Module 2- Business Math and Statistics

Takeaways
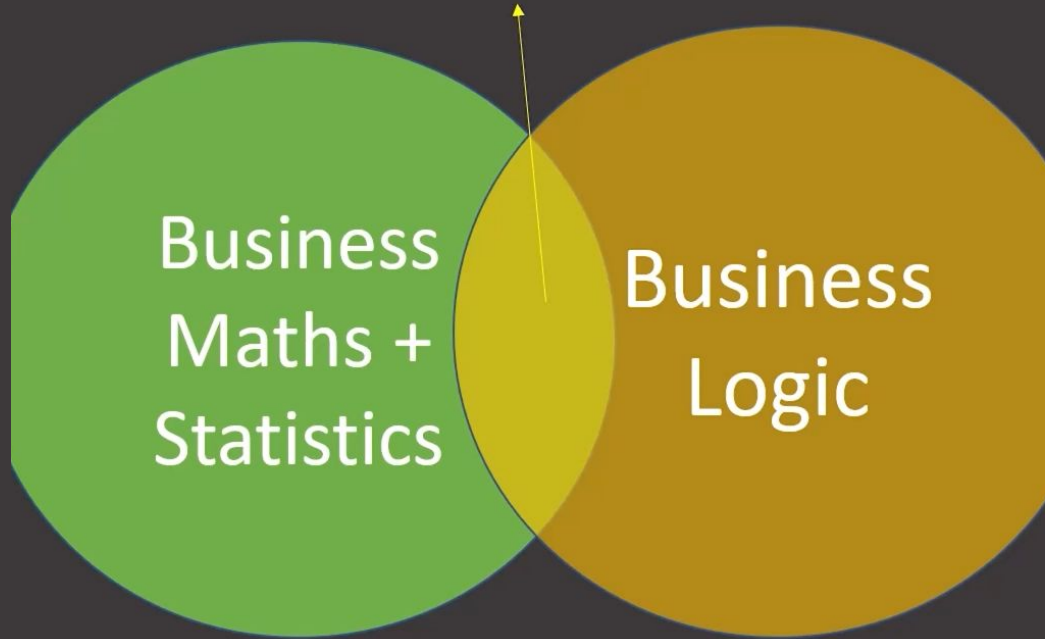
- Business Knowledge + Business Math & Statistics = **KPIs (Key Performance Indicators)**

- Formulas covered: **IF, SUM, SUMIF, COUNT, COUNTIF, AVERAGE**

- **Null values** can be handled in different ways based on the business situation:
  - Delete that data
  - Replace with the mean or median value

Business Metrics
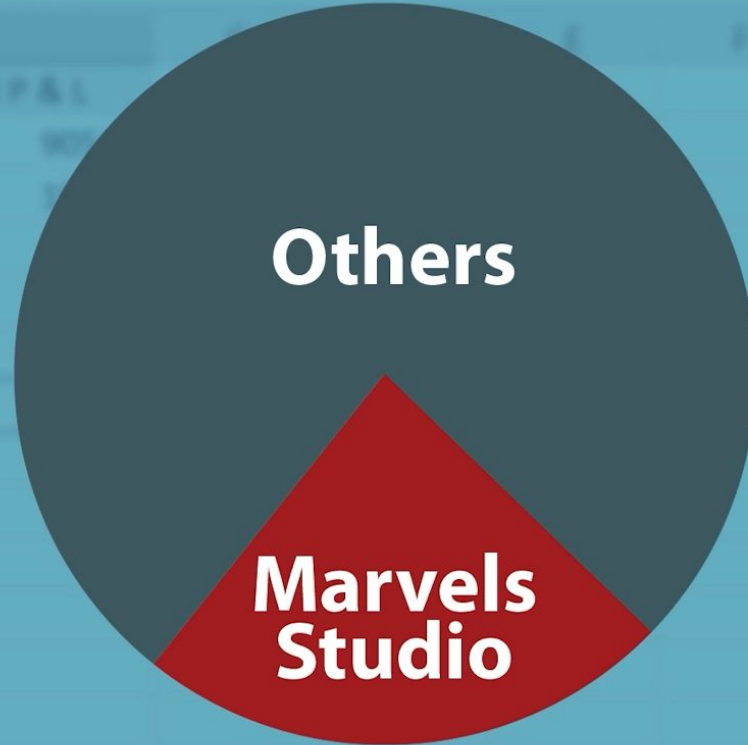
Basic Statistics

Mean

Median
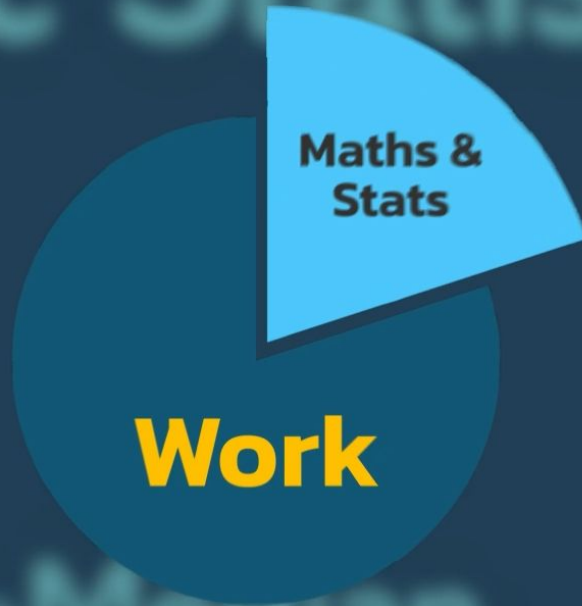
Mode

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000

Average	6250

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000

Average	6250

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000

Average	6250

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million

Nina	Rob	Rafiq	Tao	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	7500	8000	10 million

