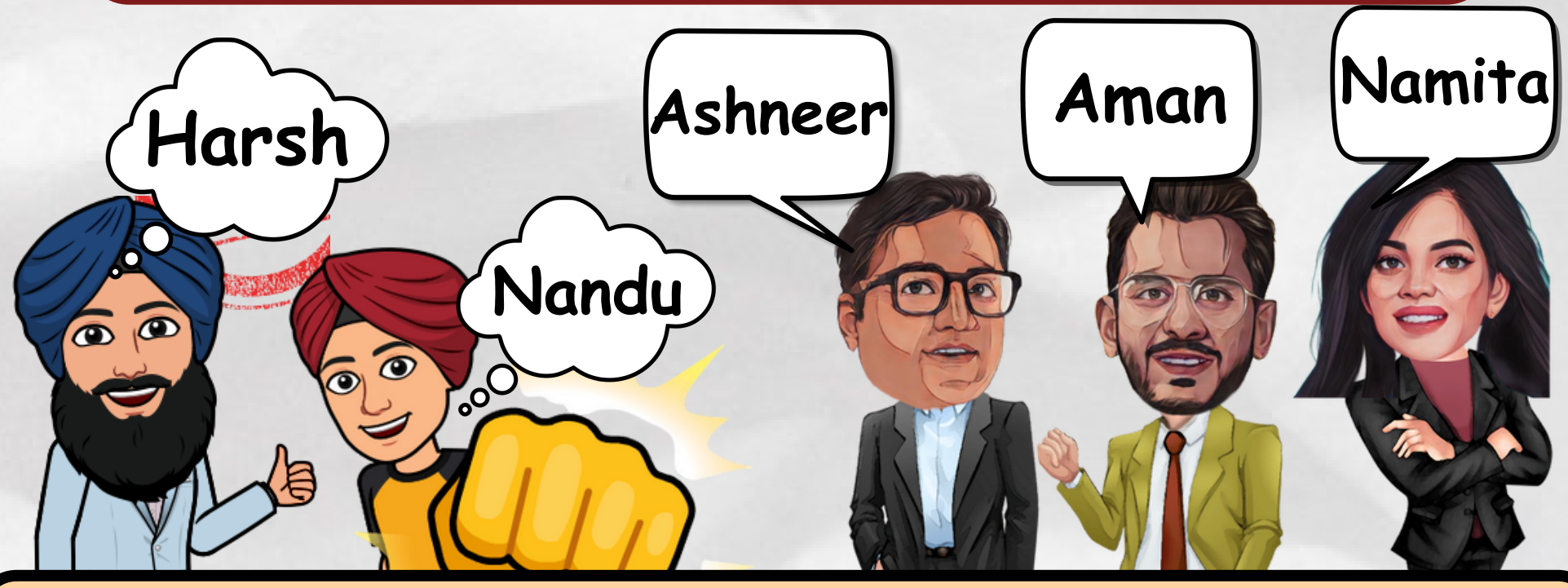
Session With Sumit

DATA SCIENCE INTERVIEW PREPARATION SERIES

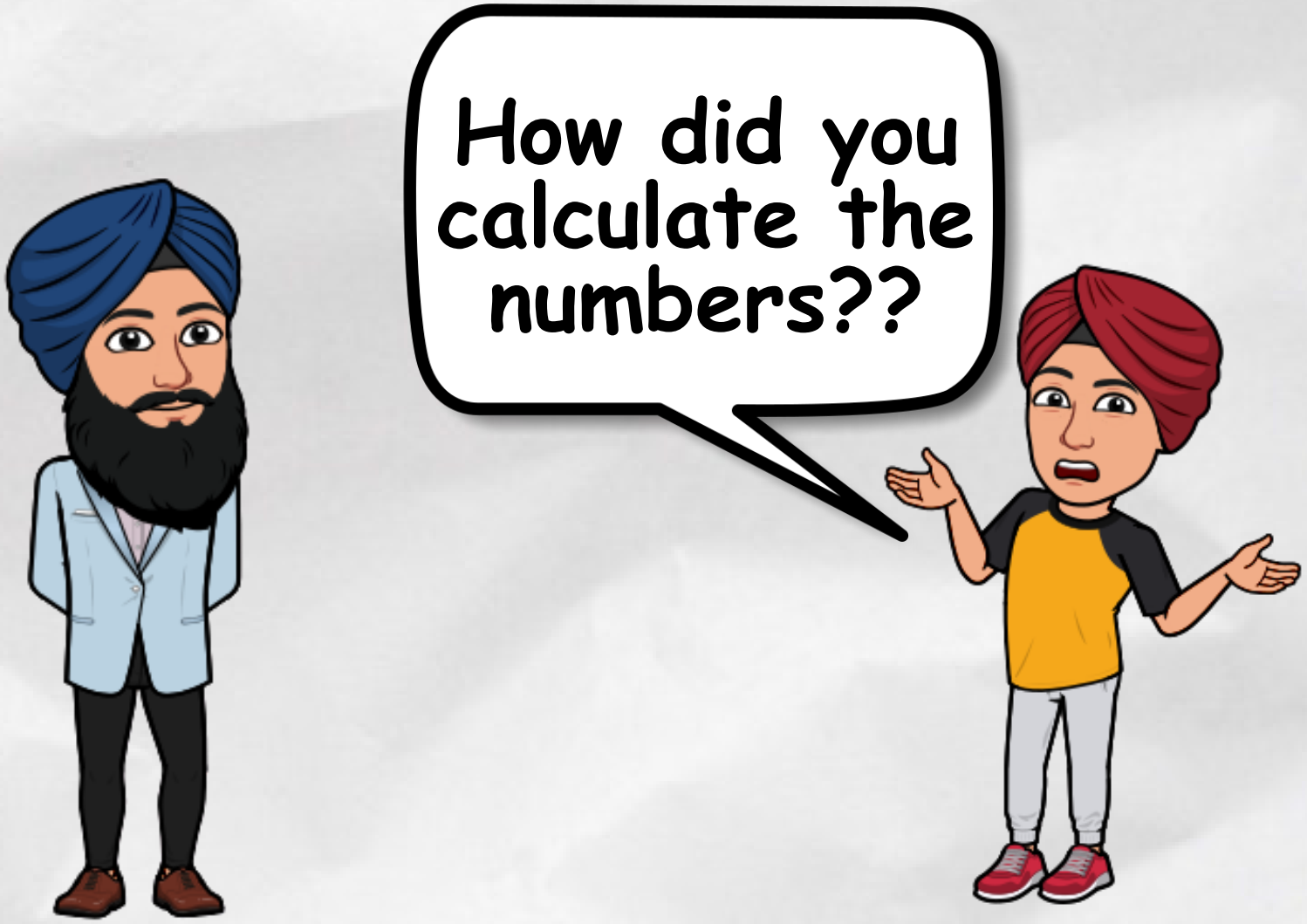


BASICS OF

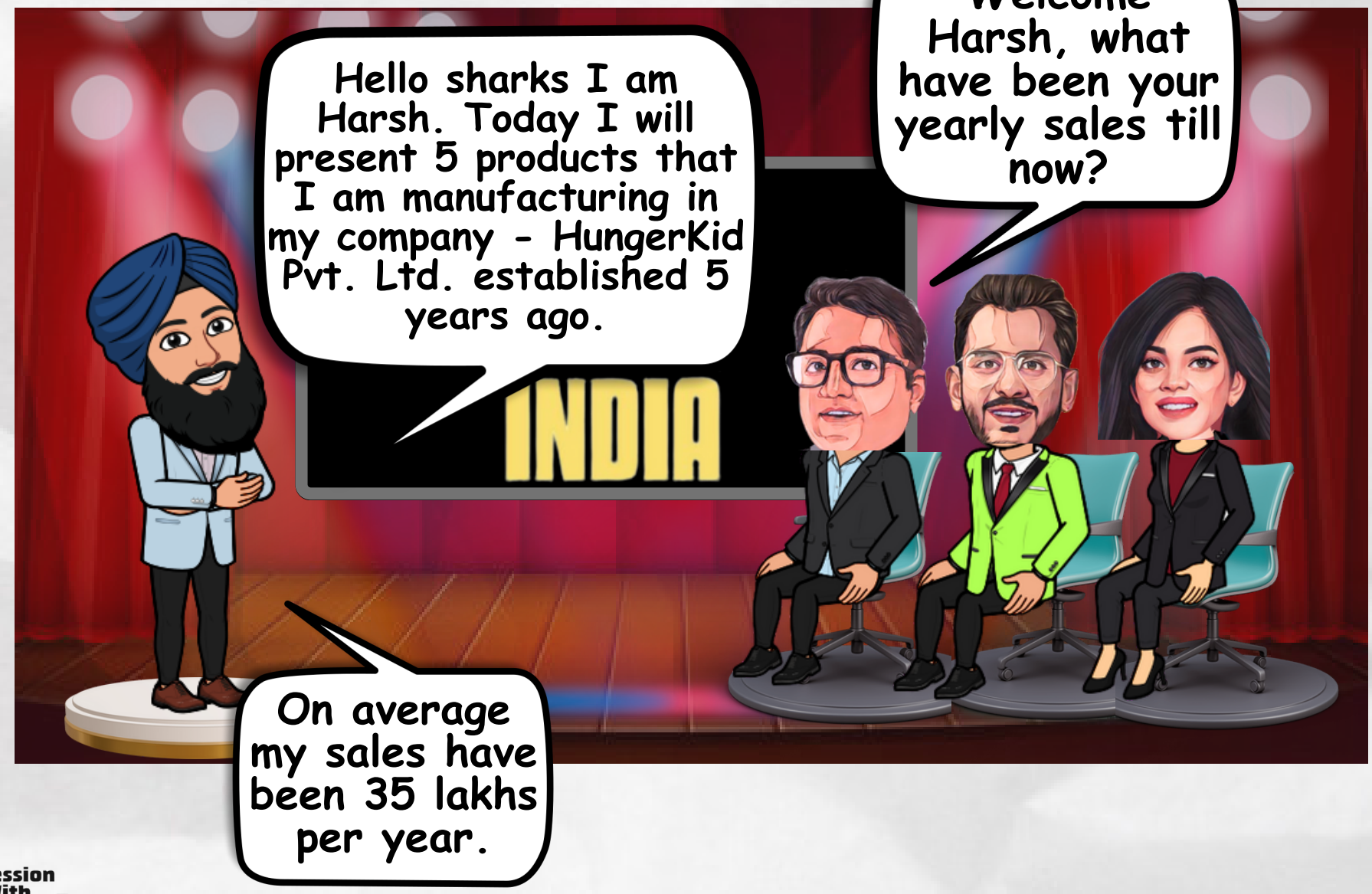
STATISTICS

Harsh, the CEO of HungerKids (A quick bites startup, producing quick bite items like chips, sandwiches, cakes, and other desi beverages) recently got an opportunity to pitch his business on Shark Tank's platform.

