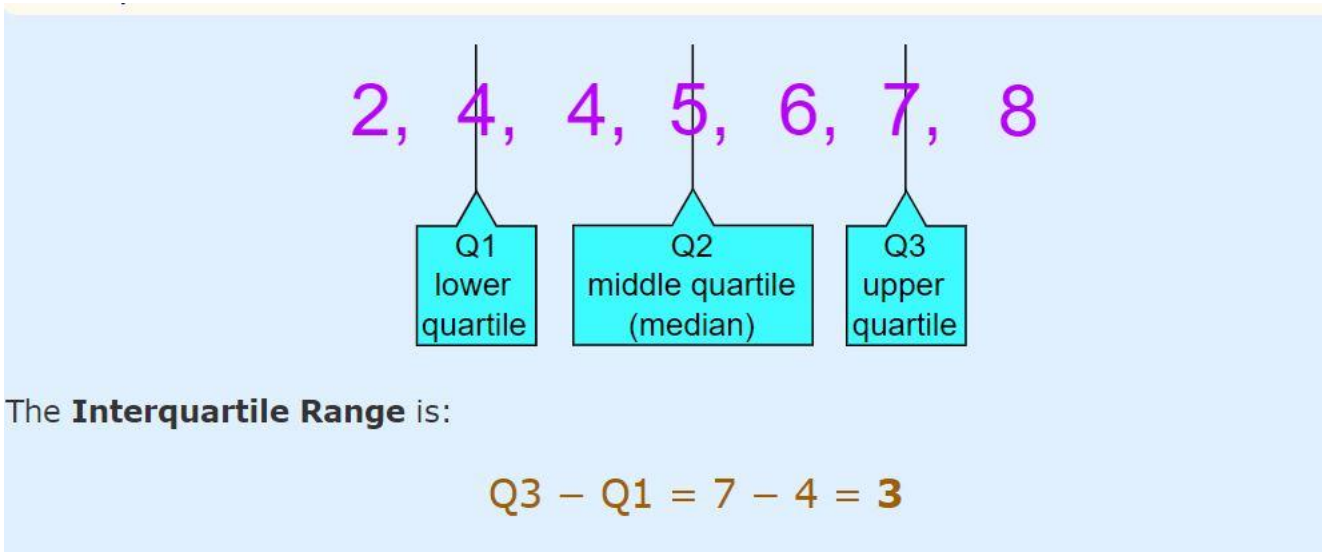
For **1, 3, 3, 4, 5, 6, 6, 7, 8, 8**:

- The 25th percentile = **3**
- The 50th percentile = **5.5**
- The 75th percentile = **7**

IQR



- The "Interquartile Range" is from Q1 to Q3:



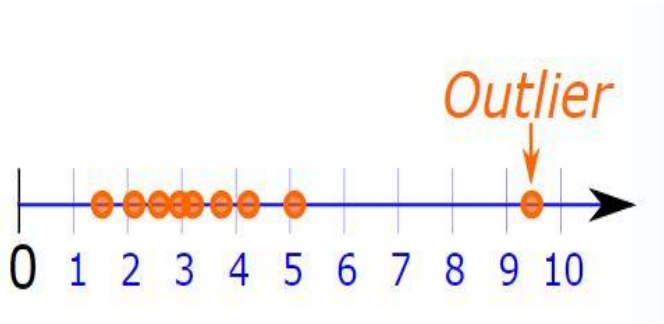
Frequency

- Frequency of an event
- Represented as “n” number of times the event occurred in an experiment
- Frequencies are often graphically represented in histograms

Number of marks	Tally marks	Frequency
1	𝍸 ||	7
2	𝍸	5
3	𝍸 |	6
4	𝍸	5
5	|||	3
Total		26

OUTLIERS

- Outliers are values that "**lie out**side" the other values.
- They can change the mean a lot, so we can either not use them (and say so) or use the median or mode instead.