Median = 7000

Nina	Rob	Rafiq	Tao	Prem	Sofia	Mohan	Elon Musk
4000	5000	6000	7000	8000	7500	8000	10 million

Median = 7500

What is Mode?

| Name	Restaurant Vote
Rob	Mexican
Rafiq	Mexican
Nina	Italian
Sofia	Thai
Mohan	Italian
Tao	Mexican
Bantu	Indian

Mode here is Mexican

Mode means most frequently occurring value in a dataset

Takeaways

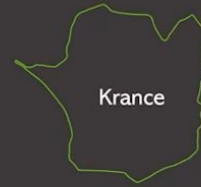- With **20% of statistics**, you can cover **80% of your work.**

- **Mean** – Synonym for **Average.**

- **Median** – Middle value of a dataset when it is ordered in ascending order.

- If the dataset has an **even number of values**, the median is the **average of the two middle values.**

- **Mode** – Most frequently occurring data value.

Variance

How far each number is from every other number in a dataset
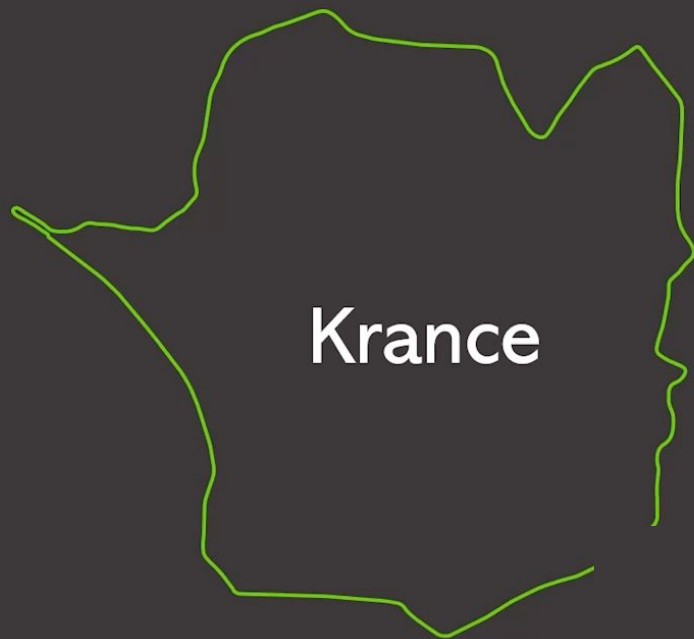
Juganda

name	yearly income
nishith	71
veeral	62
angelina	66
salma	61
Nitin	54
dhaval	67
venkat	55
Priya	60

Mean = 62

Krance

name	yearly income
mohan	99
rita	14
bhavin	75
michael	84
abdul	44
kiran	54
ahmed	98
gayatri	28

Mean = 62
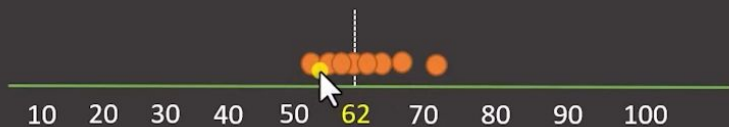
name	yearly income
nishith	71
veeral	62
angelina	66
salma	61
Nitin	54
dhaval	67
venkat	55
Priya	60
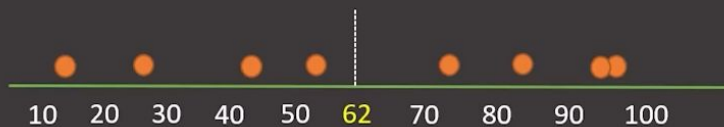
name	yearly income
mohan	99
rita	14
bhavin	75
michael	84
abdul	44
kiran	54
ahmed	98
gayatri	28

why do we square the difference in variance formula?

In the formula for variance, we calculate the average of the squared differences between each data point and the mean. The squared differences are used because they have several useful properties, such as:

1. Squaring the differences ensures that they are all positive. If we didn't square them, then the sum of the differences would always be zero, since some differences would be positive and others negative, and they would cancel each other out. By squaring them, we eliminate the negative signs and ensure that all the differences are positive.

2. Squaring the differences gives more weight to larger deviations from the mean. This is because the squared value of a large deviation is much greater than the squared value of a small deviation. By giving more weight to the larger deviations, we get a better measure of how spread out the data is.

3. Squaring the differences allows us to use calculus to find the minimum variance. Calculus requires us to differentiate the variance formula with respect to the mean, which would not be possible if we did not use squared differences.

$$variance = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\mathbb{N}}$$

$x_i = each\ value\ in\ a\ dataset$

$\bar{x} = average$

$\mathbb{N} = total\ data\ points$

Takeaways

- Variance is a measure of how spread out a distribution is. It is calculated as the **average of the squared differences from the mean.**

- The smaller the variance, the **less spread out** the data is. Conversely, the larger the variance, the **more spread out** the data is.

- Standard deviation is a measure of the amount of variation or dispersion of a set of values. It is calculated as the **square root** of the variance.

- The smaller the standard deviation, the closer the data points are to the mean. Conversely, the larger the standard deviation, the more spread out the data points are.

- The stock market's volatility is the best use case for variance and standard deviation.

Takeaways

- **Correlation is a statistical measure** that shows the degree to which two variables are related.

- A correlation coefficient can range from -1 to 1
  ↳ **-1 (perfect negative correlation) < 0 (no correlation) < 1 (perfect positive correlation)**