After the pitch was over, Harsh got back home and his 12 years old kid(Nandu) who watched the complete show, got curious to understand how he could answer all the questions so easily?



Harsh replied that he used basic Statistics. Now Nandu wanted to know about what statistics is. Harsh takes Nandu through his pitch again. Let's look at his pitch and the questions asked by the sharks



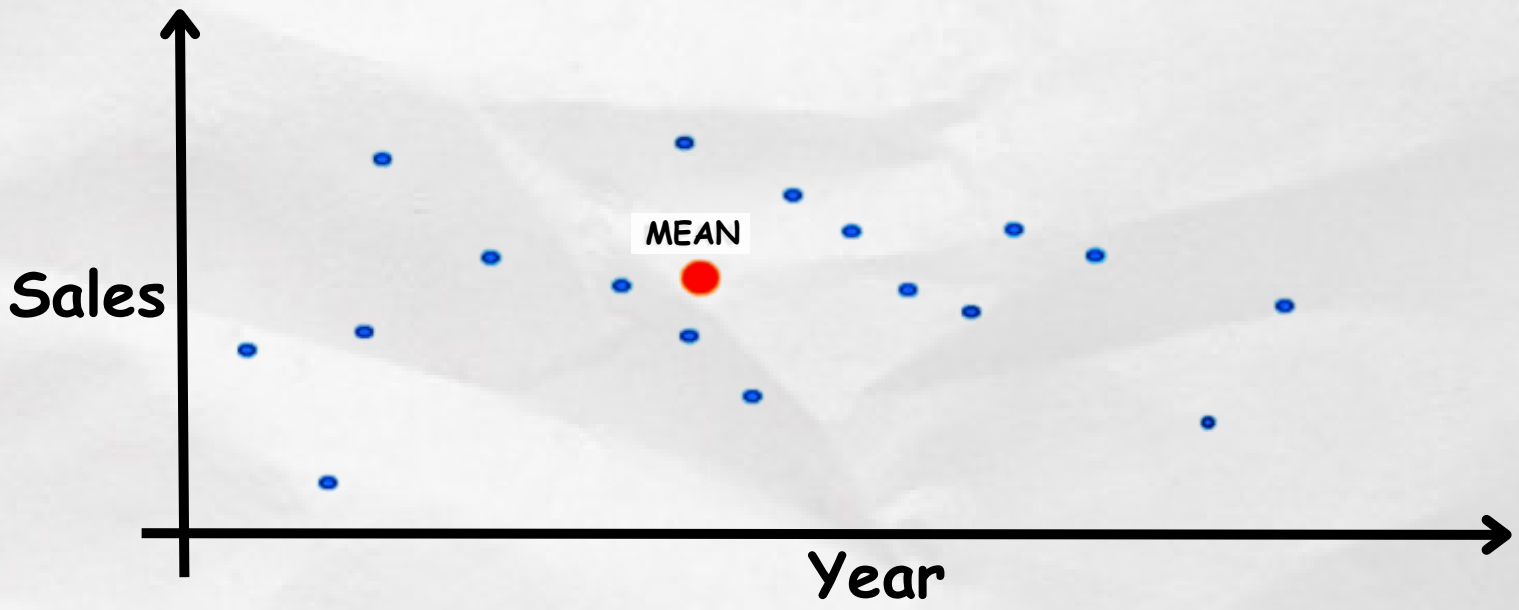
See, my yearly sale have been 10,20,25,45, 75

But what is Average?

Then why did you say 35 lakhs per year?