WHY OUTLIERS ARE CONSIDERED ?

Example: 3, 4, 4, 5 and 104

Mean: Add them up, and divide by 5 (as there are 5 numbers):

(3+4+4+5+104) / 5 = **24**

24 does not represent those numbers well at all!

Without the 104 the mean is:

(3+4+4+5) / 4 = **4**

But please tell people you are not including the outlier.

Median: They are in order, so just choose the middle number, which is **4**:

3, 4, **4**, 5, 104

Mode: 4 occurs most often, so the Mode is **4**

3, **4, 4**, 5, 104

OUTLIERS

- **TYPES OF OUTLIERS**

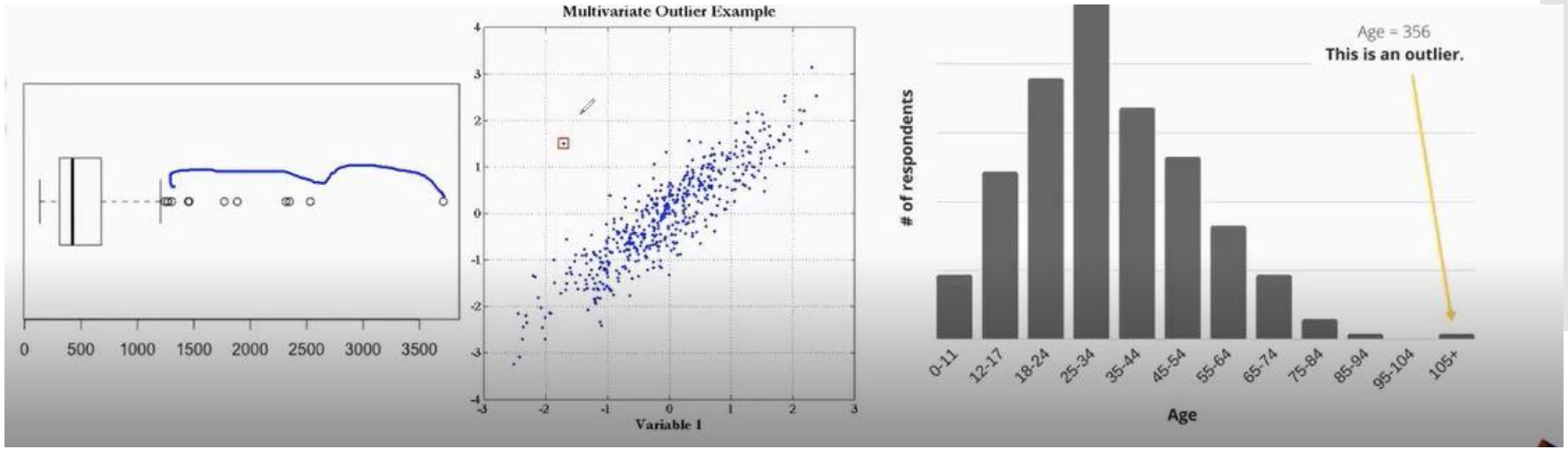
1. Univariant Outliers
2. Multivariant Outliers
3. Point Outliers
4. Contextual Outliers
5. Collective Outliers

Causes Of Outliers

- Data entry errors
- Measurement errors
- Experimental errors
- Intentional errors
- Data processing errors
- Sampling errors
- Natural errors

Detecting Outliers

- We use box plot, histograms or scatter plot to visualize and detect outliers.
- Thumb rules to detect outliers:
  - Any value beyond range of (IQR x -1.5) to (IQR x 1.5)
  - Using capping method: any value out od 5th and 95th percentile.
  - Datapoints, 3 or more standard deviation away from mean
  - Just by business understanding.



DEALING WITH OUTLIERS

- Deleting Observations
- Transforming and Binning Values
- Imputing
- Treat Outliers Separately

EDA IN PYTHON