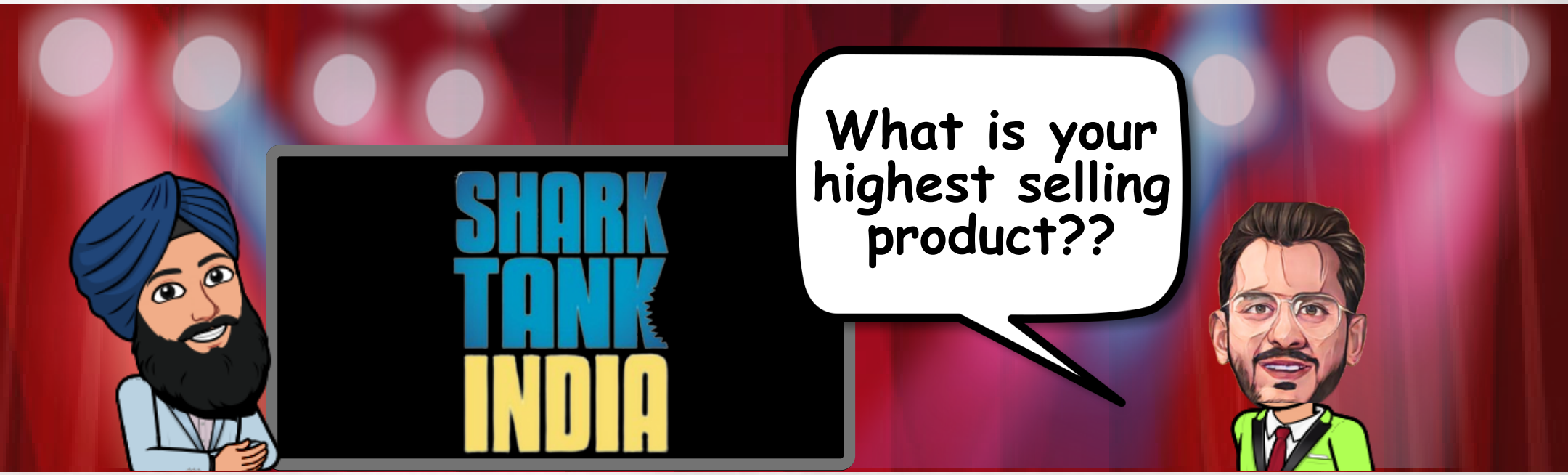
This is to present my company sales in one value or number!
Let's see how to calculate:
No. of years since the company was established = 5.
Total sales =>10+20+25+45+75
=>175
Now divide= 175/5=35 lakhs

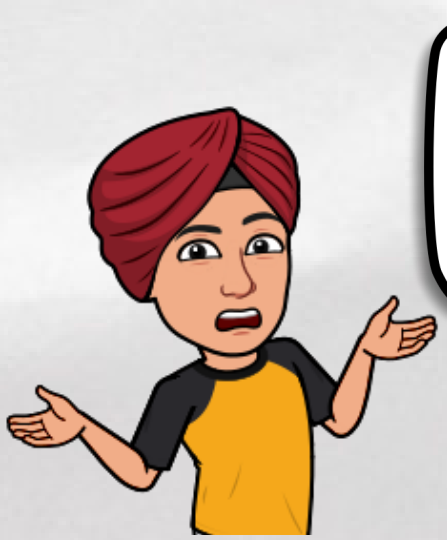
This number is called Average.

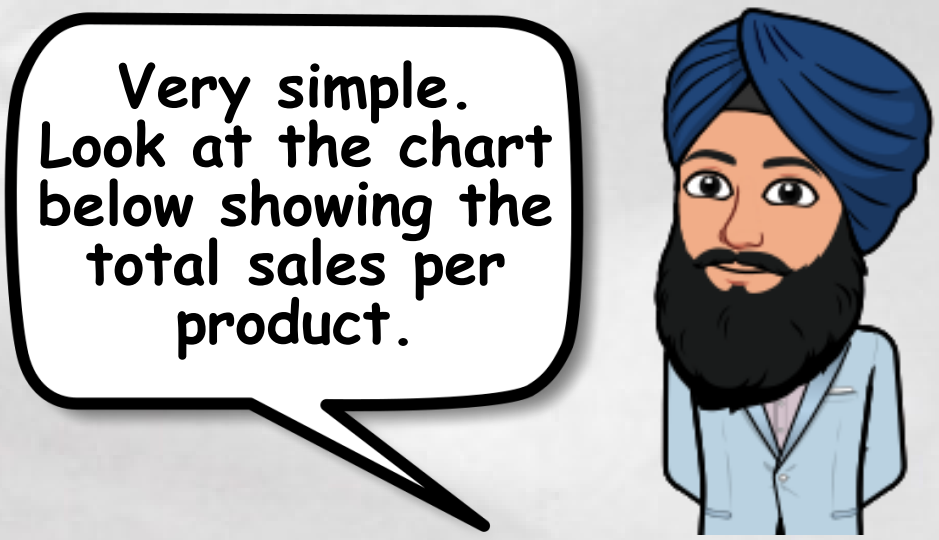


As it defines all other points in graph. So, mean or average is that central value that defines data.

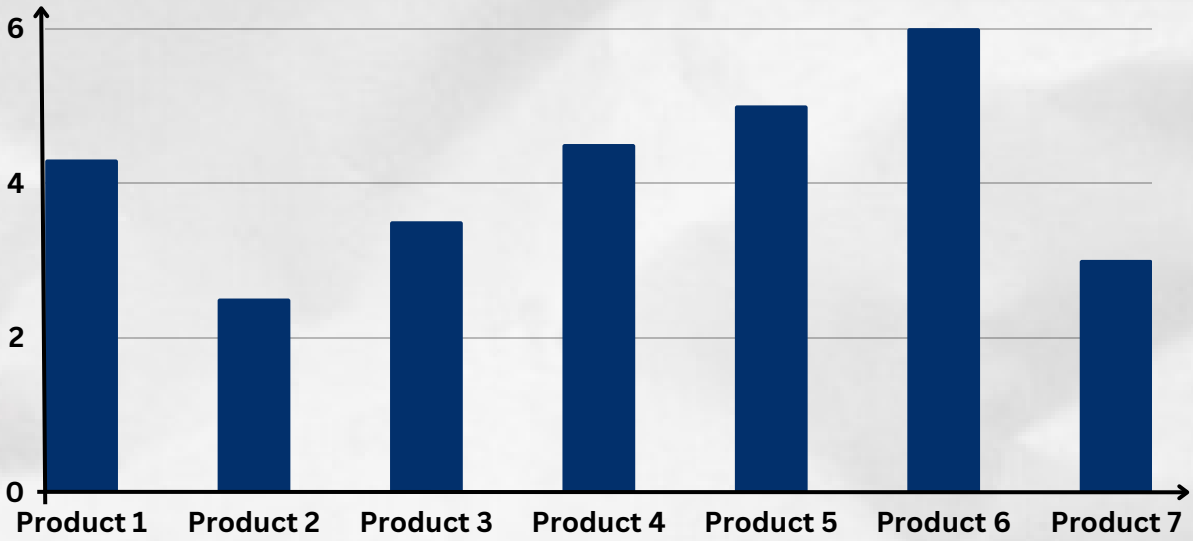
Let's revisit the pitch again..



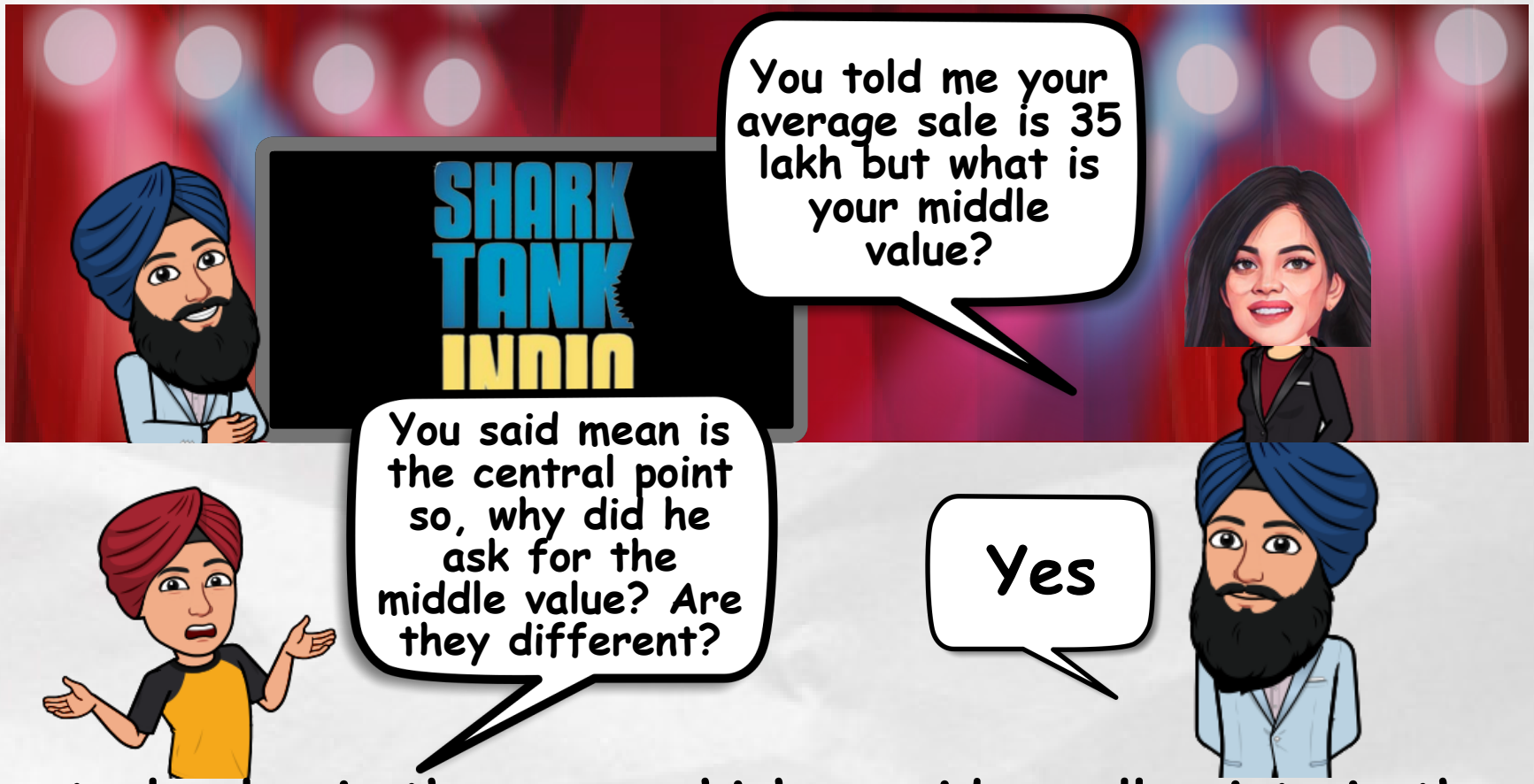




See, category 6 has the highest sales in terms of quantity. Now, in statistics, we refer to this as the mode.



Let's revisit the pitch again..



The central value is the mean which considers all points in the data.

Let me explain it.

- Yearly Sales are 10,20,25,45,75 respectively
- Now, If my sales increase, the mean will also increase.
- For eg-> if my sales have been 5,15,10,20,150
- Then my mean will be 40 (5+15+10+20+150/5)
- which is due to the extreme value.

Right? So my mean will shift if I have extremely high or extremely low values in data. It does not depict the exact middle value as it represents all data points.

For this reason, we use "median" in statistics to determine the middle value. Here,

- My median is 10,20,25,45,75 lakhs
- "Outliers" refers to values that are either extremely small or extremely high. As a result of the presence of such values, the mean will shift to the left or right.

Let's summarize-

- Statistics is a way of collecting, analyzing, and interpreting data. In simpler terms, it's a way to make sense of numbers and use them to solve problems.
- We have also summarized and described our data. This is "Descriptive Statistics" - a part of statistics.
- Descriptive statistics is a branch of statistics that involves summarizing and describing a set of data. It provides a way to analyze and understand the characteristics of a dataset, such as its central tendency, variability, and shape.

As you saw,we summarized and described everything in just one value, right ???

So, when we describe the data based on a single value around which the entire data revolves, this point is known as the "central value" or the "measure of central tendency".

These measures are:

Mean

Median

Mode

Session With Sumit

Measures of central tendency:

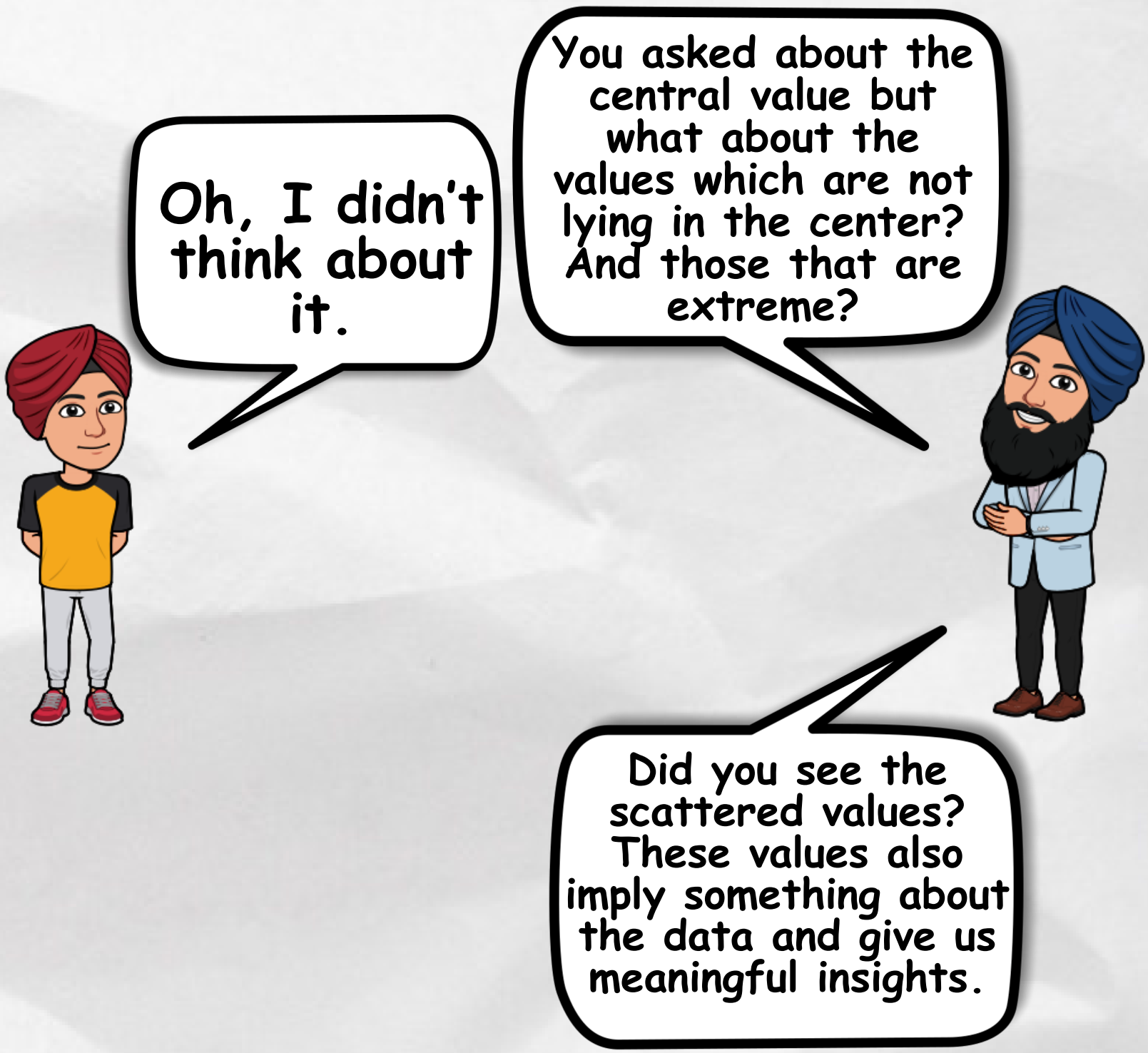
Measures of central tendency are statistical measures that describe the typical or central value of a dataset. They provide a way to summarize a large amount of data with a single value that represents the "center" of the dataset.

Mean: The mean is the arithmetic average of a dataset. It is calculated by adding up all of the values in the dataset and then dividing by the number of values. The mean can be sensitive to outliers, which are extreme values, and a few unusually high or low values can skew it.

Median: It represents the central value in a dataset when the values are sorted in numerical order. It remains unaffected by outliers and is frequently utilized when the dataset is non-normally distributed or comprises extreme values.

Mode: Mode represents the most frequently occurring value within a dataset. This statistic is ideally suited for categorical or discrete datasets, where values are non-continuous and cannot be averaged.

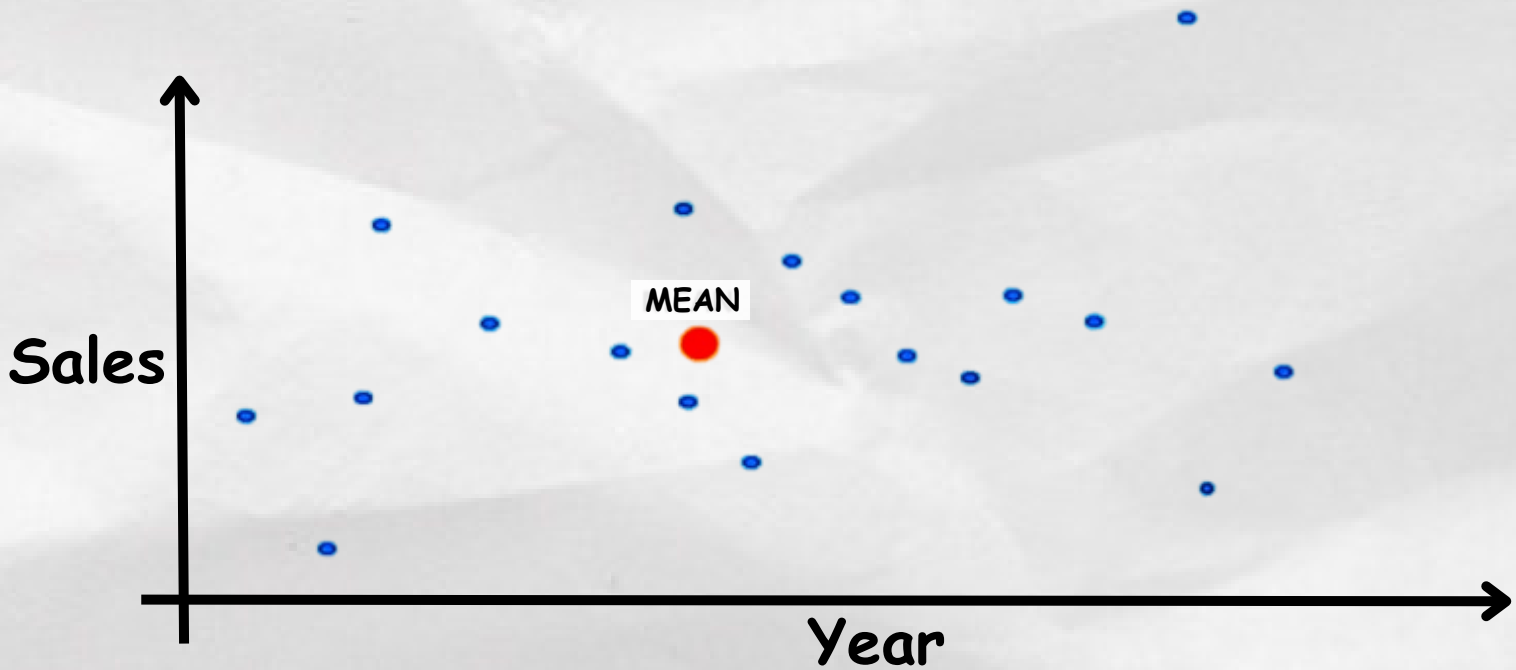
Measures of central tendency are useful in summarizing data and providing a solitary value that represents the typical or central value in a dataset.



To study them, we have one more branch of Descriptive statistics- "Measures Of Dispersion"

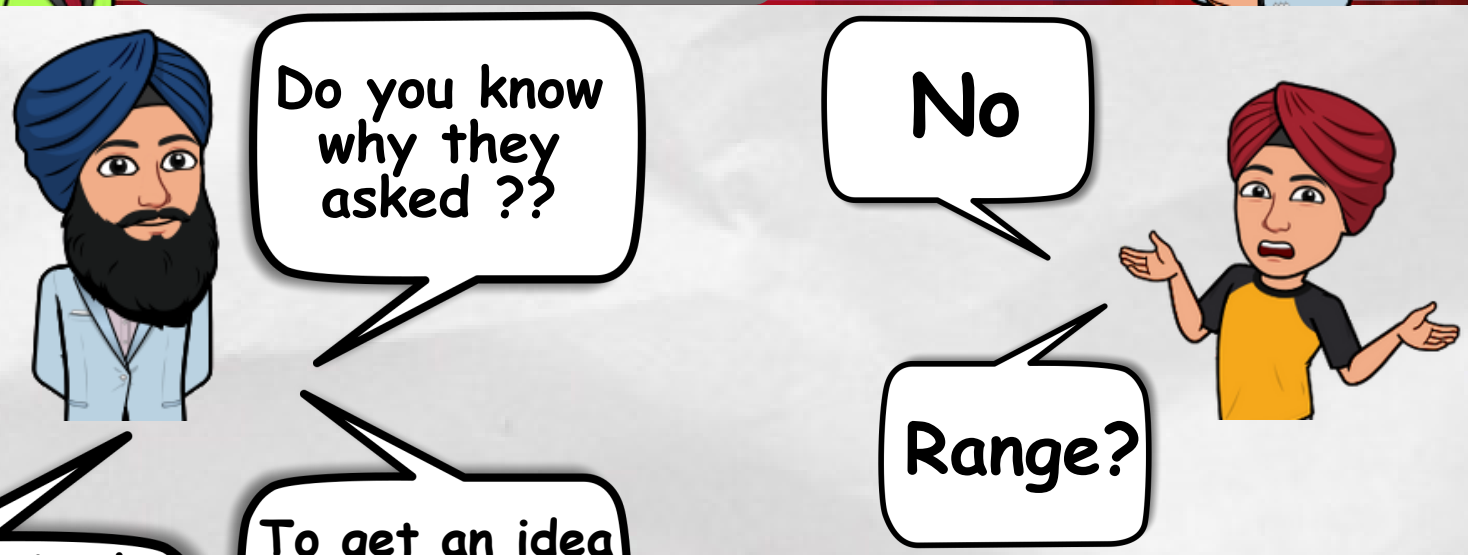
Now, let's look at the dispersed values around the Mean.

Measures of dispersion are statistical measures that describe the spread or variability of a dataset. They provide information on how spread out the data points are from the central tendency of the dataset.



Let's revisit this.
Here's the sales of the product-

Product	Period 1	Period 2	Period 3	Period 4	Period 5
A	100	150	200	250	300
B	200	250	300	350	400
C	300	350	400	450	500
D	400	450	500	550	600
E	500	550	600	650	700



It is important to measure dispersion as it provides a quick and simple way to understand the spread of a dataset.

A larger range indicates a spread out data while a smaller range indicates that the data is more tightly clustered around a central value. This info is crucial to compare different datasets or to draw conclusions about the variability of a particular variable.

So the range of my products is 300-30=270.

Here we have few measures of dispersion, i.e., standard deviation and coefficient of variation.

Let's first revisit variance:-

As the name says variance is the measure of variability. It is an important measure of dispersion in statistics as it provides info of how data points are spread out from the mean of the dataset. Basically, it measures the average of the squared differences between each data point and the mean of the dataset.

In short, how deviated my product sales are from mean.

So, let,s visit the formula first:

$$V=\frac{\sum(x-\bar{x})^2}{N}$$

Mean of Product A = (100 + 150 + 200 + 250 + 300) / 5 = 200

Variance of Product A = ((100 - 200)² + (150 - 200)² + (200 - 200)² + (250 - 200)² + (300 - 200)²) / 5
= (10000 + 2500 + 0 + 2500 + 10000) / 5
= 5100

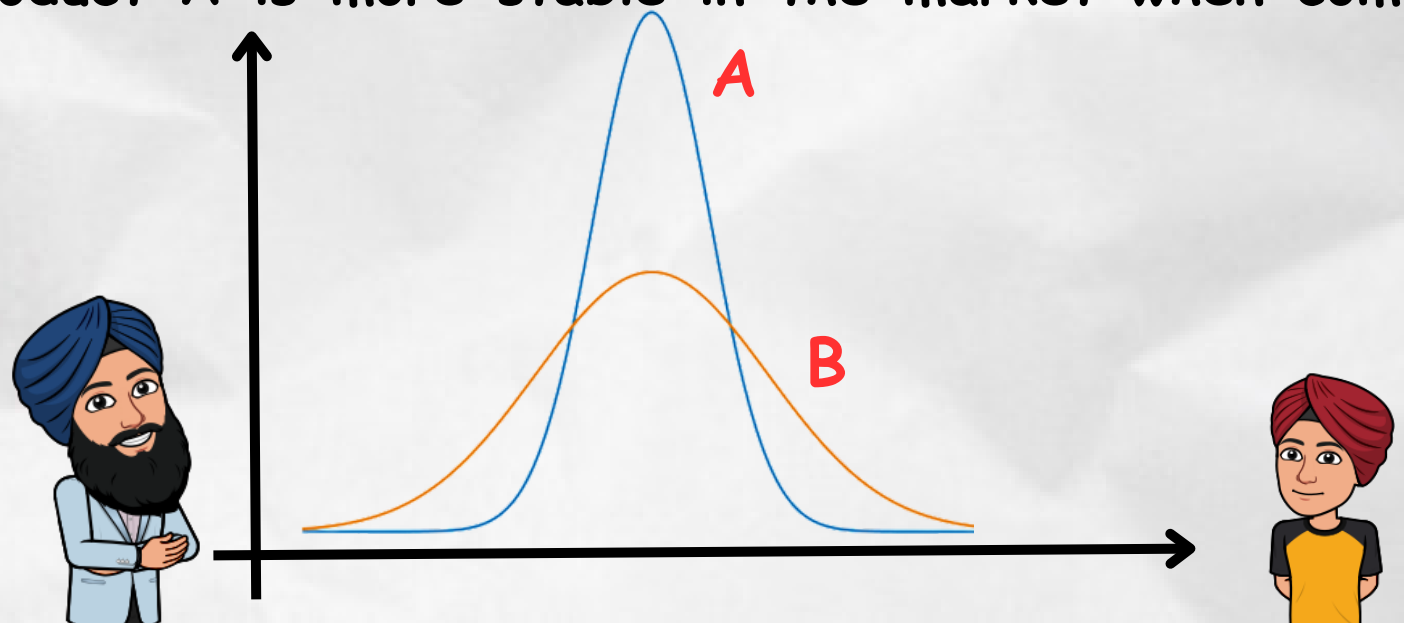
Mean of Product B= (150 + 250 + 350 + 450 + 550) / 5 = 1750/5=350

Variance of Product B= ((150 - 350)² + (250 - 350)² + (350 - 350)² + (450 - 350)² + (550 - 350)²) / 5
= 40000+10000+0+10000+40000
=100000/5=20000

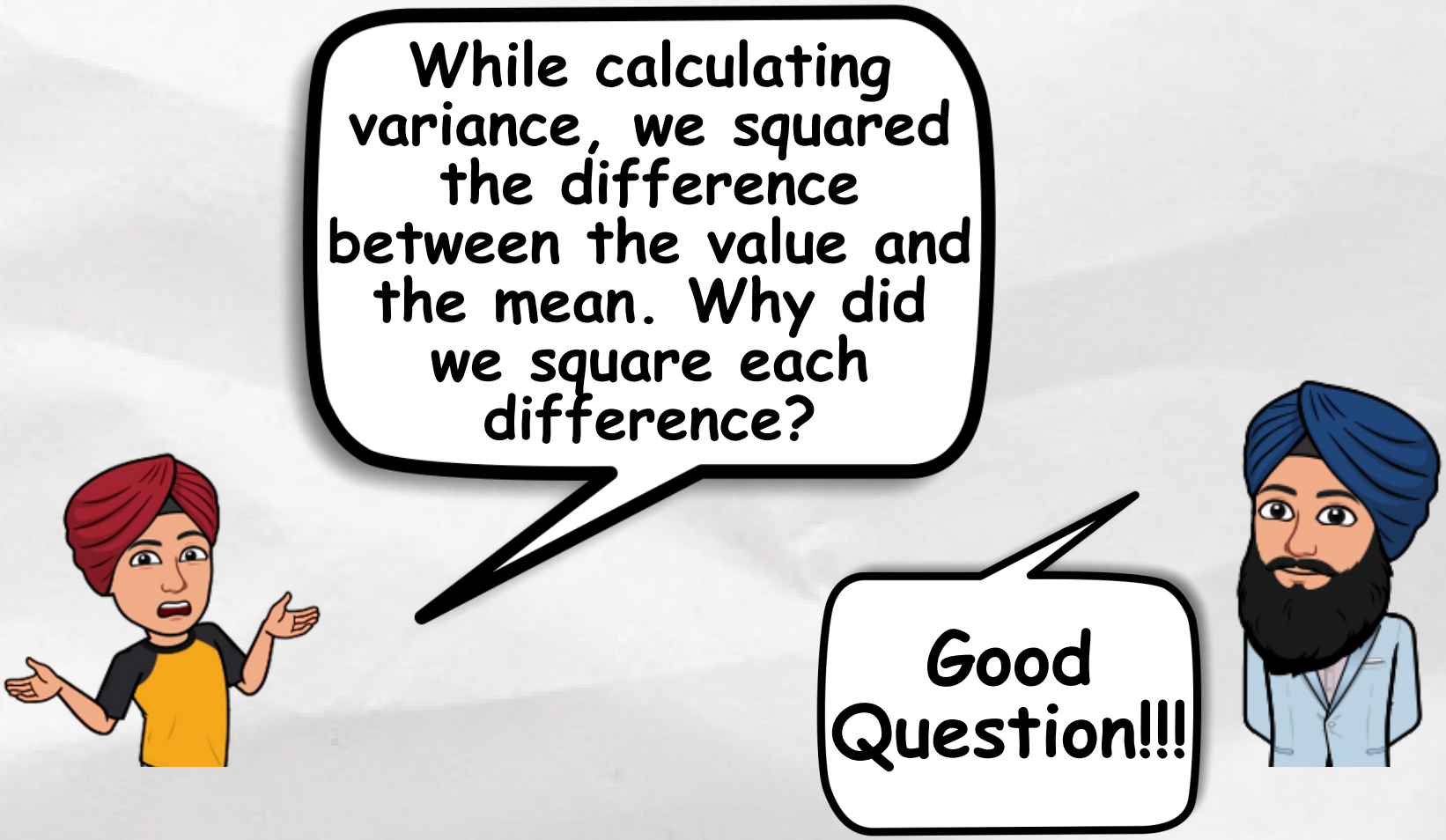
We can now say that variance of Product A < Product B

So, we can say Sales of product B is varying more than the sales of product A. For some periods, the sales of product B are very high while sometimes it's very low. Hence, product A sales are fluctuating much above and below the mean value but in the case of product A, they're more stable.

Hence, product A is more stable in the market when compared with product B.



Session With Sumit



When calculating the difference between a particular data point and the mean, the difference can be either positive or negative. When summing these differences, the positive ones can cancel out the negative ones, leading to inaccurate results. To avoid this issue, we square the values to remove the negative term.

Let's visit the definition:-

Standard deviation is a statistical measure that describes the amount of variation or dispersion of a dataset around its mean or average.

It is calculated by taking the square root of the variance, which is the average of the squared differences between each data point and the mean of the dataset.

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

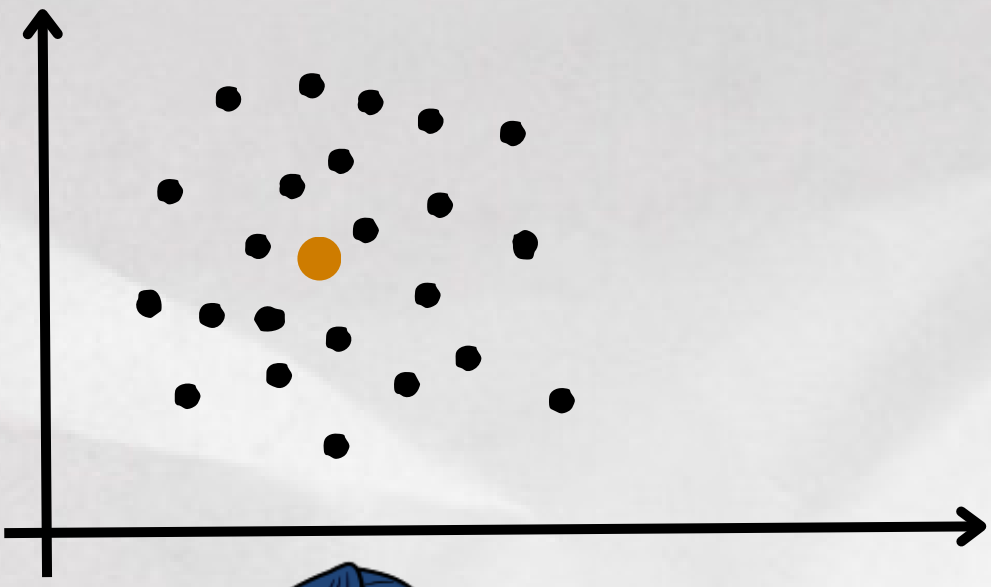
Why do we square root the value of variance?

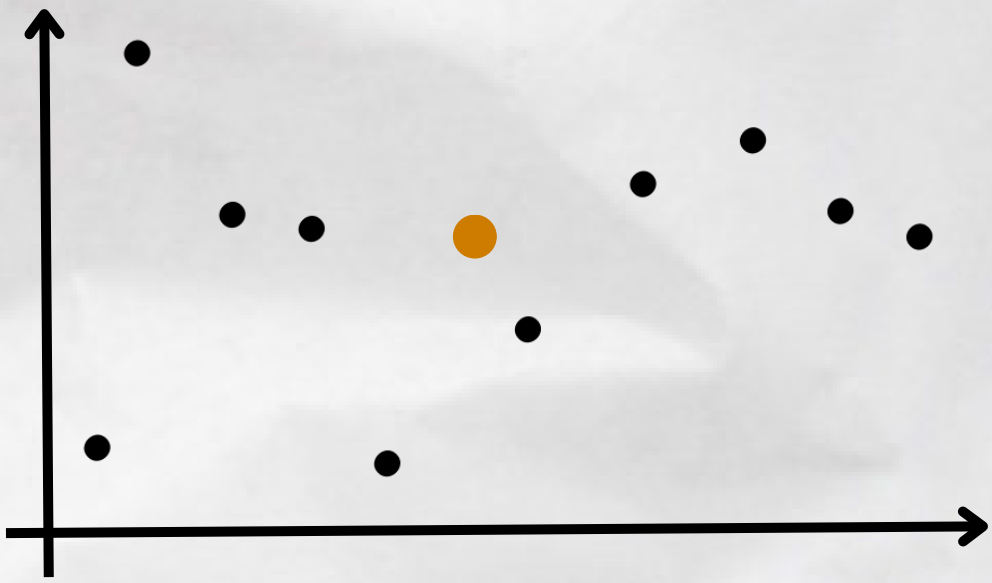
So that we get the value in absolute terms.

Understanding variance is difficult. If we look at variance of product A, it's 5100. Now, it's difficult to know what this value represents. But if we take the square root of the variance, we get 10.51 which means that the sales of product A are 10 units above and below the average sale value of product A.

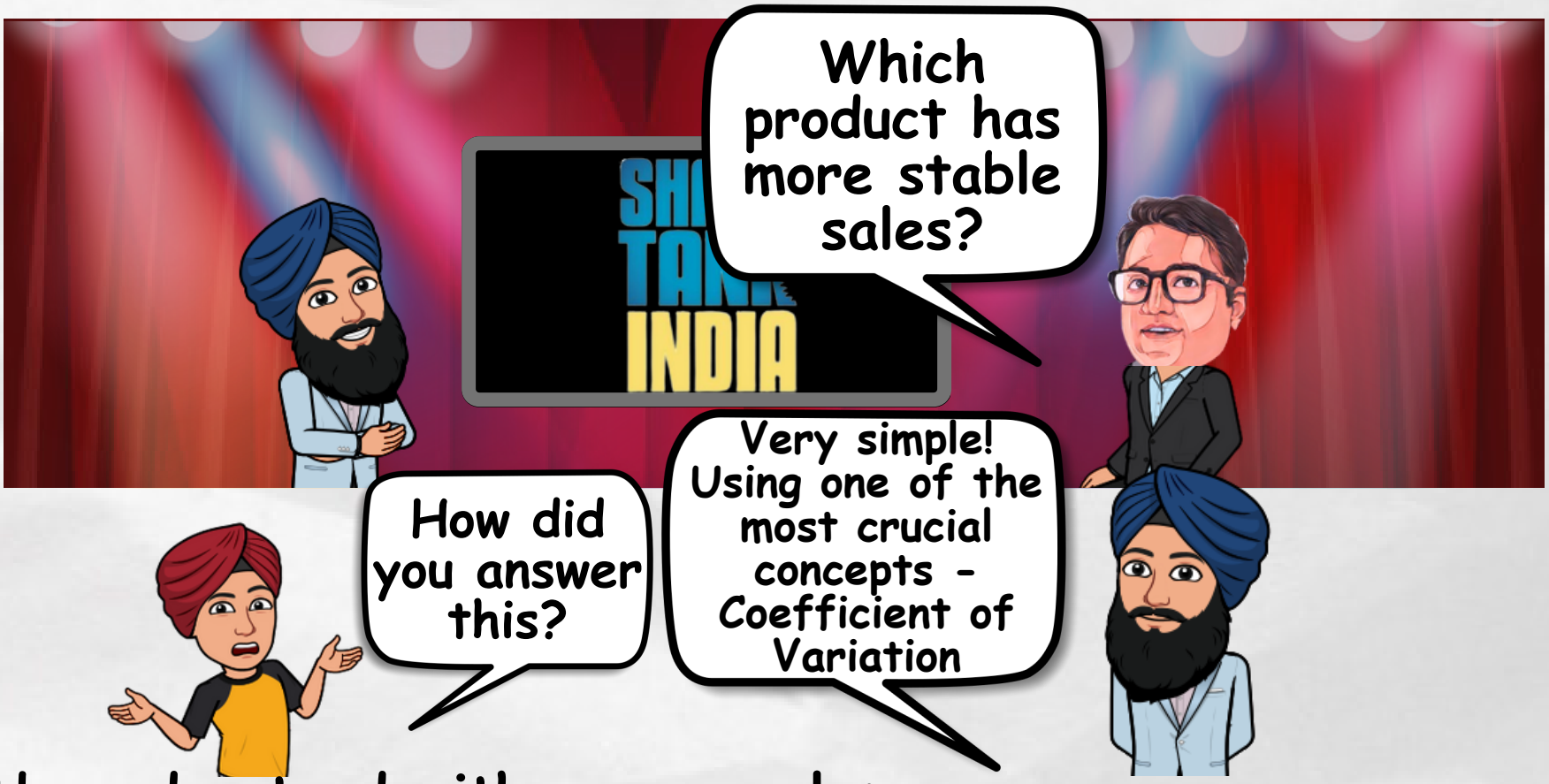
SD Of PRODUCT A, sqrt(variance)=sqrt(5100)= 10.51

Let's conclude - lesser deviation means data is clustered more towards the mean, and if the deviation is more, the data is more away from the mean and is spread widely.





Session With Sumit



Let's understand with an example:

X = [1, 2, 3]	$\overline{X} = 2$	Sx = 1
Y = [101, 102, 103]	$\overline{Y} = 102$	Sy = 1

Here series X and Y both have the same standard deviation. Now how can we compare these two series?

Which series is having more variations while which one is having fewer variations?

Simply, we will be using CV

The coefficient of variation (CV) is a statistical measure that is the ratio between the standard deviation to the mean of the data.

$$CV\ (X) = \frac{Sx}{\overline{X}} = \frac{1}{2} = 0.5$$

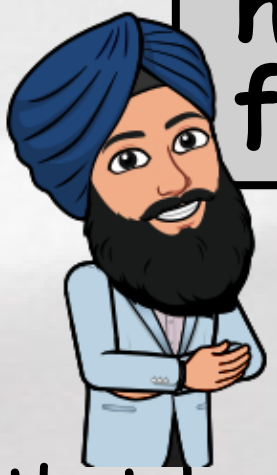
$$CV\ (Y) = \frac{Sy}{\overline{y}} = \frac{1}{102} = 0.0098$$

How to interpret CV?

Again, very simple!

COV is lower, the series is stable with less fluctuations.

COV is higher, the series is having more variations and fluctuations.



Let's take another example, consider two products, A and B.

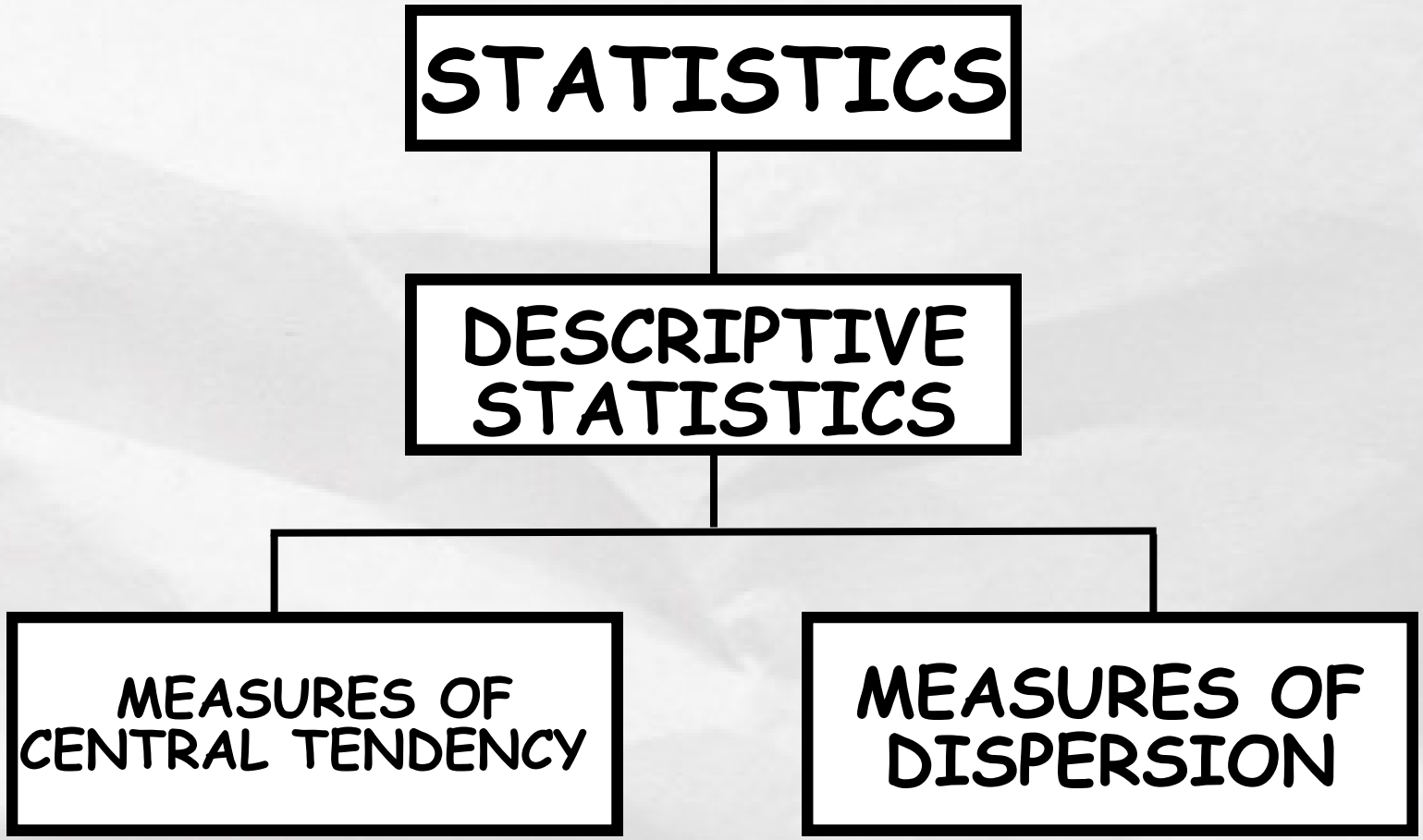
Product A | standard deviation of Sales = 5 | Average Sales : 10
Product B | standard deviation of Sales = 10 | Average Sales: 10

The coefficient of variation for Product A is 0.5 (5/10), while the coefficient of variation for Product B is 1.0 (10/10).

In this case, the higher coefficient of variation for Product B indicates that it has a higher degree of risk or variability compared to Product A.

So, we can say higher variation in sales means higher risk in investing.

That's why product A is better.


STATISTICS
DESCRIPTIVE STATISTICS
MEASURES OF CENTRAL TENDENCY
MEASURES OF DISPERSION


This is all about Descriptive Statistics where we describe and summarize the data about certain points.

If you'd like us to keep posting, support us by sharing this post.
Give it a big thumbs up and tag people who will find this helpful.

Content Designer : Hanit Kaur
Graphic Designer : Adithya Prasad
Content Lead : Sumit Shukla

With Love ❤

Session With Sumit